XII Congreso del Máster en Investigación Matemática y Doctorado en Matemáticas

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

VNIVERSITAT DE VALÈNCIA

Facultad de Matemáticas

Universitat de València

13, 14 y 15 de enero de 2025

Actas del

XII Congreso del Máster en Investigación Matemática y Doctorado en Matemáticas

Facultad de Matemáticas,
Universitat de València,
January 13th - 15th, 2025.

Preface

This book contains the contributions of the XII Congreso del Máster en Investigación y Doctorado en Matemáticas where MSc and PhD students from Universitat Politècnica de València and Universitat de València present their works.

Among the activities carried out by the InvestMat Master, there is the annual Congreso del Máster en Investigación y Doctorado en Matemáticas, which takes place in the Salón de Grados of the Faculty of Mathematics of the Universitat de València.

This congress offers the opportunity for Master and PhD students to present their research work, exchanging ideas with experts in the different research areas and improving their skills when presenting and exhibiting their work in public.

More information about the congress in https://www.uv.es/investmat/info.html.

Edited by:

– J.R. Torregrosa (Universitat Politècnica de València)
– Mª Carmen Martí (Universitat de València)
– E. López-Navarro (Universitat Politècnica de València)

I.S.B.N.: 978-84-09-75611-7

Version August 2, 2025.
Report any problems with this document to Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València, imm@imm.upv.es.

Table of Contents

The Hausdorff metric and its applications

Amaia Gastearena [♭], Belén Reverte [♮] and Aurora Sánchez. [♯]

(♭) agasiri@posgrado.upv.es
(♮) brevbad@posgrado.upv.es
(♯) asanm16a@posgrado.upv.es

1.1 Introduction

The Hausdorff metric is a mathematical tool widely used to measure the similarity between sets of points in metric spaces. Its ability to compare not only the position but also the shape of the sets makes it an essential instrument in fields such as computer vision, biomedical image processing, and pattern recognition.

This work focuses on the theoretical study of the Hausdorff metric and its applications. Starting from its mathematical definition, we analyze key properties such as stability under geometric transformations and sensitivity to outliers. We also discuss various modifications proposed to improve its performance in noisy environments, especially in tasks involving segmentation and object matching in digital images.

Finally, we address the computational complexity associated with its calculation, highlighting both its practical applications and theoretical limitations, thus establishing a solid foundation for future research in this field.

1.2 The Hausdorff Metric

We work in the Euclidean space $\mathbb{R}^n$ with standard notions of distance:

- Between points: $d(x, y) = \|x - y\|$,
- From a point to a set: $d(x, A) = \inf_{a\in A} \|x - a\|$,
- Between sets: $d(A, B) = \inf_{a\in A, b\in B} \|a - b\|$.

We also use the closed ball $B(p, r)$, and set operations such as addition $A + B = \{a + b : a \in A, b \in B\}$ and scalar multiplication $\lambda A = \{\lambda a : a \in A\}$.

Hausdorff distance: Given two non-empty compact sets $K, L \subset \mathbb{R}^n$, the Hausdorff distance is defined by:

$$\delta(K, L) = \max\left\{\sup_{x\in K}\inf_{y\in L}\|x - y\|, \sup_{x\in L}\inf_{y\in K}\|x - y\|\right\}.$$

An equivalent form is:

$$\delta(K, L) = \min\{\lambda \geq 0 \mid K \subset L + \lambda B_n,\ L \subset K + \lambda B_n\}.$$

This defines a metric on the family of non-empty compact subsets of $\mathbb{R}^n$.

To gain a more intuitive understanding of the Hausdorff distance, let us look at an example involving the distance between two compact sets in the plane.

Example 1.2.1 *Consider the triangle and the circle given by:*

$$K = \{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 2,\ 1 \leq y \leq 3 - |2(x-1)|\},$$
$$L = \{(x, y) \in \mathbb{R}^2 : (x-3)^2 + (y-1)^2 \leq 1\},$$

respectively.

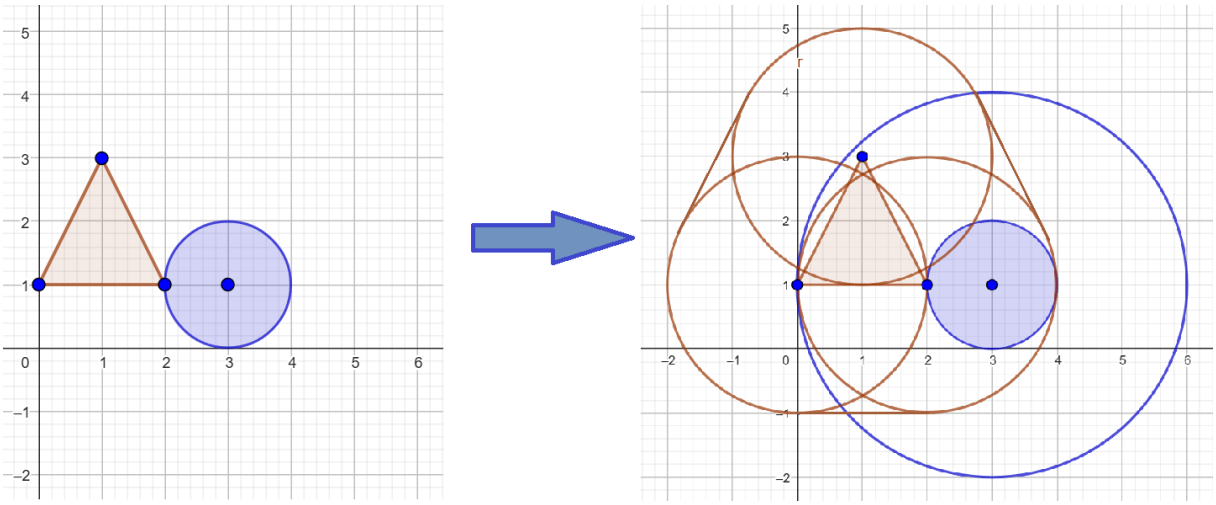


Figure 1.1: Hausdorff Distance Between Two Sets

Although the minimal distance between K and L is $d(K, L) = 0$ due to overlapping regions, the Hausdorff distance $\delta(K, L) = 2$ more accurately reflects their geometric separation. Specifically, the center of the circle L is at $(3, 1)$, and the distance to the nearest point of K is 3. Since the radius of L is 1, enlarging L by $\lambda = 2$ ensures that K is contained within $L + \lambda B_n$, and similarly L is contained within $K + 2B_n$. Thus, $\delta(K, L) = 2$.

1.3 Modifications of the Hausdorff Metric for Object Matching

The Hausdorff metric is a useful tool for measuring the similarity between sets, but it has important limitations in practical applications like object matching and shape recognition. Its high sensitivity to noise and segmentation errors can misrepresent object similarity. To address these issues, the **Modified Hausdorff Distance (MHD)** was developed, aiming for greater robustness in noisy environments.

1.3.1 Modifications of the Hausdorff Distance

In image analysis, the Hausdorff distance is computed between finite point sets. Its reliance on the maximum minimum distance makes it sensitive to isolated outlier points, which can significantly distort the measurement. The MHD reduces this influence by averaging the minimum distances, making it more stable under perturbations.

In [Dubuisson and Jain(1994)], 24 variations of the Hausdorff distance were proposed and tested. Synthetic images were first used to assess their discrimination power, followed by noise robustness tests, and finally experiments with real-world images. MHD consistently demonstrated superior performance in differentiating objects under realistic conditions.

1.3.2 Directed Distances

The Hausdorff distance can be decomposed into directed distances:

$$\delta(A, B) = \max\{h(A, B), h(B, A)\},$$

where $h(A, B) = \max_{x \in A} \min_{y \in B} \|x - y\|$ measures how far A is from B.

In practical applications, it is helpful to relax the strict maximum by ranking points by their nearest neighbor distance and selecting other percentiles:

$$h_K(A, B) = K^{\text{th}}_{a \in A} d(a, B).$$

Here, the K-th nearest distance provides greater flexibility: for instance, the 50th percentile corresponds to the median distance.

Six directed distances d_1 to d_6 are defined, using minimum, percentiles (50%, 75%, 90%), maximum, and mean distance values between points in A and B.

1.3.3 Non-Directed Distance Measures

To obtain symmetric measures, combinations of directed distances are used:

$$\begin{aligned}
f_1(d(A, B), d(B, A)) &= \min(d(A, B), d(B, A)), \\
f_2(d(A, B), d(B, A)) &= \max(d(A, B), d(B, A)), \\
f_3(d(A, B), d(B, A)) &= \frac{d(A, B) + d(B, A)}{2}, \\
f_4(d(A, B), d(B, A)) &= \frac{N_a d(A, B) + N_b d(B, A)}{N_a + N_b}.
\end{aligned}$$

Applying these to the six directed distances generates 24 possible non-directed measures. Notably, D_{18} corresponds to the classical Hausdorff distance.

directed distance	function			
	f_1	f_2	f_3	f_4
d_1	D_1	D_2	D_3	D_4
d_2	D_5	D_6	D_7	D_8
d_3	D_9	D_{10}	D_{11}	D_{12}
d_4	D_{13}	D_{14}	D_{15}	D_{16}
d_5	D_{17}	D_{18}	D_{19}	D_{20}
d_6	D_{21}	D_{22}	D_{23}	D_{24}

Figure 1.2: Possible combinations of non-directed distance measures between two point sets. Image obtained from [Dubuisson and Jain(1994)].

1.3.4 Object Matching using Distance Measures

An effective distance for object matching must satisfy two key properties:

1. **Discrimination Power**: ability to distinguish different objects,
2. **Sensitivity to Noise**: distance should increase with increasing differences.

In [Dubuisson and Jain(1994)], only D_{18} was found to satisfy all properties of a metric. Other measures either violated the triangle inequality or failed the separation property, posing challenges for robust object matching.

Selection of Distance Measures Based on Discrimination Power

Experiments on synthetic images revealed that several measures (e.g., D_1–D_9, D_{13}, D_{17}, D_{21}) had poor discrimination power, often yielding a distance of zero between distinct objects. Measures based on operator f_2 (maximum) exhibited better separation. Consequently, D_{10}, D_{14}, D_{18}, and D_{22} were selected for detailed evaluation.

Behavior Against Noise

Behavior against noise was assessed using four noise models: random perturbations (K_n), added lines (K_a), removed lines (K_d), and pixel noise (K_u).

For D_{18} (standard Hausdorff distance), Figure 1.3 shows that it remains high even with small noise levels and does not vary significantly as noise increases, making it overly sensitive to outliers.

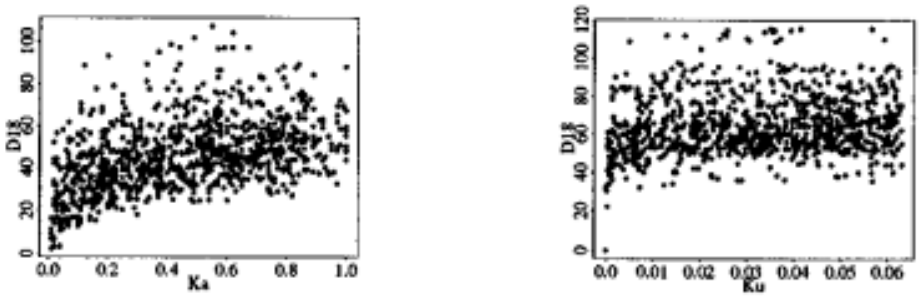


Figure 1.3: Behavior of the Hausdorff distance (D_{18}) under K_n and K_p noise models. Image obtained from [Dubuisson and Jain(1994)].

Generalized distances like D_{10} and D_{14} allow partial matching and are less sensitive to isolated points. However, as seen in Figure 1.4, they can report zero distance even when significant noise is present.

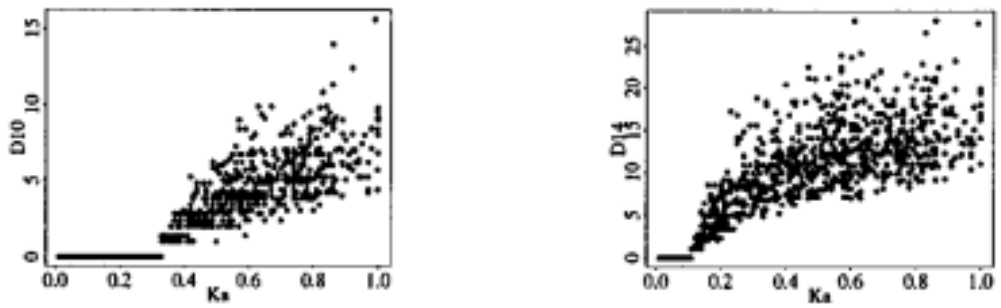


Figure 1.4: Behavior of distances D_{10} and D_{14} under pixel noise K_p. Image obtained from [Dubuisson and Jain(1994)].

The Modified Hausdorff Distance D_{22} displayed the best behavior, increasing steadily with noise (see Figure 1.5), providing a good balance between sensitivity and robustness.

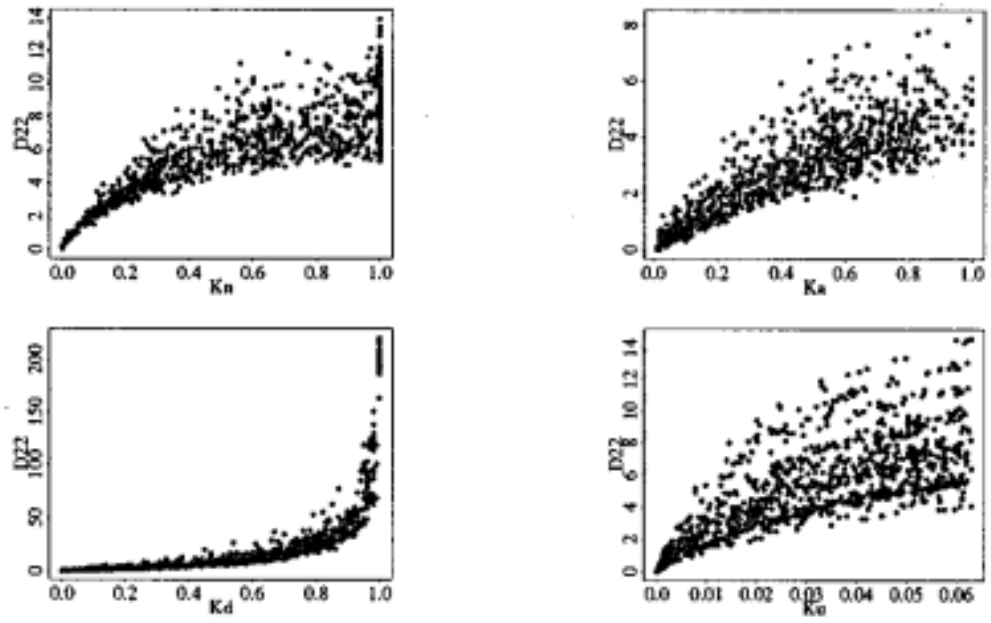


Figure 1.5: Behavior of the Modified Hausdorff Distance (D_{22}) under different noise models. Image obtained from [Dubuisson and Jain(1994)].

1.3.5 Application to Real Images

MHD was applied to real-world object edge images (e.g., moving vehicles). As shown in Figure 1.6, four edge images were analyzed, with small perturbations introduced.

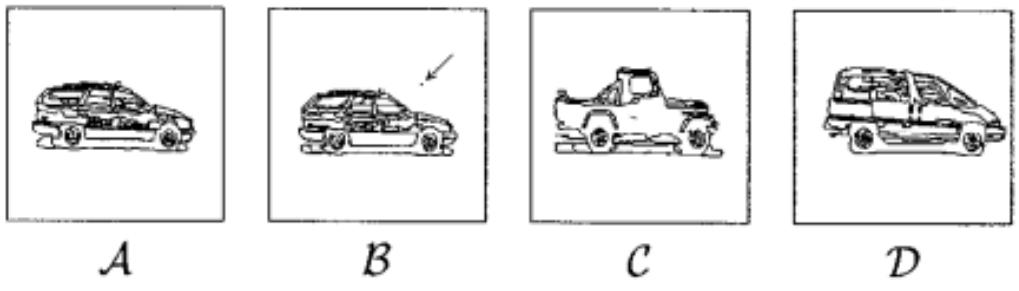


Figure 1.6: Edge images of four real objects used for object matching tests. Image obtained from [Dubuisson and Jain(1994)].

Results showed that the classical Hausdorff distance (D_{18}) was overly sensitive to small outliers, while MHD (D_{22}) correctly grouped similar objects, as illustrated in Figure 1.7.

D_{10}

	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
$\mathcal{A}$	0	**2**	5	3
$\mathcal{B}$	**2**	0	7	2
$\mathcal{C}$	5	7	0	6
$\mathcal{D}$	3	2	6	0

D_{14}

	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
$\mathcal{A}$	0	**3**	10	6
$\mathcal{B}$	**3**	0	10	7
$\mathcal{C}$	10	10	0	12
$\mathcal{D}$	6	7	12	0

D_{18}

	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
$\mathcal{A}$	0	**32**	22	32
$\mathcal{B}$	**32**	0	107	25
$\mathcal{C}$	22	107	0	36
$\mathcal{D}$	32	25	36	0

D_{22} (MHD)

	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
$\mathcal{A}$	0	**1**	6	4
$\mathcal{B}$	**1**	0	6	4
$\mathcal{C}$	6	6	0	6
$\mathcal{D}$	4	4	6	0

Figure 1.7: Comparison of distance measures D_{10}, D_{14}, D_{18}, and D_{22} applied to real images. Image obtained from [Dubuisson and Jain(1994)].

1.3.6 Results

The Modified Hausdorff Distance (MHD) demonstrated clear advantages:

- It grows steadily with object dissimilarity,
- It is robust to outliers and small segmentation errors.

These properties make it the preferred distance measure for object matching tasks in noisy environments.

1.4 Computational Complexity of the Hausdorff Distance

1.4.1 Problem Presentation

The Hausdorff distance measures how far two sets are from each other, but its computation becomes complex when dealing with infinite or semi-algebraic sets. While finite set distances are computable in polynomial time, infinite sets often require advanced algebraic techniques, particularly when described by polynomial constraints.

1.4.2 Semi-Algebraic Sets

Semi-algebraic sets are subsets of $\mathbb{R}^n$ defined by polynomial inequalities and equations. Although powerful for representing complex geometries, their computation is complicated by the presence of real (possibly irrational) coefficients, requiring specialized transformations to work over rationals or integers.

1.4.3 General Decision Algorithm

The decision problem studied is: Given two semi-algebraic sets A, B and threshold t, determine if

$$d_H(A, B) \leq t.$$

This translates into a quantified logical formula:

$$\forall a \in A, \exists b \in B \text{ such that } \|a - b\| \leq t.$$

Solving such formulas involves significant complexity due to the alternation of quantifiers ($\forall$, $\exists$).

1.4.4 Algebraic Complexity

Problems involving real polynomial constraints are classified into complexity classes:

- $\forall\exists\mathbb{R}$: Universal-existential theory over the reals.
- $\forall\exists_<\mathbb{R}$: Same, but restricted to strict inequalities without negations.

Determining $d_H(A, B) \leq t$ places the problem in $\forall\exists_<\mathbb{R}$.

1.5 Problems and Results

1.5.1 Main Result

Theorem 1 *The Hausdorff distance decision problem for semi-algebraic sets is $\forall\exists_<\mathbb{R}$-complete.*

This result shows that deciding whether $d_H(A, B) \leq t$ is as hard as solving the most difficult problems in the complexity class $\forall\exists_<\mathbb{R}$. Consequently, it surpasses well-known classes like NP and $\exists\mathbb{R}$ in computational difficulty, highlighting the inherent challenge of computing the Hausdorff distance between general semi-algebraic sets.

1.5.2 Margin Reduction and Instance Construction

To further understand the computational hardness, beyond exact decision, the proof introduces a technique called **margin reduction**.

The key idea is to create instances where the Hausdorff distance between the sets A and B can only take two extremely separated values:

- Either the distance is less than a given threshold t,
- Or the distance is at least $t \cdot 2^{2^n}$, where n is the number of variables involved.

This large separation guarantees that even approximating $d_H(A, B)$ within a reasonable factor becomes computationally infeasible, further reinforcing the problem's hardness.

It is important to note that while margin reduction strengthens the inapproximability result, the basic $\forall\exists_<\mathbb{R}$-hardness is already established independently by a more fundamental construction.

1.5.3 Computational Implications

The margin reduction technique leads to the following important corollary:

Corollary 1 *There is no polynomial-time algorithm that approximates the Hausdorff distance within a margin $f(n) = 2^{2^{o(n)}}$, unless $P = \forall\exists_<\mathbb{R}$.*

This shows that not only is the exact computation hard, but even rough approximations are impossible under standard complexity assumptions.

Moreover, the Hausdorff distance problem remains $\forall\exists_<\mathbb{R}$-complete even if the sets A and B are described by particularly simple algebraic formulas:

- Either by a conjunction of quadratic polynomial equations,
- Or by a single polynomial equation of degree at most four (quartic).

Thus, syntactic simplicity of the formulas does not make the problem computationally easier.

1.6 Proof Strategy of the Main Result

The proof strategy proceeds as follows:

- Start from an arbitrary instance of a $\forall\exists_<\mathbb{R}$ problem, given by a second-order logical formula:
$$\phi := \forall X \in \mathbb{R}^n, \exists Y \in \mathbb{R}^m : \psi(X, Y),$$
where ψ is a quantifier-free formula involving strict inequalities.
- Define two sets in $\mathbb{R}^{n+m}$:
  - A as the set of (x, y) pairs satisfying $\psi(x, y)$,
  - B as the entire domain $[-C, C]^n \times \{0\}^m$, where C is a sufficiently large constant.
- The critical link is:
$$d_H(A, B) \leq t \quad \text{if and only if} \quad \phi \text{ is true.}$$
If ϕ is true, A densely covers B, making the distance zero or very small. If ϕ is false, there is a region missing from A (an **open ball**), causing the Hausdorff distance to be strictly positive.
- To ensure that false instances of ϕ generate a real "gap" in A (rather than isolated points), the construction guarantees that the counterexample set contains an **open ball** of positive radius. This step is essential to ensure $d_H(A, B) > 0$ when ϕ is false.
- Additionally, preprocessing steps ensure that all formulas respect the **Strict-UETR** conditions:
  - All inequalities are strict $(<, >)$,
  - No explicit negations $(\neg)$ are used.

Thus, the construction successfully reduces any $\forall\exists_<\mathbb{R}$ instance to a Hausdorff distance decision problem between semi-algebraic sets. Consequently, the Hausdorff distance problem is shown to be $\forall\exists_<\mathbb{R}$-complete.

1.6.1 Conclusions

- For finite point sets, d_H is computable in $O(ab)$, or $O((a + b)\log(a + b))$ using Voronoi diagrams.
- For convex polygons or simplicial complexes, efficient algorithms exist.
- For semi-algebraic sets, computing d_H is $\forall\exists_<\mathbb{R}$-complete, and even approximation is infeasible unless major breakthroughs in complexity theory occur.

1.7 Application of the Hausdorff Distance in Image Processing

The Hausdorff distance is widely used in computer vision for tasks like shape comparison, object recognition, and image alignment [Karimi and Salcudean(2019)]. In the medical field, it evaluates discrepancies between geometric structures, particularly in segmentation validation.

1.7.1 Introduction to Image Processing

Image processing analyzes digital images to extract information. It ranges from low-level tasks like adjusting brightness to complex ones like object detection. We focus here on border detection and segmentation, key areas where the Hausdorff distance applies.

Border Detection. Edge detection identifies intensity changes that correspond to object boundaries [Torre and Poggio(1986)]. Techniques include Sobel, Prewitt, and Canny methods, essential for applications like medical imaging.

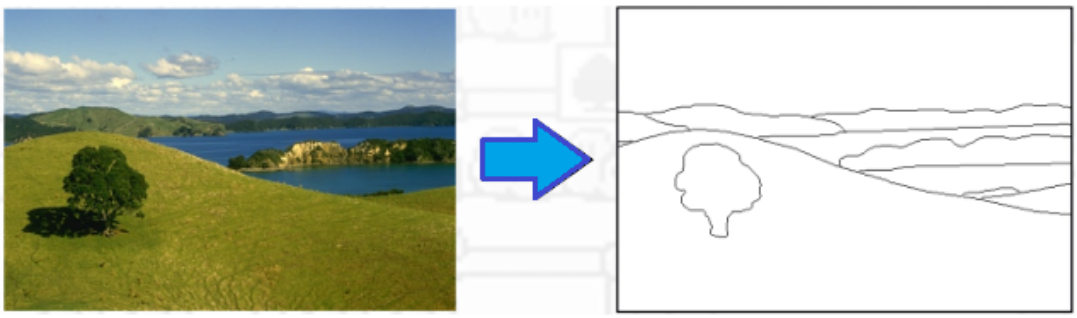


Figure 1.8: Example of edge detection.

Image Segmentation. Segmentation divides an image into meaningful regions, facilitating object detection [Haralick and Shapiro(1985)]. Methods include thresholding, clustering, and deep learning architectures like U-Net and Mask R-CNN.

Figure 1.9: Example of image segmentation.

1.7.2 The Hausdorff Distance in Image Processing

Shape comparison is critical in image matching, where the Hausdorff distance measures the proximity between two point sets without requiring exact point correspondences. While robust to small variations, it is highly sensitive to outliers [Zhao et al.(2005)Zhao, Shi, and Deng]. Modifications like the Modified Hausdorff Distance mitigate this issue, as discussed in Section 1.3.

Given two segmentations (ground truth and prediction sets X and Y), the Hausdorff distance quantifies their similarity [Karimi and Salcudean(2019)].

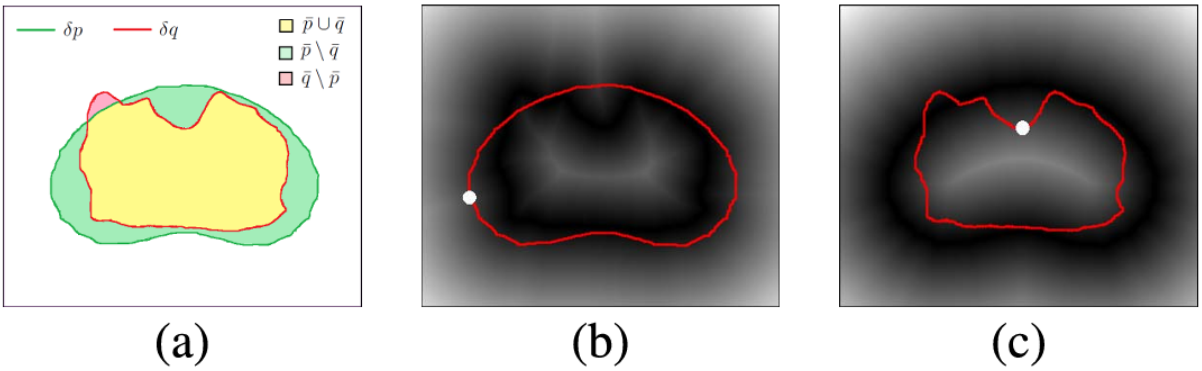


Figure 1.10: Ground-truth and predicted segmentations [Karimi and Salcudean(2019)].

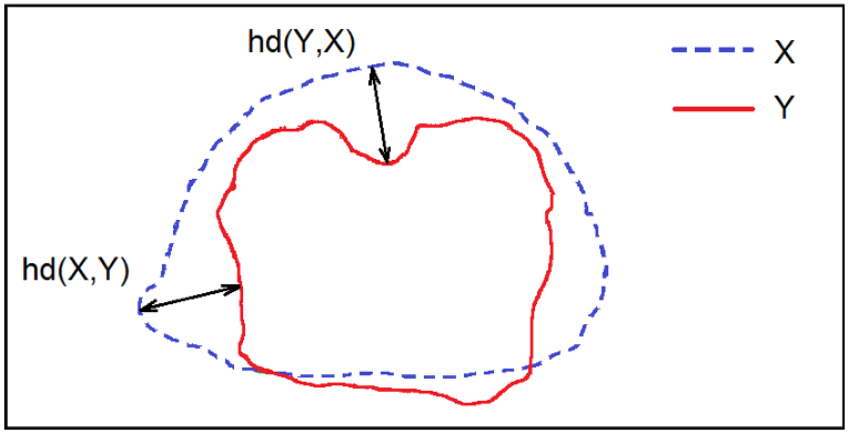


Figure 1.11: Schematic of Hausdorff distance between sets X and Y [Karimi and Salcudean(2019)].

Algorithms for Computing the Hausdorff Distance

Naive Algorithm. The NaiveHD algorithm computes all pairwise distances with complexity $\mathcal{O}(n \cdot m)$ [Chen et al.(2017)Chen, He, Wu, and Hou].

ITK Algorithm. The Insight Segmentation and Registration Toolkit (ITK) method improves efficiency using distance transforms and spatial indexing [Segmentation and (ITK)(n.d.)], achieving complexity $\mathcal{O}(N + n)$.

1.7.3 The Hausdorff Distance in Medical Imaging

An example from [El-Banby et al.(2024)El-Banby, Salem, Tafweek, and El-Azziz] uses the Hausdorff distance to validate a deep U-Net model for breast cancer detection in mammography.

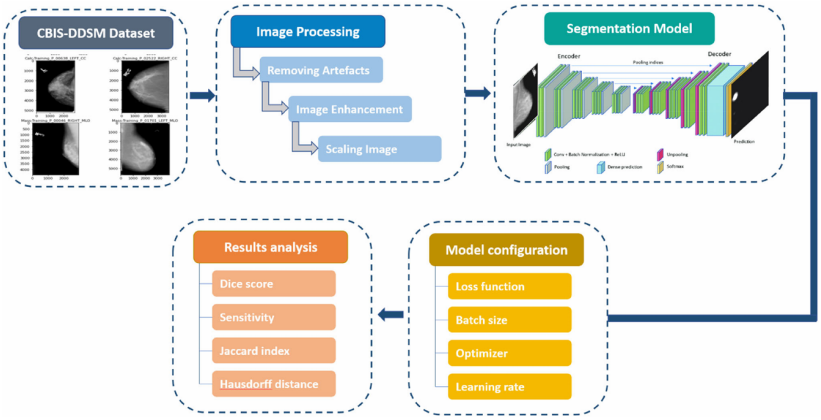


Figure 1.12: Segmentation process flow for abnormalities detection [El-Banby et al.(2024)El-Banby, Salem, Tafweek, and El-Azziz].

The model uses preprocessing (artifact removal, CLAHE, filtering), data augmentation, and a U-Net architecture for segmentation, evaluated on CBIS-DDSM and INbreast datasets.

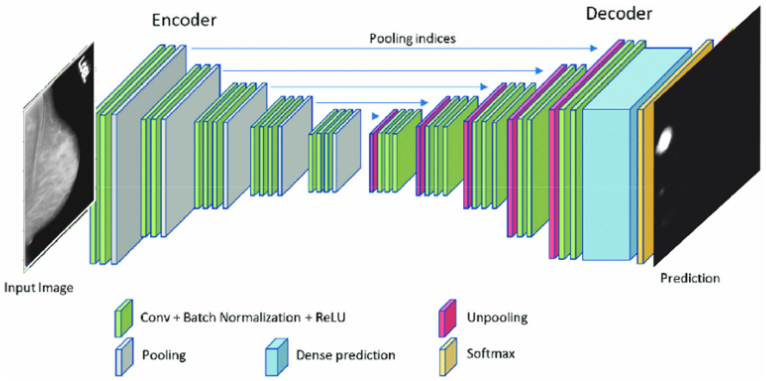


Figure 1.13: U-Net architecture [El-Banby et al.(2024)El-Banby, Salem, Tafweek, and El-Azziz].

Evaluation metrics include the Dice score, Jaccard coefficient, and the Hausdorff distance:

$$\text{Dice score} = \frac{2|\text{GT} \cap P|}{|\text{GT}| + |P|},$$

$$\text{Jaccard coefficient} = \frac{|\text{GT} \cap P|}{|\text{GT} \cup P|},$$

$$\text{Hausdorff distance} = \max\left(\sup_{p \in P} \inf_{gt \in \text{GT}} d(p, gt), \sup_{gt \in \text{GT}} \inf_{p \in P} d(gt, p) \right).$$

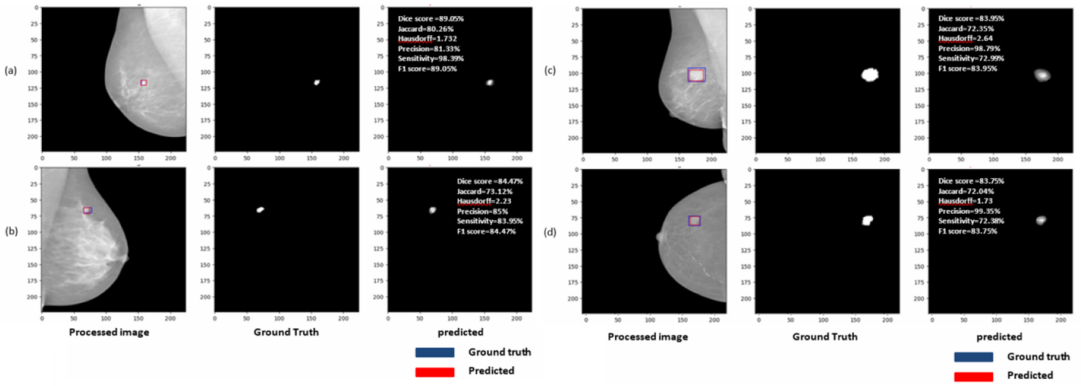


Figure 1.14: Results using Dice, Jaccard, and Hausdorff metrics [El-Banby et al.(2024)El-Banby, Salem, Tafweek, and El-Azziz].

The Hausdorff distance enables precise boundary evaluation, crucial for medical diagnostics where minor misalignments are significant.

1.7.4 Conclusion

The Hausdorff distance is fundamental for image comparison in medical imaging, offering valuable insight into boundary discrepancies. Despite its sensitivity to outliers, modifications have enhanced its robustness, allowing its use in various medical applications, from organ segmentation to model validation.

References

[Chen et al.(2017)Chen, He, Wu, and Hou] Yilin Chen, Fazhi He, Yiqi Wu, and Neng Hou. A local start search algorithm to compute exact hausdorff distance for arbitrary point sets. *Pattern Recognition*, 67:139–148, 2017.

[Dubuisson and Jain(1994)] M.-P. Dubuisson and A.K. Jain. A modified hausdorff distance for object matching. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 566–568 vol.1, 1994. doi: 10.1109/ICPR.1994.576361.

[El-Banby et al.(2024)El-Banby, Salem, Tafweek, and El-Azziz] Ghada M El-Banby, Nourhan S Salem, Eman A Tafweek, and Essam N Abd El-Azziz. Automated abnormalities detection in mammography using deep learning. *Complex & Intelligent Systems*, 10(5):7279–7295, 2024.

[Haralick and Shapiro(1985)] Robert M Haralick and Linda G Shapiro. Image segmentation techniques. *Computer vision, graphics, and image processing*, 29(1):100–132, 1985.

[Karimi and Salcudean(2019)] Davood Karimi and Septimiu E Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on medical imaging*, 39(2):499–513, 2019.

[Segmentation and (ITK)(n.d.)] Insight Segmentation and Registration Toolkit (ITK). Hausdorff distance image filter documentation, n.d. URL `https://itk.org/Doxygen312/html/classitk_1_1HausdorffDistanceImageFilter.html`.

[Torre and Poggio(1986)] Vincent Torre and Tomaso A Poggio. On edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):147–163, 1986.

[Zhao et al.(2005)Zhao, Shi, and Deng] Chunjiang Zhao, Wenkang Shi, and Yong Deng. A new hausdorff distance for image matching. *Pattern Recognition Letters*, 26(5):581–586, 2005.

On the Cahn-Hilliard-Navier-Stokes Equations and the Implicit-Explicit Schemes

Andreu Martorell Garcés [♭]

(♭) andreu.martorell@uv.es

1.1 Introduction

The isentropic compressible Cahn-Hilliard-Navier-Stokes (CHNS) equations model the dynamics of binary fluid mixtures, capturing both phase separation and fluid flow. These equations result in a system of fourth-order partial differential equations that are numerically challenging to solve due to the presence of high-order spatial derivatives.

The primary objective of this work is to study an efficient numerical scheme for solving initial-boundary value problems involving these equations, [9, 1]. Specifically, in [9] is explained a second-order Implicit-Explicit (IMEX) Runge-Kutta method, where the convective terms are treated explicitly, while the stiff, high-order terms implicitly. A key advantage of this scheme is that it only requires solving linear systems at each implicit stage, which can be done efficiently using a multigrid solver [9, 4]. Numerical experiments are performed to validate the accuracy and efficiency of the proposed method.

The numerical difficulties inherent in solving the CHNS equations stem from the stiffness introduced by the second- to fourth-order spatial derivatives. When discretized using finite differences, these terms result in matrices with large eigenvalues, leading to stiffness. In the context of solving a system of ordinary differential equations of the form $z' = f(z)$, where the Jacobian $f'(z)$ has eigenvalues with large negative real parts, stiffness occurs. In such cases, explicit time-stepping methods are constrained to very small time steps, typically proportional to $\frac{1}{|\lambda|}$, where λ denotes the eigenvalue with the largest absolute value. Implicit methods, however, allow for significantly larger time steps, making them more suitable for stiff problems.

In many practical situations, the right-hand side of the equation can be decomposed as $f(z) = g_e(z) - g_c(z)$, where $g_e(z)$ and $g_c(z)$ are expansive and contractive terms, respectively, and both are strictly convex functions. In such cases, an IMEX scheme that treats $g_c(z)$ implicitly and $g_e(z)$ explicitly provides a gradient-stable method, as is shown in [7].

The CHNS system naturally contains both diffusive (contractive) and anti-diffusive (expansive) components due to the Cahn-Hilliard equation and the diffusion produced by the Cauchy stress tensor. As such, it is advantageous to treat the stiff, contractive components implicitly and the remaining terms explicitly, specially if they do not significantly contribute to stiffness. This approach leads to stable and efficient time integration schemes well-suited for the complex dynamics modeled by the CHNS system.

The following work is structured as follows: in Section 1.2 the isentropic compressible CHNS are stated; in Section 1.3 implicit-explicit Runge–Kutta finite difference numerical schemes for

the two-dimensional equations are proposed; finally in Section 1.4 we perform some numerical experiments to assess the efficiency of the method.

1.2 Cahn-Hilliard-Navier-Stokes Equations

The models considered are based on [1]. Let c_i be the mass concentration of species $i = 1, 2$, and define $c = c_1 - c_2$, ρ as the mixture density, and $\boldsymbol{v}$ as the bulk velocity. The domain $\Omega \subset \mathbb{R}^3$ represents the fluid-filled region, and $\varepsilon > 0$ is a parameter related to the thickness of the diffuse interface.

The Ginzburg-Landau free energy in a subregion $V \subseteq \Omega$ is given by:

$$E(\rho, c) = \int_V \left(\rho f(\rho, c) + \frac{\varepsilon}{2} |\nabla c|^2 \right) dx,$$

where $f(\rho, c) = f_e(\rho) + \psi(c)$, and the double-well potential is defined by $\psi(c) = \frac{1}{4}(c^2 - 1)^2$.

The governing equations are the isentropic compressible Cahn-Hilliard-Navier-Stokes system with gravity:

$$\begin{cases} \rho_t + \operatorname{div}(\rho \boldsymbol{v}) = 0, \\ (\rho \boldsymbol{v})_t + \operatorname{div}(\rho \boldsymbol{v} \otimes \boldsymbol{v}) = \rho \boldsymbol{g} + \operatorname{div}(\mathbb{T}), \\ (\rho c)_t + \operatorname{div}(\rho c \boldsymbol{v}) = \Delta \mu. \end{cases}$$

Here, the stress tensor, $\mathbb{T}$, and the chemical potential, μ, are given, respectively, by:

$$\mathbb{T} = \nu(c)(\nabla \boldsymbol{v} + \nabla \boldsymbol{v}^T) + (\lambda(c) \operatorname{div}(\boldsymbol{v}) - p(\rho, c))\mathbb{I} + \frac{\varepsilon}{2} |\nabla c|^2 \mathbb{I} - \varepsilon (\nabla c \otimes \nabla c),$$

$$\mu = \psi'(c) - \frac{\varepsilon}{\rho} \Delta c.$$

The system is closed by the initial and boundary conditions:

$$(\rho, \boldsymbol{v}, c)|_{t=0} = (\rho_0, \boldsymbol{v}_0, c_0), \quad \boldsymbol{v}|_{\partial\Omega} = \nabla c \cdot \boldsymbol{n}|_{\partial\Omega} = \nabla \mu \cdot \boldsymbol{n}|_{\partial\Omega} = 0. \tag{1.1}$$

In [1] is proved that the isentropic compressible Cahn-Hilliard-Navier-Stokes equations with gravitation admit weak solutions, with renormalization of ρ in the sense of Di Perna and Lions, in any interval $[0, T]$ for $T > 0$, provided the following conditions are satisfied:

$$\gamma > \frac{3}{2}, \quad 0 \leq \rho_0 \in L^\gamma(\Omega), \quad \rho_0 |\boldsymbol{v}_0|^2 \in L^1(\Omega), \quad c_0 \in H^1(\Omega).$$

Therefore, consider ν, λ as constant parameters, $p(\rho) = C_p \rho^\gamma$ with $C_p > 0$ constant and $\gamma > \frac{3}{2}$. The system of equations is given by:

$$\begin{cases} \rho_t + \operatorname{div}(\rho \boldsymbol{v}) = 0, \\ (\rho \boldsymbol{v})_t + \operatorname{div}\big(\rho \boldsymbol{v} \otimes \boldsymbol{v} + p(\rho)\mathbb{I}\big) = \rho \boldsymbol{g} + (\nu + \lambda) \nabla \operatorname{div}(\boldsymbol{v}) \\ \qquad + \nu \Delta \boldsymbol{v} + \dfrac{\varepsilon}{2} \nabla |\nabla c|^2 - \varepsilon \operatorname{div}(\nabla c \otimes \nabla c), \\ (\rho c)_t + \operatorname{div}(\rho c \boldsymbol{v}) = \Delta \mu. \end{cases}$$

In the remainder of the present work, we restrict our analysis of this system to the two-dimensional case, following the approach provided by [9]. Assuming that the velocity field is denoted by $\boldsymbol{v} = (v_1, v_2)$, the governing equations in 2D become:

$$\begin{aligned}
\rho_t + (\rho v_1)_x + (\rho v_2)_y &= 0,\\
(\rho v_1)_t + (\rho v_1^2 + C_p\rho^\gamma)_x + (\rho v_1 v_2)_y &= \frac{\varepsilon}{2}(c_y^2 - c_x^2)_x - \varepsilon(c_x c_y)_y + \\
&\qquad \nu\Delta v_1 + (\nu+\lambda)((v_1)_{xx} + (v_2)_{xy}),\\
(\rho v_2)_t + (\rho v_2^2 + C_p\rho^\gamma)_y + (\rho v_1 v_2)_x &= \rho g + \frac{\varepsilon}{2}(c_x^2 - c_y^2)_y - \varepsilon(c_x c_y)_x + \\
&\qquad \nu\Delta v_2 + (\nu+\lambda)((v_1)_{xy} + (v_2)_{yy}),\\
(\rho c)_t + (\rho c v_1)_x + (\rho c v_2)_y &= \Delta\left(\psi'(c) - \frac{\varepsilon}{\rho}\Delta c\right).
\end{aligned} \tag{1.2}$$

1.3 Numerical Schemes

Numerical schemes for the Cahn-Hilliard equation can be found in [5, 6], while methods for the quasi-incompressible Cahn-Hilliard-Navier-Stokes are presented in [8, 11, 13].

Our purpose is to follow the approach provided in [9] for the two-dimensional case of the compressible isentropic CHNS equation. In [9] it is introduced a second order IMEX-RK scheme, where convective terms are treated explicitly, and only linear systems need to be solved.

1.3.1 Spatial Semidiscretization

This section presents finite difference numerical methods for the efficient approximation of solutions to the two-dimensional equations introduced in 1.2. Thus, consider $\Omega = (0,1)^2$ and the equispaced computational grid given by the M^2 nodes $\boldsymbol{x}_{i,j} = \left(\left(i-\frac{1}{2}\right)h, \left(j-\frac{1}{2}\right)h\right)$ for $i,j = 1,\cdots,M$, where $h = \frac{1}{M}$ and we denote by (x,y) the spatial variable $\boldsymbol{x}$.

We denote by

$$u = (\rho, \boldsymbol{m}, q), \quad \boldsymbol{m} = (m_1, m_2) = (\rho v_1, \rho v_2), \quad q = \rho c,$$

the vector of conserved variables and aim to approximate equation (1.2) by a spatial semi-discretization consisting of $4M^2$ ordinary differential equations

$$u'_{k,i,j}(t) = \mathcal{L}(U(t))_{k,i,j}, \quad k = 1,\cdots,4, \;\; i,j = 1,\cdots,M,$$

for $4M^2$ unknowns $u'_{k,i,j} \in \mathbb{R}^4$ which are approximations of $u_k(\boldsymbol{x}_{i,j}, t)$ and form the $4M^2$ (column) vector function $U(t)$ by using lexicographical order so that

$$U = \begin{bmatrix} \varrho \\ \varrho * V_1 \\ \varrho * V_2 \\ \varrho * C \end{bmatrix}, \quad (\varrho * S)_i = \varrho_i S_i,$$

$$\rho(\boldsymbol{x}_{i,j}, t) \approx \varrho_{M(i-1)+j}(t), \quad v_k(\boldsymbol{x}_{i,j}, t) \approx (V_k)_{M(i-1)+j}(t), \;\; k = 1,2, \quad c(\boldsymbol{x}_{i,j}, t) \approx C_{M(i-1)+j}(t).$$

For the sake of notation and simplicity, we adopt a slight abuse of notation, identifying, for example, $\varrho_{i,j} \equiv \varrho_{M(i-1)+j}$. We also use superscripts for M^2-block notation, e.g., $U^1 = \varrho$.

The nonzero terms in the spatial semidiscretization

$$\mathcal{L}(U) = \mathcal{C}(U(t)) + \mathcal{L}_1(U(t)) + \mathcal{L}_2(U(t)) + \mathcal{L}_3(U(t)) + \mathcal{L}_4(U(t)),$$

are for the convective terms:

$$\mathcal{C}(U)_{1,i,j} \approx -\left((\rho v_1)_x + (\rho v_2)_y\right)(x_{i,j},t), \qquad \mathcal{C}(U)_{2,i,j} \approx -\left((\rho v_1^2 + \rho^\gamma)_x + (\rho v_1 v_2)_y\right)(x_{i,j},t),$$

$$\mathcal{C}(U)_{3,i,j} \approx -\left((\rho v_1 v_2)_x + (\rho v_2^2 + \rho^\gamma)_y\right)(x_{i,j},t), \quad \mathcal{C}(U)_{4,i,j} \approx -\left((\rho c v_1)_x + (\rho c v_2)_y\right)(x_{i,j},t),$$

and for the diffusive terms:

$$\mathcal{L}_1(U)_{3,i,j} \approx \rho(x_{i,j},t)G,$$

$$\mathcal{L}_2(U)_{2,i,j} \approx \left(\frac{\varepsilon}{2}(c_y^2 - c_x^2)_x - \varepsilon(c_x c_y)_y\right)(x_{i,j},t),$$

$$\mathcal{L}_2(U)_{3,i,j} \approx \left(\frac{\varepsilon}{2}(c_x^2 - c_y^2)_y - \varepsilon(c_x c_y)_x\right)(x_{i,j},t).$$

$$\mathcal{L}_3(U)_{4,i,j} \approx \Delta\left(\psi'(c) - \frac{\varepsilon}{\rho}\Delta c\right)(x_{i,j},t),$$

$$\mathcal{L}_4(U)_{2,i,j} \approx \big(\nu((v_1)_{xx} + (v_1)_{yy}) + (\nu+\lambda)\big((v_1)_{xx} + (v_2)_{xy}\big)\big)(x_{i,j},t),$$

$$\mathcal{L}_4(U)_{3,i,j} \approx \big(\nu((v_2)_{xx} + (v_2)_{yy}) + (\nu+\lambda)\big((v_1)_{xy} + (v_2)_{yy}\big)\big)(x_{i,j},t).$$

Let us see how we approximate each of the above operators.

The convective term $\mathcal{C}$ is obtained through finite differences of numerical fluxes obtained by WENO5 reconstructions [2, 3] on Global Lax–Friedrichs flux splittings [12], which is fifth-order accurate for finite difference schemes, based on point values.

The approximation of the only nonzero term of the operator $\mathcal{L}_1$ is taken pointwise, i.e.,

$$\mathcal{L}_1(U)_{3,i,j} \approx \rho(x_{i,j},t)g.$$

The operator $\mathcal{L}_2$, which involves the derivatives of c in the conservation of momenta, is approximated using finite differences which are second-order accurate at interior points and first-order accurate at boundary points. Indeed, consider the finite difference operators for functions on $M \times M$ grids, to approximate first-order derivatives along the x dimension (analogously in the y-direction):

$$D_x^{1*} f_{i,j} = \begin{cases} \frac{f_{i,j}}{h}, & i = 1, \\ \frac{f_{i,j} - f_{i-1,j}}{h}, & 1 < i < M, \\ -\frac{f_{i-1,j}}{h}, & i = M, \end{cases} \quad D_x^1 f_{i,j} = \begin{cases} \frac{f_{i+1,j} - f_{i,j}}{h}, & i < M, \\ 0, & i = M, \end{cases}$$

$$D_x\, f_{i,j} = \begin{cases} \frac{f_{i+1,j} - f_{i-1,j}}{2h}, & 1 < i < M, \\ \frac{f_{i+1,j} - f_{i,j}}{h}, & i = 1, \\ \frac{f_{i,j} - f_{i-1,j}}{h}, & i = M, \end{cases} \qquad D_x^* f_{i,j} = \begin{cases} \frac{f_{i+1,j} - f_{i-1,j}}{2h}, & 1 < i < M, \\ \frac{f_{i+1,j} - f_{i,j}}{2h}, & i = 1, \\ \frac{f_{i,j} - f_{i-1,j}}{2h}, & i = M, \end{cases}$$

and the shift operator

$$S_x f_{i,j} = \begin{cases} f_{i+1,j}, & i < M, \\ 0, & i = M. \end{cases}$$

The properties of these operators are the following:

1. $D_x^{1*} f_{i,j}$ is a second-order accurate approximation for $f_x(x_{i-\frac{1}{2},j})$, when $f_{i,j} = f(x_{i,j})$ and $f \in C^3$, with $f(x_{0,j}) = f(x_{M,j}) = 0$. This operator is used to approximate pure double derivatives.

2. $D_x^1 f_{i,j}$ is a second-order accurate approximation for $f_x(x_{i+\frac{1}{2},j})$, when $f_{i,j} = f(x_{i,j})$ and $f \in C^3$, with $f_x(x_{M+\frac{1}{2},j}) = 0$. This operator is also used to approximate pure double derivatives. The two operators are related by $D_x^{1*} = -(D_x^1)^T$.

3. $D_x\, f_{i,j}$ is a second-order accurate approximation for $f_x(x_{i,j})$ for $1 < i, j < M$, and first-order accurate otherwise, assuming $f_{i,j} = f(x_{i,j})$ and $f \in C^3$.

4. $D_x^* f_{i,j}$ is a second-order accurate approximation for $f_x(x_{i,j})$ when $1 < i, j < M$ or $j = 1, M$, and first-order accurate otherwise, assuming $f_{i,j} = f(x_{i,j})$, $f \in C^3$, and $f(x_{i,\frac{1}{2}}) = f(x_{i,M+\frac{1}{2}}) = 0$.

We consider the above second-order accurate approximations at interior points $1 < i, j < M$, and first-order accurate at boundary points. Let

$$c_{i,j} = \frac{(\rho c)_{i,j}}{\rho_{i,j}} \approx c(x_{i,j}),$$

with boundary conditions on c taken into account as in equation (1.1). Then:

$$(c_x^2)_x(x_{i,j}) \approx \left(D_x^{1*}\left(D_x^1 C * D_x^1 C\right)\right)_{i,j}, \quad (c_y^2)_y(x_{i,j}) \approx \left(D_y^{1*}\left(D_y^1 C * D_y^1 C\right)\right)_{i,j},$$

$$(c_y^2)_x(x_{i,j}) \approx \left(D_x\left(D_y^* C * D_y^* C\right)\right)_{i,j}, \quad (c_x^2)_y(x_{i,j}) \approx \left(D_y\left(D_x^* C * D_x^* C\right)\right)_{i,j},$$

and

$$(c_x c_y)_x(x_{i,j}) \approx \frac{1}{2}\left(D_x^{1*}\left(D_x^1 C * \left(S_x D_y^* C + D_y^* C\right)\right)\right)_{i,j},$$
$$(c_y c_x)_y(x_{i,j}) \approx \frac{1}{2}\left(D_y^{1*}\left(D_y^1 C * \left(S_y D_x^* C + D_x^* C\right)\right)\right)_{i,j}.$$

On the other hand, the operator $\mathcal{L}_3$, which is related to the Cahn-Hilliard equation, needs a special treatment. This is because only negative definite terms should be treated implicitly, and the term $(\psi'(c))_{xx}$ changes sign in $(-1, 1)$. Therefore, due to Eyre [7], a splitting $\psi' = \phi_+ + \phi_-$ is considered in such a way that ϕ_+ is treated implicitly and ϕ_- explicitly. Both function are

$$\phi_+(c) = 2c, \quad \phi_-(c) = c^3 - 3c,$$

which verifies that

$$\phi_+'(c) = 2 > 0, \quad \phi_-'(c) = 3(c^2 - 1) \leq 0,$$

for every $c \in [-1, 1]$.

For approximating $\Delta(\phi_\pm(c))(x_{i,j}, t)$, we consider the matrix

$$\mathcal{M}_\pm(\mathcal{C}) = -(I_M \otimes D_1^T)\, D(\lambda_x) D(\lambda_\pm^x)(I_M \otimes D_1) - (D_1^T \otimes I_M)\, D(\lambda_y) D(\lambda_\pm^y)(D_1 \otimes I_M),$$

where I_M denotes the $M \times M$ identity matrix, and $\otimes$ is the Kronecker product, D is the diagonal operator on $M \times M$ matrices given by

$$(D(v)w)_{i,j} = v_{i,j}\, w_{i,j}, \text{ for } i, j = 1, \cdots, M,\ v, w \in \mathbb{R}^{M \times M},$$

$$D_1 = \frac{1}{h}\begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{M\times M},$$

and $\lambda_{\pm}^{x}$ and $\lambda_{\pm}^{y}$ is defined by the midpoints values of $\phi_{\pm}$ between adjacent grid cells in the x- and y-directions, respectively.

Thus, a second-order accurate approximation of the operator $\mathcal{L}_3$ is given by

$$\mathcal{L}_3(U)_{4,i,j} \approx \left(\mathcal{M}_+(C)C + \mathcal{M}_-(C)C - \varepsilon\Delta_h\left(D(\varrho)^{-1}\Delta_h C\right)\right)_{i,j},$$

where $\left(\mathcal{M}_{\pm}(C)C\right)_{i,j} \approx \Delta\left(\phi_{\pm}(c)\right)(x_{i,j},t)$ and Δ_h is the discrete laplacian operator.

Finally, it remains to approximate the operator $\mathcal{L}_4$, which involves the derivatives of $\boldsymbol{v}$ in the conservation of momenta. We use the following finite difference approximation for $(v_k)_{xx}, (v_k)_{yy}$:

$$w''(x_i) = \begin{cases} \dfrac{1}{h^2}\left(\dfrac{4}{3}w(x_{i+1}) - \dfrac{4}{3}w(x_i)\right), & i = 1, \\ \dfrac{1}{h^2}\left(w(x_{i+1}) - 2w(x_i) + w(x_{i-1})\right), & 1 < i < M, \\ \dfrac{1}{h^2}\left(-4w(x_i) + \dfrac{4}{3}w(x_{i-1})\right), & i = M, \end{cases}$$

for any $w \in \mathcal{C}^4$ such that $w(0) = w(1) = 0$. This approximation is second-order accurate for interior points $1 < i, j < M$, and first-order accurate at the boundaries under the no-slip boundary conditions for v_k, $k = 1, 2$.
These approximations lead to the $k = 2, 3$ blocks $\mathcal{L}_k^4(U)$ of $\mathcal{L}_4(U)$, expressed as:

$$\begin{bmatrix} \mathcal{L}_2^4(U) \\ \mathcal{L}_3^4(U) \end{bmatrix} = \begin{bmatrix} (2\nu+\lambda)I_M \otimes E + \nu E \otimes I_M & (\nu+\lambda)D \otimes D \\ (\nu+\lambda)D \otimes D & \nu I_M \otimes E + (2\nu+\lambda)E \otimes I_M \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix},$$

where matrices E and D are defined as follows:

$$E = \frac{1}{h^2}\begin{bmatrix} -\frac{4}{3} & \frac{4}{3} & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -2 & 1 \\ 0 & \cdots & 0 & \frac{4}{3} & -\frac{4}{3} \end{bmatrix}, \quad D = \frac{1}{h}\begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix},$$

which fail to be symmetric due to the boundary conditions.

1.3.2 IMEX schemes

For obtaining a Linearly Implicit Explicit schemes we use the technique of doubling variables and a partitioned Runge-Kutta schemes [9, 4, 10]. Consider the operator

$$\tilde{\mathcal{L}}(\tilde{U}, U) = \mathcal{C}(\tilde{U}) + \mathcal{L}_1(U) + \mathcal{L}_2(U) + \tilde{\mathcal{L}}_3(\tilde{U}, U) + \mathcal{L}_4(U),$$

where $\tilde{U}$ is treated explicitly and U implicitly and

$$\tilde{\mathcal{L}}_3(\tilde{U}, U)_{4,i,j} = \left(\mathcal{M}_+(\tilde{C})C + \mathcal{M}_-(\tilde{C})\tilde{C} - \varepsilon\Delta_h\left(D(\varrho)^{-1}\Delta_h C\right)\right)_{i,j},$$

We have that the initial value problem

$$\begin{aligned} U' &= \mathcal{L}(U), \\ U(0) &= U_0, \end{aligned} \tag{1.3}$$

is equivalent to

$$\tilde{U}' = \tilde{\mathcal{L}}(\tilde{U}, U),$$

$$U' = \tilde{\mathcal{L}}(\tilde{U}, U),$$

$$\tilde{U}(0) = U(0) = U_0.$$

A partitioned Runge-Kutta scheme, in which there are two different s stages Butcher tableaus, one explicit and one (diagonally) implicit

$$\begin{array}{c|c} \tilde{\delta} & \tilde{\alpha} \\ \hline & \tilde{\beta}^T \end{array}, \quad \tilde{\alpha}_{i,j} = 0, \quad j \geq i, \qquad \begin{array}{c|c} \delta & \alpha \\ \hline & \beta^T \end{array}, \quad \alpha_{i,j} = 0, \quad j > i,$$

can be applied to (1.3). It can be seen that if both Butcher tableaus yield second order accurate schemes and $\beta = \tilde{\beta}$, then the resulting partitioned Runge-Kutta scheme is second-order accurate [10]. Therefore, making this assumption, it is shown that there is no need of doubling variables resulting in the recursion

$$\tilde{U}^{(i)} = \tilde{U}^n + \Delta t \sum_{j<i} \tilde{\alpha}_{i,j} \tilde{\mathcal{L}}(\tilde{U}^{(j)}, U^{(j)}),$$

$$U^{(i)} = U^n + \Delta t \sum_{j<i} \alpha_{i,j} \tilde{\mathcal{L}}(\tilde{U}^{(j)}, U^{(j)}) + \Delta t \alpha_{i,i} \tilde{\mathcal{L}}(\tilde{U}^{(i)}, U^{(i)}),$$

$$U^{n+1} = U^n + \Delta t \sum_{j=1}^{s} \beta_j \tilde{\mathcal{L}}(\tilde{U}^{(j)}, U^{(j)}),$$

for $i = 1, ..., s$.

1.3.3 System to solve

At each stage, one needs to solve

$$U^{(i)} = U^n + \Delta t \sum_{j<i} \alpha_{i,j} \tilde{\mathcal{L}}(\tilde{U}^{(j)}, U^{(j)}) + \Delta t \alpha_{i,i} \tilde{\mathcal{L}}(\tilde{U}^{(i)}, U^{(i)}),$$

for $U^{(i)}$, where

$$U^n = \begin{bmatrix} \varrho^n \\ M_1^n \\ M_2^n \\ \mathcal{Q}^n \end{bmatrix}, \qquad U^{(i)} = \begin{bmatrix} \varrho^{(i)} \\ M_1^{(i)} \\ M_2^{(i)} \\ \mathcal{Q}^{(i)} \end{bmatrix} = \begin{bmatrix} \varrho^{(i)} \\ \varrho^{(i)} * V_1^{(i)} \\ \varrho^{(i)} * V_2^{(i)} \\ \varrho^{(i)} * C^{(i)} \end{bmatrix}.$$

As we shall see, although $\mathcal{L}_2, \tilde{\mathcal{L}}_3, \mathcal{L}_4$ are not linear, only linear systems have to be solved.

Notice that $\varrho^{(i)}$ is explicitly computable. Indeed, with block superscript notation for the operators and $\tilde{\mathcal{L}}$ variables, we get

$$\varrho^{(i)} = \varrho^n + \Delta t \sum_{j<i} \alpha_{i,j} \tilde{\mathcal{L}}_j^1 + \Delta t \alpha_{i,i} \tilde{C}^1(\tilde{U}^{(i)}).$$

For the fourth variable $\mathcal{Q}^{(i)}$, since $\varrho^{(i)}$ is already computed, this can be cast for the $C^{(i)}$ variables, resulting in the following linear system

$$\left(D(\varrho^{(i)}) - \Delta t \alpha_{i,i} \mathcal{M}_+(\tilde{C}^{(i)}) + \Delta t \alpha_{i,i} \varepsilon \Delta_h D(\varrho)^{-1} \Delta_h\right) C^{(i)} =$$

$$= \mathcal{Q}^n + \Delta t \sum_{j<i} \alpha_{i,j} \mathcal{K}_j^4 + \Delta t \alpha_{i,i} \left(\tilde{C}^4(\tilde{U}^{(i)}) + \mathcal{M}_-(\tilde{C}^{(i)}) \tilde{C}^{(i)}\right).$$

If $\varrho_i^k > 0$ for all k, then the matrix of this system is symmetric and positive definite, since it is the sum of a diagonal positive matrix and two symmetric and positive semidefinite matrices.

Finally, since $\varrho^{(i)}$ and $C^{(i)}$ are known, then $M_1^{(i)}, M_2^{(i)}$ can be computed by solving a linear system for $V_1^{(i)}$ and $V_2^{(i)}$, which is

$$\left(\begin{bmatrix} D(\varrho^{(i)}) & 0 \\ 0 & D(\varrho^{(i)}) \end{bmatrix} \right.$$

$$\left. - \Delta t \alpha_{i,i} \begin{bmatrix} (2\nu+\lambda) I_M \otimes E + \nu E \otimes I_M & (\nu+\lambda) D \otimes D \\ (\nu+\lambda) D \otimes D & \nu I_M \otimes E + (2\nu+\lambda) E \otimes I_M \end{bmatrix} \right) \begin{bmatrix} V_1^{(i)} \\ V_2^{(i)} \end{bmatrix}$$

$$= \begin{bmatrix} M_1^n \\ M_2^n \end{bmatrix} + \Delta t \sum_{j<i} \alpha_{i,j} \begin{bmatrix} \mathcal{K}_j^2 \\ \mathcal{K}_j^3 \end{bmatrix} + \Delta t \alpha_{i,i} \begin{bmatrix} \mathcal{C}^2(\tilde{U}^{(i)}) + \mathcal{L}_1^2(\tilde{U}^{(i)}) + \mathcal{L}_2^2(\tilde{U}^{(i)}) \\ \mathcal{C}^3(\tilde{U}^{(i)}) + \mathcal{L}_1^3(\tilde{U}^{(i)}) + \mathcal{L}_2^3(\tilde{U}^{(i)}) \end{bmatrix},$$

If $\varrho_i^k > 0$ for all k, then the matrix of this system should be close to symmetric and positive definite, since the matrix

$$- \begin{bmatrix} (2\nu+\lambda) I_M \otimes E + \nu E \otimes I_M & (\nu+\lambda) D \otimes D \\ (\nu+\lambda) D \otimes D & \nu I_M \otimes E + (2\nu+\lambda) E \otimes I_M \end{bmatrix}$$

is the discretization of the self-adjoint elliptic operator

$$-\left((\nu+\lambda) \nabla \mathrm{div} \boldsymbol{v} + \nu \Delta \boldsymbol{v}\right),$$

under the boundary conditions (1.1).

1.3.4 Time-Step Selection

The time step taken is

$$\Delta t = \mathrm{CFL}^* \cdot \frac{\Delta x}{cs},$$

with CFL* is some number and the maximun of characteristic speeds, cs, is computed as

$$c_s = \max\left\{ \left|V_{k,j}^{(i)}\right| + \sqrt{C_p\gamma\left(\varrho_j^{(i)}\right)^{\gamma-1}} : i = 1,\dots,s,\ \ k = 1,2,\ \ j = 1,\dots,M^2 \right\}. \tag{1.4}$$

The scheme is not ensured to be bound preserving, that is, it might happen that the density become negative or the c-variable be outside $[-1,1]$. Purely convective models can lead to the formation of vacuum regions, which poses significant numerical challenges.

To address this, we adopt an adaptive time-stepping strategy: the time step Δt is reduced by 0.5 whenever $|c|$ exceeds a predefined threshold (1.5 in our experiments), and gradually increased to a prescribed maximum when the solution remains within acceptable bounds. Although, there is not a proof that this strategy always succeeds, the simulations $|c|$ has been always below the threshold, at the expense of local reductions of the parameter CFL* [9].

1.4 Numerical experiments

The objective of the experiments in this section are:

1. Showing that the order of the global errors in some experiments coincides with the expected design order of the scheme used to obtain them.

2. Showing that some IMEX schemes can perform time steps Δt with the same stability restrictions as the purely convective subsystem.

3. Testing the behaviour of different issues for the alogrithms, such as conservation.

For all numerical experiments, the adiabatic constant γ is set to $5/3$, the constant C_p to 1, the gravity $g = -10$ and CFL= 0.4 with Δt provided by (1.4). In addition, the following second order Butcher tableaus,

$$\text{*-DIRKSA}\quad \begin{array}{c|cc} 0 & 0 & 0 \\ 1+s & 1+s & 0 \\ \hline & s & 1-s \end{array}, \qquad \begin{array}{c|cc} 1-s & 1-s & 0 \\ 1 & s & 1-s \\ \hline & s & 1-s \end{array}, \quad s = \frac{1}{\sqrt{2}},$$

are considered.

1.4.1 Order test

This test is designed to verify that the *-DIRKSA method attains second-order accuracy in terms of the global errors. To this end, a forcing term is added to the equations so that the solution is prescribed. In particular, the solution in this case is

$$\begin{aligned}
\rho^*(x,y,t) &= \frac{\cos(2\pi x)\cos(\pi y)(t+1)}{10} + \frac{5}{4}, \\
v_1^*(x,y,t) &= -\sin(\pi x)\sin(\pi y)\left(2t^2-1\right), \\
v_2^*(x,y,t) &= \sin(\pi x)\sin(2\pi y)\left(t^2+1\right), \\
c^*(x,y,t) &= \frac{3}{4} - \frac{\cos(\pi x)\cos(\pi y)(t-1)}{10}.
\end{aligned}$$

Notice that the functions satisfy the boundary conditions. The equations to be solved are:

$$\rho_t + (\rho v_1)_x + (\rho v_2)_y = \rho^*_t + (\rho^* v^*_1)_x + (\rho^* v^*_2)_y,$$

$$\begin{aligned}(\rho v_1)_t + (\rho v_1^2 + C_p\rho^\gamma)_x + (\rho v_1 v_2)_y &= \frac{\varepsilon}{2}(c_y^2 - c_x^2)_x - \varepsilon(c_x c_y)_y + \nu\Delta v_1 \\ &+ (\nu+\lambda)((v_1)_{xx} + (v_2)_{xy}) + (\rho^* v^*_1)_t + (\rho^*(v^*_1)^2 + C_p(\rho^*)^\gamma)_x + (\rho^* v^*_1 v^*_2)_y \\ &- \frac{\varepsilon}{2}((c^*_y)^2 - (c^*_x)^2)_x + \varepsilon(c^*_x c^*_y)_y - \nu\Delta v^*_1 - (\nu+\lambda)((v^*_1)_{xx} + (v^*_2)_{xy}),\end{aligned}$$

$$\begin{aligned}(\rho v_2)_t + (\rho v_1 v_2)_x + (\rho v_2^2 + C_p\rho^\gamma)_y &= \rho g + \frac{\varepsilon}{2}(c_x^2 - c_y^2)_y - \varepsilon(c_x c_y)_x + \nu\Delta v_2 \\ &+ (\nu+\lambda)((v_1)_{xy} + (v_2)_{yy}) + (\rho^* v^*_2)_t + (\rho^* v^*_1 v^*_2)_x + (\rho^*(v^*_2)^2 + C_p(\rho^*)^\gamma)_y - \rho^* g \\ &- \frac{\varepsilon}{2}\left((c^*_x)^2 - (c^*_y)^2\right)_y + \varepsilon(c^*_x c^*_y)_x - \nu\Delta v^*_2 - (\nu+\lambda)((v^*_1)_{xy} + (v^*_2)_{yy}),\end{aligned}$$

$$\begin{aligned}(\rho c)_t + (\rho c v_1)_x + (\rho c v_2)_y &= \Delta\left(\psi(c) - \frac{\varepsilon}{\rho}\Delta c\right) + (\rho^* c^*)_t + (\rho^* c^* v^*_1)_x \\ &+ (\rho^* c^* v^*_2)_y - \Delta\left(\psi(c^*) - \frac{\varepsilon}{\rho^*}\Delta c^*\right).\end{aligned}$$

The parameters that have been used for this test are: $\nu = 01$, $\lambda = 0.1$ and $\varepsilon = 10^{-4}$.

For $M \times M$ grids with $M = 2^\ell$, $\ell = 3, \ldots, 8$, the global errors of the approximations $u^n_{k,i,j}$ obtained using the *-DIRKSA method at time $t^n = T = 0.01$ are computed as

$$e_M = \frac{1}{M^2}\sum_{k=1}^{4}\sum_{i,j=1}^{M}\left|u^n_{k,i,j} - u_k(x_{i,j}, T)\right|,$$

and are reported in Table 1.1. The observed experimental order of convergence,

$$\mathrm{EOC}_M = \log_2\left(\frac{e_M}{e_{2M}}\right),$$

approaches 2, confirming the second-order accuracy of the scheme.

M	e_M	EOC_M
8	1.9828e-02	2.21
16	4.2964e-03	2.04
32	1.0422e-03	1.97
64	2.6617e-04	1.97
128	6.7802e-05	1.98
256	1.7148e-05	–

Table 1.1: Computed orders of convergence of global errors.

1.4.2 Test of total mass conservation

In this test, we show that the current scheme preserve the total mass, that is,

$$\int_\Omega \rho(x,t)\ dx = \int_\Omega \rho(x,0)\ dx, \text{ and } \int_\Omega q(x,t)\ dx = \int_\Omega q(x,0)\ dx, \tag{1.5}$$

for almost all $t \in (0, T)$.

In order to check it, we consider $\rho_0 = 1$, $v_0 = 0$, and c_0 as a uniform random sample of zero mean and 10^{-10} standard deviation. We approximate (1.5) by computing

$$\int_\Omega \rho(x, t_n)\ dx - \int_\Omega \rho(x, 0)\ dx \approx \mathrm{err}_\rho(t_n) = \sum_{i,j=1}^{M} \varrho_{i,j}^n - \sum_{i,j=1}^{M} \varrho_{i,j}^0,$$

$$\int_\Omega q(x, t_n)\ dx - \int_\Omega q(x, 0)\ dx \approx \mathrm{err}_q(t_n) = \sum_{i,j=1}^{M} \mathcal{Q}_{i,j}^n - \sum_{i,j=1}^{M} \mathcal{Q}_{i,j}^0.$$

In Figure 1.1 the results are shown for parameters $\nu = 10^{-3}$, $\lambda = \varepsilon = 10^{-4}$, $M = 256$.

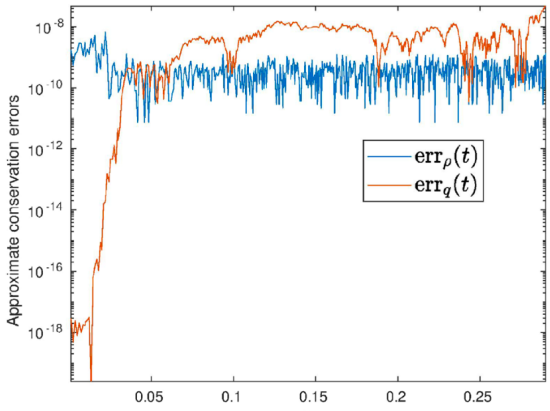


Figure 1.1: Mass conservation errors.

References

[1] Abels, H., Feireisl, E., *On a diffuse interface model for a two-phase flow of compressible viscous fluids*, Indiana Univ. Math. J. **57**(2), (2008), pp. 659–698.

[2] Baeza, A., Bürger, R., Mulet, P., Zorio, D., *On the efficient computation of smoothness indicators for a class of WENO reconstructions*, Journal of Scientific Computing, **80**(2), 1240–1263 (2019).

[3] Baeza, A., Bürger, R., Mulet, P., Zorio, D., *WENO reconstructions of unconditionally optimal high order*, SIAM Journal on Numerical Analysis, **57**(6), 2760–2784 (2019).

[4] Boscarino, S., Bürger, R., Mulet, P., Russo, G., Villada, L. M., *Linearly implicit IMEX Runge–Kutta methods for a class of degenerate convection–diffusion problems*, SIAM J. Sci. Comput., **37**(2), B305–B331 (2015).

[5] Elliott, C. M., *The Cahn–Hilliard model for the kinetics of phase separation*. In *Mathematical Models for Phase Change Problems (Óbidos, 1988)*, Internat. Ser. Numer. Math., vol. 88, pp. 35–73. Birkhäuser, Basel, 1989.

[6] Elliott, C. M., French, D. A., *Numerical studies of the Cahn–Hilliard equation for phase separation*, IMA Journal of Applied Mathematics, **38**(2), 97–128 (1987).

[7] Eyre, D. J., *Unconditionally gradient stable time marching the Cahn–Hilliard equation*, MRS Proceedings, **529**, 39–46 (1998).

[8] Jacqmin, D., *Calculation of two-phase Navier–Stokes flows using phase-field modeling*, Journal of Computational Physics, **155**(1), 96–127 (1999).

[9] Mulet, P., *Implicit-Explicit Schemes for Compressible Cahn–Hilliard–Navier–Stokes Equations*, Journal of Scientific Computing, **101**(2), 36 (2024).

[10] Pareschi, L., Russo, G., *Implicit–explicit Runge–Kutta schemes and applications to hyperbolic systems with relaxation*, J. Sci. Comput., **25**(1/2), 129–155 (2005).

[11] Shen, J., Yang, X., *Numerical Approximations of Allen–Cahn and Cahn–Hilliard Equations*, Discrete and Continuous Dynamical Systems, **28**(4), 1669–1691 (2010).

[12] Shu, C.-W., *High order weighted essentially nonoscillatory schemes for convection dominated problems*, SIAM Review, **51**(1), 82–126 (2009).

[13] Yue, P. T., Feng, J. J., Liu, C., Shen, J., *A diffuse-interface method for simulating two-phase flows of complex fluids*, Journal of Fluid Mechanics, **515**, 293–317 (2004).

Modelos SEIR: Estimación de parámetros mediante algoritmos genéticos para la toma de decisiones en salud pública

Cortes, José Julián [♭], Salamanca, Brian Smith 4[♭]

(♭) jjcormuo@posgrado.upv.es
(♮) bsaldur@posgrado.upv.es

1.1. Introduction

Dengue, a vector-borne disease, presents a significant incidence in tropical and subtropical regions, posing a global public health challenge [23]. Mathematical modeling is employed to understand its complex transmission dynamics—involving interactions among humans, *Aedes* mosquito vectors, and environmental factors—and to inform the design of public health policies [14]. This approach allows for the quantitative exploration of propagation mechanisms and the assessment of the potential impact of various intervention strategies.

Initial compartmental epidemiological models, such as the SIR (Susceptible-Infected-Recovered) model, established a fundamental framework for analyzing disease transmission. A primary limitation of this model is that it does not incorporate the incubation period; that is, the latent phase between exposure to the pathogen and the point at which an individual becomes infectious. To explicitly represent this latency phase, SEIR (Susceptible-Exposed-Infected-Recovered) models were subsequently developed. These models include an additional 'Exposed' (E) compartment for individuals who are infected but not yet infectious, a relevant feature for describing diseases with a significant incubation period, such as dengue [17].

In vector-borne diseases like dengue, SEIR models are often extended to SEIR-SEI structures, explicitly differentiating the human population (SEIR) from the vector population (SEI) [4]. This structure allows for modeling bidirectional transmission: from infected vectors to susceptible humans and from infected humans to susceptible vectors. Variations exist, such as SIR-SI models [19] or simple SEIR models [15]. The choice among these depends on the study objectives and data availability, reflecting a trade-off between biological realism and model complexity.

The practical application of these SEIR-SEI models fundamentally depends on the accuracy of their parameters (e.g., transmission rates, infectious periods, development rates). Direct measurement of many of these parameters is complex; therefore, their estimation from observed epidemiological data (calibration) is essential [18]. This process faces challenges such as data quality (e.g., noise, underreporting) and parameter identifiability. Ensuring that parameters can be uniquely and reliably estimated from available data constitutes a significant methodological problem [16].

To address the parameter estimation problem in complex models, metaheuristics such as Genetic Algorithms (GA) [7] and Particle Swarm Optimization (PSO) are employed. These global search methods efficiently explore high-dimensional parameter spaces without requiring derivatives and exhibit robustness against local optima. They have been applied to estimate parameters in *Aedes* vector population models [10], calculate the basic reproduction number (R_0) for dengue [24, 15], and in general calibration studies of SEIR/SEI models [2, 21]. Furthermore, GAs have been used in problems related to optimal control [5, 6].

Dengue dynamics are influenced by climatic variables, primarily temperature and precipitation, which affect vector biology and viral replication [11]. Integrating these factors into the models is necessary to capture observed seasonality. This is achieved through functional dependencies of parameters on climate [8, 9] or through statistical models linking climate and incidence [3, 22]. The influence of phenomena such as the El Niño-Southern Oscillation (ENSO) is particularly relevant in regions like Latin America [12, 1]. Climate models project potential changes in dengue incidence due to global warming [3], also highlighting the role of extreme events [13].

In this context, the present study aims to achieve the following specific objectives: first, to apply an adapted SEIR-SEI model to simulate the dynamics of dengue transmission in a specific urban environment; second, to employ genetic algorithms to estimate key model parameters by fitting the model to observed historical data; third, to focus the analysis on the municipality of Villavicencio (Meta, Colombia), a region characterized by its endemicity [12, 13]; and fourth, to explore the implications of the calibrated model's results for planning public health interventions aimed at mitigating the impact of dengue.

1.2. Model Formulation and Methodology

SEIR-SEI Model Structure

Compartmental models are standard tools in mathematical epidemiology. The basic SIR (Susceptible-Infected-Recovered) model assumes an immediate transition from the susceptible to the infectious state. An extension of this model is the SEIR (Susceptible-Exposed-Infected-Recovered) model, which incorporates an ."Exposed"(E) compartment to represent the latent incubation period, during which an infected individual is not yet infectious. This approach is applicable to diseases such as dengue [17].

In the specific case of vector-borne diseases like dengue, where the dynamics of the *Aedes aegypti* mosquito are fundamental, an SEIR-SEI structure is commonly used [4]. This structure combines an SEIR model for the human population (H) with an SEI model for the vector population (V). The human compartments include Susceptible (S_H), Exposed (E_H), Infectious (I_H), and Recovered (R_H). The vector compartments encompass Susceptible (S_V), Exposed (E_V), and Infectious (I_V). It is assumed that infected vectors remain in that state for their entire lives. This model allows for representing bidirectional transmission, where susceptible humans are infected by infectious vectors (I_V), and susceptible vectors become infected by biting infectious humans (I_H).

The system of ordinary differential equations (ODEs) describing the transmission dynamics in

the human population in this study is:

$$\begin{aligned}
\frac{dS_H}{dt} &= \mu_H N_H - \beta_{VH}(t)\,\frac{S_H\, I_V}{N_V} - \mu_H S_H,\\
\frac{dE_H}{dt} &= \beta_{VH}(t)\,\frac{S_H\, I_V}{N_V} - (\sigma_H + \mu_H)E_H,\\
\frac{dI_H}{dt} &= \sigma_H E_H - (\gamma_H + \mu_H)I_H,\\
\frac{dR_H}{dt} &= \gamma_H I_H - \mu_H R_H.
\end{aligned}$$

Here, $N_H = S_H + E_H + I_H + R_H$ is the total human population, considered constant or slowly varying due to the natural birth and death rate μ_H. The term $\beta_{VH}(t)$ is the effective time-dependent transmission rate from vector to human, N_V is the total vector population (considered constant or calculated), σ_H is the rate at which exposed humans become infectious (inverse of the intrinsic incubation period), and γ_H is the recovery rate of infectious humans.

For the vector population, the dynamics are represented by:

$$\begin{aligned}
\frac{dS_V}{dt} &= \Lambda_V - \beta_{HV}(t)\,\frac{S_V I_H}{N_H} - \mu_V S_V,\\
\frac{dE_V}{dt} &= \beta_{HV}(t)\,\frac{S_V I_H}{N_H} - (\sigma_V + \mu_V)E_V,\\
\frac{dI_V}{dt} &= \sigma_V E_V - \mu_V I_V.
\end{aligned}$$

In this system, Λ_V is the recruitment rate of new susceptible vectors (birth rate), $\beta_{HV}(t)$ is the effective time-dependent transmission rate from human to vector, μ_V is the natural mortality rate of the vector, and σ_V is the rate at which exposed vectors become infectious (inverse of the extrinsic incubation period, EIP). The total vector population is given by $N_V = S_V + E_V + I_V$.

Data Acquisition and Study Area

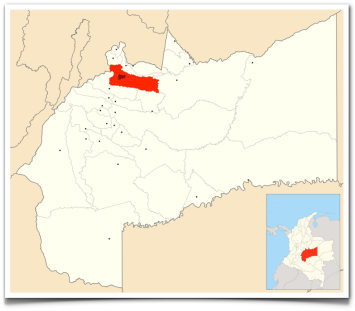


Figura 1.1: Map of the Department of Meta (Colombia), highlighting the location of the municipality of Villavicencio (in red). The inset shows the location of the department within Colombia.

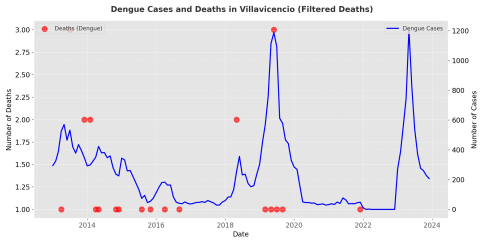
The data used in this study come from official Colombian sources for the period 2010-2022. Epidemiological data, consisting of monthly records of dengue cases and associated mortality, were obtained from the National Institute of Health (INS). Demographic data, corresponding to annual population projections, come from the National Administrative Department of Statistics

(DANE). Climatic variables, including monthly measurements of mean temperature and cumulative precipitation, were supplied by the Institute of Hydrology, Meteorology and Environmental Studies (IDEAM).

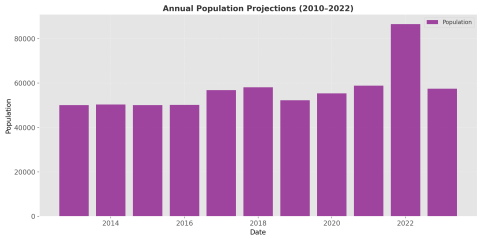
Cuadro 1.1: Pearson correlations between monthly climatic variables and reported dengue cases for several Colombian municipalities (period 2010–2022, n = number of months with complete data).

Municipality	**r(Temp., Cases)**	**r(Precip., Cases)**	**n**
Villavicencio (Meta)	0.65	0.42	122
Espinal (Tolima)	0.62	0.50	105
Sincelejo (Sucre)	0.40	0.33	98
Leticia (Amazonas)	0.60	0.45	110
Girardot (Cundinamarca)	0.57	0.38	99

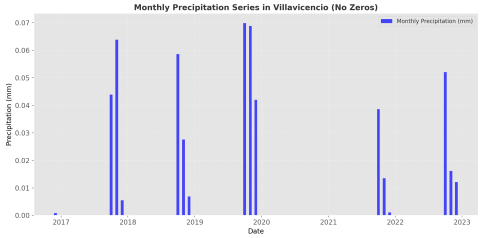
The municipality of Villavicencio, capital of the department of Meta (see Figure 1.1), was selected as the case study. This choice was based on the availability of relatively complete and consistent epidemiological, demographic, and climatic records for the period of interest. Furthermore, Villavicencio exhibits persistent reports of dengue and an observed correlation between case incidence and climatic variables, particularly temperature, as shown in Table 1.1, making it a suitable site for calibrating an SEIR-SEI model with climate forcing.



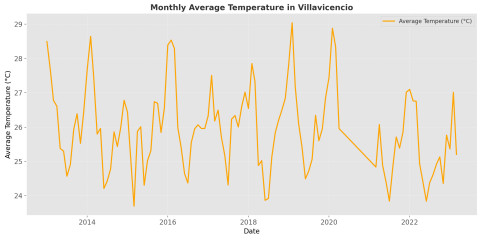
(a) Dengue Cases and Deaths



(b) Annual Population Projections



(c) Monthly Precipitation



(d) Mean Monthly Temperature

Figura 1.2: Overview of the data for Villavicencio: (a) Dengue cases and deaths (left axis: deaths, right axis: cases). (b) Annual population projections. (c) Monthly precipitation (mm). (d) Mean monthly temperature (°C). Period approx. 2010-2022.

Figure 1.2 presents an overview of the key data for Villavicencio used in this study: (a) the time series of monthly dengue cases and reported deaths, (b) annual population projections, (c) cumulative monthly precipitation, and (d) mean monthly temperature. These data form the basis for model calibration and the assessment of climatic influence.

Parameter Estimation using Genetic Algorithms

As mentioned in the Introduction, model calibration, i.e., the estimation of unknown parameters from observed data, is an essential step [18]. In this work, Genetic Algorithms (GA) [7], a class of metaheuristics inspired by natural evolution, are used to perform this task. GAs are suitable for optimizing parameters in complex and nonlinear systems like SEIR-SEI models [2].

The basic idea of a GA involves maintaining a population of candidate solutions (sets of parameters) and iteratively applying genetic operators to improve the population's fitness. The fitness of each solution is measured by an objective function that quantifies the discrepancy between the model output (using those parameters) and the actual observed data. The general steps are:

Initial Definitions: An initial population $P^{(0)} = \{x_1^{(0)}, x_2^{(0)}, \ldots, x_N^{(0)}\}$ of N individuals is defined, where each x_i is a vector of parameters to be estimated. A fitness function $f : x \mapsto \mathbb{R}$ is defined such that $f(x_i^{(t)}) \geq 0$, where higher (or lower, depending on the formulation) values indicate a better fit.

Step 1: Evaluation: In each generation t, the fitness of each individual is calculated: $f_i^{(t)} = f(x_i^{(t)})$ for $i = 1, \ldots, N$.

Step 2: Selection: Individuals are selected from $P^{(t)}$ to form an intermediate population $M^{(t)} = \{m_1^{(t)}, \ldots, m_N^{(t)}\}$. The selection probability of an individual $x_i^{(t)}$ is often proportional to its relative fitness: $P(x_i^{(t)}) = f_i^{(t)} / \sum_{j=1}^{N} f_j^{(t)}$.

Step 3: Crossover: Selected pairs of individuals (parents) from $M^{(t)}$ are combined to generate a new population of offspring $Y^{(t)} = \{y_1^{(t)}, \ldots, y_N^{(t)}\}$. The crossover operation C exchanges genetic material (parameter values) between the parents: $(y_{2k-1}^{(t)}, y_{2k}^{(t)}) = C(m_{2k-1}^{(t)}, m_{2k}^{(t)})$.

Step 4: Mutation: A mutation operation M is applied to each offspring $y_i^{(t)}$ with a small probability μ. Mutation introduces small random alterations to the parameters: $z_i^{(t)} = M(y_i^{(t)})$ with probability μ, or $z_i^{(t)} = y_i^{(t)}$ otherwise. This generates the population $Z^{(t)} = \{z_1^{(t)}, \ldots, z_N^{(t)}\}$.

Step 5: Replacement and Iteration: The population for the next generation $P^{(t+1)}$ is formed from $Z^{(t)}$ (often $P^{(t+1)} = Z^{(t)}$ in simple GAs, or using replacement strategies such as elitism, which preserves the best individuals). The process repeats from Step 1 until a termination criterion is met (e.g., maximum number of generations, fitness convergence).

In this study, the objective function to be minimized by the GA is the Root Mean Square Error (RMSE) between the dengue cases simulated by the model ($\text{Cases}_{\text{model}}(t)$) and the actual observed cases ($\text{Cases}_{\text{real}}(t)$) over a time period T:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\big(\text{Cases}_{\text{model}}(t) - \text{Cases}_{\text{real}}(t)\big)^2}.$$

The result of the optimization process is the set of parameters that yields the lowest RMSE, representing the best fit found by the algorithm. Table 1.2 shows examples of other infectious diseases where this combination of SEIR models and GA has been applied.

Cuadro 1.2: Examples of infectious diseases modeled with SEIR variants and calibrated using Genetic Algorithms.

Disease	**Application of SEIR with GA (Conceptual Examples)**
COVID-19	Parameter fitting (transmission, recovery) for peak prediction [2].
Influenza	Calibration of seasonal dynamics, evaluation of optimal vaccination strategies.
Measles	Estimation of contact rates, optimization of public health interventions.
Ebola	Fine-tuning complex models to understand transmission patterns.
Dengue	Optimization of intervention timing and vector control strategies [15, 6].

Model Implementation Details

To reflect the influence of external factors on transmission, seasonal or climatic forcing is incorporated. This is achieved by allowing the transmission rates $\beta_{VH}(t)$ and/or $\beta_{HV}(t)$ to vary over time. This variation can be modeled as a periodic function (e.g., $\beta_{VH}(t) = \beta^0_{VH}[1 + \alpha\cos(2\pi t/T)]$) or be directly driven by monthly climate data, such as temperature, following functional relationships established in the literature [11, 10]. Additionally, to emulate the occurrence of new outbreaks or case importation, some implementations allow for the periodic reintroduction of a small number of infectious individuals (I_H) or infectious vectors (I_V) into the system.

Cuadro 1.3: Reference values, estimation ranges, and initial conditions considered for the SEIR-SEI dengue model in Villavicencio.

Parameter / Data	**Value / Range / Observation**
Human population (N_H)	$\approx$ 50,000 - 85,000 (variable annually, Fig. 1.2b)
Initial $S_H(0)$	$\approx 0{,}95 \times N_H(0)$
Initial $I_H(0)$	Based on reported cases at start / small value (e.g., 10-100)
Initial $R_H(0)$	Fraction based on seroprevalence or 0
Initial $E_H(0)$	0 or small fraction of $I_H(0)$
Initial $S_V(0)$	Proportion of N_H (e.g., 1-10 times) or based on carrying capacity K_V
Initial $E_V(0), I_V(0)$	Small fractions of total vectors (e.g., 1 %)
β_{HV} (human $\to$ vector)	Estimated by GA (range e.g., [0.1, 1.0] day^{-1})
β_{VH} (vector $\to$ human)	Estimated by GA (range e.g., [0.1, 1.0] day^{-1})
σ_H (human incubation^{-1})	Fixed (e.g., 1/5 day^{-1}) or Estimated (range e.g., [1/10, 1/3] day^{-1})
σ_V (vector incubation^{-1})	Fixed (e.g., 1/10 day^{-1}) or Estimated (range e.g., [1/14, 1/7] day^{-1})
γ_H (human recovery^{-1})	Fixed (e.g., 1/7 day^{-1}) or Estimated (range e.g., [1/10, 1/4] day^{-1})
μ_H (human mortality/birth)	Fixed (e.g., $\approx 1/(75 \times 365)$ day^{-1})
μ_V (vector mortality^{-1})	Fixed (e.g., 1/14 day^{-1}) or Estimated (range e.g., [1/21, 1/7] day^{-1})
Λ_V (vector birth rate)	Often $\Lambda_V = \mu_V N_V$ for disease-free equilibrium
Forcing parameters (α, T)	Estimated or based on climate data
Reintroduction parameters (ϵ, freq.)	Defined for specific scenarios

Numerical integration of the ODE system is performed using standard methods, such as higher-order Runge-Kutta algorithms (e.g., implemented in the `solve_ivp` function from the SciPy library in Python), calculating the evolution of the compartments in discrete time steps (daily or monthly). Model initialization requires defining the compartment values at $t = 0$. Typically,

$S_H(0)$ is set close to the total population N_H, with a small initial fraction of individuals in $E_H(0)$ and/or $I_H(0)$ based on historical data from the beginning of the simulation period. Similarly, initial conditions are defined for the vector population $S_V(0), E_V(0), I_V(0)$, often assuming a proportion relative to the human population or an equilibrium state. Table 1.3 summarizes the parameter ranges explored by the GA and the initial conditions used or considered in this type of model.

The combination of seasonal forcing (driven by climate) and possible periodic reintroductions are key mechanisms that allow SEIR-SEI models to generate multi-wave dynamics or recurrent outbreaks, instead of a single epidemic that depletes susceptibles, thus seeking greater correspondence with the incidence patterns observed in real data such as those in Figure 1.2a.

1.3. Numerical experiments

The calibrated SEIR-SEI model, incorporating seasonal forcing and periodic reintroductions, was used to simulate the dynamics of dengue in Villavicencio. Figure 1.3 compares the simulated fraction of infected individuals (I_H/N_H, solid line) with the observed data (converted to fraction, dashed line) over time. The final model simulation captures the occurrence of multiple epidemic waves, a characteristic present in the real data for the analyzed period.

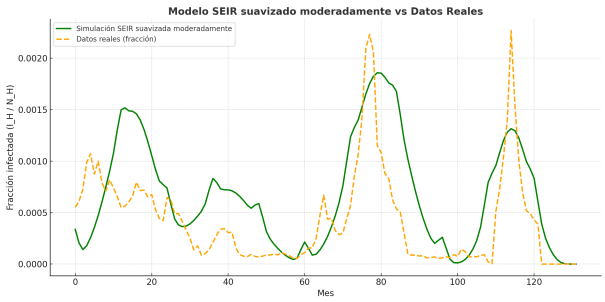


Figura 1.3: Comparison of the fraction of the infected population (I_H/N_H) simulated by the SEIR-SEI model with seasonality and reintroductions (solid line) versus the observed data in Villavicencio (dashed line) over time (months).

The quantitative fit of the final calibrated model to the observed data from Villavicencio resulted in a Root Mean Square Error (RMSE) of 0.00047, a Mean Absolute Error (MAE) of 0.00037, and a Mean Absolute Percentage Error (MAPE) that was not interpretable due to the presence of near-zero values in the observed data.

Table 1.4 presents the final values of the key SEIR-SEI model parameters estimated through the calibration process with genetic algorithms, along with their 95

Cuadro 1.4: Final estimated parameters of the calibrated SEIR-SEI model for Villavicencio, with 95 % Confidence Intervals (CI).

Parameter	Estimated value	95 % CI	Unit
β_{HV} (human $\rightarrow$ vector)	0.52	[0.44, 0.60]	days^{-1}
β_{VH} (vector $\rightarrow$ human)	0.39	[0.32, 0.47]	days^{-1}
σ_H (human incubation)	0.18	[0.14, 0.22]	days^{-1}
σ_V (vector incubation)	0.22	[0.18, 0.26]	days^{-1}
γ_H (human recovery)	0.14	[0.12, 0.17]	days^{-1}
μ_V (vector mortality)	0.071	[0.065, 0.079]	days^{-1}

During the calibration process, the performance of different genetic algorithm variants in minimizing the RMSE was evaluated. Figure 1.4 illustrates the convergence trajectory (RMSE as a function of generations) for some of the tested variants. Table 1.5 summarizes the minimum RMSE achieved and the average execution time for the selected variants, indicating differences in optimization efficiency and effectiveness. The NSGA-II-lite and SPEA2-lite variants achieved the lowest RMSE values in these comparative tests.

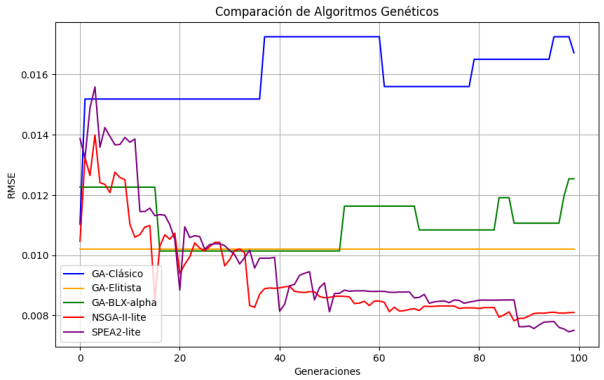


Figura 1.4: Comparison of convergence (RMSE as a function of generations) for five genetic algorithm variants used for parameter estimation of the SEIR-SEI model.

Cuadro 1.5: Performance comparison between Genetic Algorithm variants: Minimum RMSE achieved during optimization and average execution time.

GA Variant	Minimum RMSE	Average Time (s)
Standard GA (Classic)	0.0167	0.25
Elitist GA (N1)	0.0102	0.22
NSGA-II-lite (N2)	0.0078	0.25
SPEA2-lite (N3)	0.0075	0.29

To evaluate the relative contribution of the model components that enable the generation of multiple waves, simulations were performed under hypothetical scenarios, selectively removing seasonal forcing and/or periodic reintroductions. Table 1.6 compares the goodness-of-fit metrics (RMSE and R^2) obtained in these scenarios with those of the complete final model. The numerical results show that omitting either of these components substantially deteriorates the model's fit to the observed data, increasing the RMSE and reducing R^2.

Cuadro 1.6: Evaluation of the SEIR-SEI model fit under different hypothetical scenarios, removing key components.

Scenario	RMSE	R^2
Without reintroduction	0.0167	0.78
Without seasonality	0.0153	0.80
Without reintroduction or seasonality	0.0191	0.73
Seasonality + Reintroduction (Final Model)	0.0083	0.91

Numerical experiments indicate that the joint incorporation of seasonal forcing (or climate-dependent parameters) and periodic reintroductions is necessary for the SEIR-SEI model to reproduce the multi-wave dynamics observed in the real data from Villavicencio (Table 1.6). The absence of one or both factors results in a single-wave pattern or a significantly poorer fit. The use of genetic algorithms allowed exploration of the parameter space and obtaining sets of values that reduce the discrepancy between simulations and observations (Table 1.4), achieving a considerable quantitative fit with the final model (R^2=0.91). Models that consider climate-dependent parameters (reflected in seasonality) and allow for additional external infections (reintroductions) align better with the real data.

1.4. Conclusions

The developed SEIR-SEI model, which incorporates seasonal forcing and periodic reintroductions, demonstrated the capability to capture the observed dengue transmission dynamics in Villavicencio. The model parameters were calibrated using genetic algorithms, achieving a quantitative fit to historical incidence data.

The results from the calibrated model suggest that public health interventions implemented in a timely manner, based on the estimated periods of maximum vector proliferation and human transmission, can reduce and delay dengue epidemic peaks. The model analysis identified three main intervention strategies. The first consists of the systematic elimination of mosquito breeding sites, strategically initiated (e.g., day 30 of the simulated period) to minimize the increase in the adult vector population (expected between days 45-60); this action includes inspecting and eliminating stagnant water in high-risk areas, supported by community outreach. The second strategy is staggered fumigation, starting approximately two weeks before the expected rainfall peaks (e.g., day 45), with applications at regular intervals (e.g., days 45, 60, 75) and entomological monitoring (e.g., ovitraps) to adjust frequency if necessary. The third strategy is based on surveillance and early warning systems, monitoring key indicators (climate data, cases) from the beginning, defining alert thresholds (e.g., a 20

Simulations indicate that the coordinated implementation of these measures (breeding site elimination, timely fumigation, and active surveillance) generates a synergistic effect, resulting in dengue incidence peaks of lower magnitude and later onset. The success of these interventions also depends on active community participation (e.g., eliminating containers, reporting symptoms), requiring continuous communication strategies. Additional sensitivity analyses could refine the optimal start time for each policy. Model-based planning can help optimize resource allocation for vector control.

Future work is recommended to refine the seasonal forcing functions by incorporating more

detailed climate data. The application of multi-objective optimization approaches could explicitly evaluate the trade-off between the cost of interventions and their epidemiological impact. Finally, validating the model with datasets from different regions or years, and extending it to consider multiple dengue virus serotypes, would constitute relevant methodological advancements.

Bibliografía

[1] Cai, W., Santoso, A., Wang, G., Yeh, S.-W., An, S.-I., Cobb, K. M., Collins, M., Guilyardi, E., Jin, F.-F., Kug, J.-S., Lengaigne, M., McPhaden, M. J., Takahashi, K., Timmermann, A., Vecchi, G. A., Watanabe, M.,& Wu, L. (2015). *ENSO and greenhouse warming*. Nature Climate Change, **5**(9), 849–859. https://doi.org/10.1038/nclimate2743

[2] Chang, C., Pongsumpun, P.,& Tang, I. (2023). *SEIR-SEI Model with Genetic Algorithms for Dengue Transmission: A Case Study*. International Journal of Epidemiological Modeling, **11**(2), 45–62.

[3] Colón-González, F. J., Caminade, C., Lowe, R., Tompkins, A. G.,& Ebi, K. L. (2018). *Limiting global-mean temperature increase to 1.5–2 °C could reduce the incidence and spatial spread of dengue fever in Latin America*. Proceedings of the National Academy of Sciences, **115**(24), 6243–6248. https://doi.org/10.1073/pnas.1718945115

[4] Esteva, L.& Vargas, C. (1998). *Analysis of a dengue disease transmission model*. Mathematical Biosciences, **150**(2), 131–151. https://doi.org/10.1016/S0025-5564(98)10003-2

[5] Florentino, H. O., Reis, C. A., Pato, M. V., Jones, D. L.,& Cerasuolo, M. (2014). *Multi-objective genetic algorithm applied to optimal control of dengue epidemic model*. In Proceedings of the 19th World Congress of the International Federation of Automatic Control (IFAC), 6830–6835. https://doi.org/10.3182/20140824-6-ZA-1003.01063

[6] Florentino, H. O., Reis, C. A., Pato, M. V., Jones, D. L.,& Cerasuolo, M. (2018). *Genetic algorithm for optimization of the Aedes aegypti control strategies*. Pesquisa Operacional, **38**(3), 457–478. https://doi.org/10.1590/0101-7438.2018.038.03.0457

[7] Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. The University of Michigan Press.

[8] Huber, J. H., Childs, M. L., Caldwell, J. M.,& Mordecai, E. A. (2018). *Seasonal temperature variation influences climate suitability for dengue, chikungunya, and Zika transmission*. PLOS Neglected Tropical Diseases, **12**(5), e0006451. https://doi.org/10.1371/journal.pntd.0006451

[9] Lana, R. M., Cota, W., Marques, F. S., Gomes, M. F. C., Codeço, C. T., da Silva, M. H. L.,& Kraenkel, R. A. (2021). *Assessing the role of climate fluctuations on dengue dynamics in Brazil using a mathematical model*. Scientific Reports, **11**(1), 18019. https://doi.org/10.1038/s41598-021-97409-z

[10] Liu, Y., Wang, X., Tang, S.,& Cheke, R. A. (2023). *Integrating multiple meteorological factors into a five-stage model for Aedes albopictus population dynamics: Pa-*

rameter estimation and analysis. PLOS Neglected Tropical Diseases, **17**(4), e0011247. https://doi.org/10.1371/journal.pntd.0011247

[11] Mordecai, E. A., Cohen, J. M., Evans, M. V., Gudapati, P., Johnson, L. R., Lippi, C. A., Maffei, K., Martins, K. F., Ryan, S. J., Savage, V., Smith, M. S. A., Stewart-Ibarra, A. M., Thomas, M. B.,& Weikel, D. P. (2017). *Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models*. PLOS Neglected Tropical Diseases, **11**(4), e0005568. https://doi.org/10.1371/journal.pntd.0005568

[12] Muñoz, E., Poveda, G., Arbeláez, M. P.,& Vélez, I. D. (2020). *Spatiotemporal dynamics of dengue in Colombia in relation to the combined effects of local climate and ENSO*. medRxiv. https://doi.org/10.1101/2020.08.24.20181032 (Preprint)

[13] Muñoz, E., Stewart-Ibarra, A. M., Arbeláez, M. P., Vélez, I. D.,& Poveda, G. (2024). *Extreme climate events, landscape attributes, and socioeconomic conditions shape dengue risk in Colombia*. medRxiv. https://doi.org/10.1101/2024.04.02.24304484 (Preprint)

[14] Murray, J. D. (2002). *Mathematical Biology I: An Introduction* (3rd ed.). Springer.

[15] Navarro Valencia, V. A., Díaz, Y., Pascale, J. M., Boni, M. F.,& Sanchez-Galan, J. E. (2023). *Using Compartmental Models and Particle Swarm Optimization to Assess Dengue Basic Reproduction Number R0 for the Republic of Panama in the 1999–2022 Period*. Applied Sciences, **13**(8), 5088. https://doi.org/10.3390/app13085088

[16] Osei-Kyei, R., Johnson, L. R.,& Mordecai, E. A. (2018). *Practical unidentifiability of a simple vector-borne disease model: Implications for parameter estimation and intervention assessment*. Journal of the Royal Society Interface, **15**(143), 20180226. https://doi.org/10.1098/rsif.2018.0226

[17] Phaijoo, G. R.,& Gurung, D. B. (2024). *Sensitivity Analysis of SEIR SEI Model of Dengue Disease*. arXiv. https://doi.org/10.48550/arXiv.2401.17196 (Preprint)

[18] Rahman, M. M., Mathiphan, S., Viriyayudhakorn, S.,& Theeramunkong, T. (2021). *SEIR-SEI-EnKF: A New Model for Estimating and Forecasting Dengue Outbreak Dynamics*. IEEE Access, **9**, 158086–158096. https://doi.org/10.1109/ACCESS.2021.3129997

[19] Samat, N. A.,& Mohd Imam Ma'arof, S. H. (2014). *Disease Mapping based on Stochastic SIR-SI Model for Dengue and Chikungunya in Malaysia*. AIP Conference Proceedings, **1635**, 227–232. https://doi.org/10.1063/1.4903587

[20] Samat, N. A.,& Mohd Imam Ma'arof, S. H. (2016). *New Approach to Calculate the Denominator for the Relative Risk Equation*. Jurnal Sains Kesihatan Malaysia (Malaysian Journal of Health Sciences), **13**(2), 67–73.

[21] Side, S., Rangkuti, Y. M., Pane, D. G.,& Sinaga, M. S. (2018). *Stability Analysis Susceptible, Exposed, Infected, Recovered (SEIR) Model for Spread of Dengue Fever in Medan*. IOP Conference Series: Journal of Physics, **954**(1), 012018. https://doi.org/10.1088/1742-6596/954/1/012018

[22] Stewart-Ibarra, A. M., Ryan, S. J., Kanyangarara, M., Happy, K. L., Morin, C. W., Pan, W. K., Dutra, H. L., Munayco, C. V., Bausch, W. F.,& Bausch, D. G. (2024). *Climate-driven dengue cases and Aedes aegypti levels in the Peruvian Amazon*. PLOS Neglected Tropical Diseases, **18**(4), e0012096. https://doi.org/10.1371/journal.pntd.0012096

[23] World Health Organization. (2020). *Dengue and Severe Dengue—Key Facts*. URL: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

[24] Windarto, Khan, M. A.,& Fatmawati. (2020). *Parameter estimation and fractional derivatives of dengue transmission model*. AIMS Mathematics, **5**(3), 2156–2171. https://doi.org/10.3934/math.2020178

Relation between Fourier multipliers in $\mathbb{R}^N$, $\mathbb{T}^N$ and $\mathbb{Z}^N$

Boza, Santiago [♭], Isert, Daniel [♭], Matta, Larry Andrés [♭] Ramis, Bernat[♭] Ibáñez, Jorge Santiago[♭], Vila, Carlos[♭],

(♭) santiago.boza@upc.edu
(♭) daisa2@uv.es
(♭) lamattap9@gmail.com
(♭) bramis@student.ethz.ch
(♭) jorgesib@ucm.es
(♭) cvilaper@alumnos.unex.es

1.1 Introduction

In this paper we will see how to relate operators by Fourier transform in three different contexts: in $\mathbb{R}^N$, $\mathbb{T}^N$ and $\mathbb{Z}^N$.

We will start seeing how the Fourier transform is defined in the three contexts, and how to recover them by means of the inverse formula.

Let $f \in L^1(\mathbb{R}^N)$, its Fourier transform is defined by

$$\hat{f}(\xi) = \int_{\mathbb{R}^N} f(x) e^{-2\pi i x\cdot\xi}\, dx,\ \xi \in \mathbb{R}^N.$$

The inverse Fourier transform of $\hat{f}(\xi)$, when $\hat{f}(\xi) \in L^1(\mathbb{R}^N)$, is defined by

$$f(x) = (\hat{f})^\vee(x) = \int_{\mathbb{R}^N} \hat{f}(\xi) e^{2\pi i x\cdot\xi}\, d\xi,\ \text{a.e. } x \in \mathbb{R}^N.$$

Identifying $\mathbb{T}^N$ with the interval $[-\frac{1}{2}, \frac{1}{2}]^N$ we have that:
For a given function $f \in L^1(\mathbb{T}^N)$, its Fourier coefficients are given by

$$\hat{f}(n) = \int_{\mathbb{T}^N} f(x) e^{-2\pi i x\cdot n}\, dx,\ n \in \mathbb{Z}^N.$$

We can recover the function f from its Fourier coefficients under some regularity conditions by

$$f(x) = \sum_{n\in\mathbb{Z}^N} \hat{f}(n) e^{2\pi i x\cdot n},\ x \in \mathbb{T}^N.$$

Finally, given a sequence $a = \{a(n)\}_{n\in\mathbb{Z}^N} \in \ell^1(\mathbb{Z}^N)$ we have that:
Its Fourier transform is the periodic and continuous function given by

$$\hat{a}(\xi) = \sum_{n\in\mathbb{Z}^N} a(n) e^{-2\pi i \xi\cdot n},\ \xi \in \mathbb{T}^N.$$

We can recover the sequence a from its Fourier transform $\hat{a}$ by:

$$a(n) = \int_{\mathbb{T}^N} \hat{a}(\xi) e^{2\pi i \xi \cdot n},\ n \in \mathbb{Z}^N.$$

Let us continue by recalling the fundamental property of the Fourier transform respect to the convolution product in all three continuous, discrete and periodic contexts. Here K denotes a convolution kernel, and $\hat{K}$ its transform.

$$\widehat{(K * f)}(\xi) = \hat{K}(\xi)\hat{f}(\xi) = m(\xi)\hat{f}(\xi),\ \xi \in \mathbb{R}^N.$$

$$\widehat{(K * f)}(n) = \hat{K}(n)\hat{f}(n) = m(n)\hat{f}(n),\ n \in \mathbb{Z}^N.$$

$$\widehat{(K_d \star a)}(\xi) = \hat{K}(\xi)\hat{f}(\xi) = m(\xi)\hat{f}(\xi),\ \xi \in \mathbb{T}^N.$$

There are many different works in the classic literature that relate convolution operators in $\mathbb{R}^N$, $\mathbb{T}^N$, y $\mathbb{Z}^N$. These operators can be defined by the action of the corresponding multipliers on the side of the Fourier transform, they are the so-called Fourier multipliers.

Let m be a continuous function in $\mathbb{R}^N$, we define

$$(Cf)(x) := (T_m f)(x) = \int_{\mathbb{R}^N} m(\xi)\hat{f}(\xi) e^{2\pi i x \cdot \xi}\ d\xi,\ x \in \mathbb{R}^N,$$

for a continuous function f defined in $\mathbb{R}^N$.

$$(Pg)(x) := (\tilde{T}_m g)(x) = \sum_{k \in \mathbb{Z}^N} m(k)\hat{g}(k) e^{2\pi i k \cdot x}, x \in \mathbb{T}^N,$$

for a periodic function g defined in $\mathbb{T}^N$.

$$(Da)(n) := \int_{[-1/2,1/2]^N} m(\xi) P(\xi) e^{2\pi i n \cdot \xi},\ n \in \mathbb{Z}^N,$$

for a sequence $a = \{a(n)\}_n$ in $\mathbb{Z}^N$ and $P(\xi) = \sum_m a(m) e^{-2\pi i m \cdot \xi}$.

We are going to consider $m \in L^\infty(\mathbb{R}^N)$, for it is a necessary condition in order to get that the operator is well-defined, and in consequence the multiplier is essentially bounded. We define the operator T_m that is bounded in $L^2(\mathbb{R}^N)$ by

$$\widehat{(T_m f)}(\xi) = m(\xi)\hat{f}(\xi).$$

By Plancherel's theorem ($\|f\|_2 = \|\hat{f}\|_2$), so $T_m f$ is well defined if $f \in L^2$ and, moreover

$$\|T_m f\|_2 \leq \|m\|_\infty \|f\|_2.$$

In fact, respect to the norm of the operator T_m in L^2, $\|T_m\|$, we have that

$$\|T_m\| = \|m\|_\infty.$$

When T_m can be extended as a bounded operator in L^p, it is said that m is a Fourier multiplier in L^p. Usually, the operators defined by Fourier multipliers are defined under a dense class, such as Schwartz class in L^p, and they are extended by density to the whole space L^p.

In $\mathbb{R}^N$, under the assumption that $m \in L^\infty$, we want to see when the convolution operator of the kernel $K \in \mathcal{S}'$ (tempered distributions), defined for functions $f \in \mathcal{S}$ (Schwartz class) by

$$T_m f = f * K$$

extends to a bounded operator in L^p, where $\hat{K} = m$.

If G is one of the groups $\mathbb{R}^N$, $\mathbb{T}^N$ or $\mathbb{Z}^N$, we will denote by $\mathcal{M}_p(G)$ the space of the multipliers that define bounded operators in $L^p(\mathbb{R}^N)$, $\ell^p(\mathbb{Z}^N)$ y $L^p(\mathbb{T}^N)$, respectively.

1.2 Restriction of multipliers in $\mathbb{R}^N$ to $\mathbb{Z}^N$

In this section we will see that, given a bounded multiplier in its continuous version and under some certain assumptions for the multiplier, we can obtain that the multiplier in its periodic version is also bounded. This fact is not quite surprising since the second form of the multiplier is the restriction to the integer numbers of the first form in the side of the Fourier transform.

In order to prove such Theorem, we will require the following two Lemmas, we present their proofs here for the sake of completeness. The first one is a technical convergence lemma that can be found in [3, Lemma 3.9, Chapter 7].

Lemma 1.2.1. *Let f be a continuous and periodic function in $\mathbb{R}^N$, then*

$$\lim_{\epsilon\to 0}\epsilon^{N/2}\int_{\mathbb{R}^N} f(x)e^{-\epsilon\pi|x|^2}dx = \int_{Q^N} f(x)dx \text{ , with } Q^N=[0,1]^N. \tag{1.2.1}$$

Proof. Observe that we have the equality if we take a trigonometric polynomial with the form $f(x)=e^{2\pi im\cdot x}$, where $m\in\mathbb{Z}^N$ and $x\in\mathbb{R}^N$. This is because the function f is an exponential type function of Fourier integrand, that in (1.2.1) is integrated as a gaussian function. Knowing that the Fourier transform of a Gaussian is again a Gaussian, then, since

$$\int_{\mathbb{R}^N} e^{-\epsilon\pi|x|^2}e^{2\pi im\cdot x}dx = \epsilon^{-N/2}e^{-\pi|m|^2/\epsilon},$$

we have that, in (1.2.1),

$$\lim_{\epsilon\to 0}\epsilon^{N/2}\int_{\mathbb{R}^N} f(x)e^{-\epsilon\pi|x|^2}dx = \lim_{\epsilon\to 0} e^{-\pi|m|^2/\epsilon}\begin{cases}1 & \text{if } m=0\in\mathbb{Z}^N\\ 0 & \text{if } m\neq 0\in\mathbb{Z}^N.\end{cases}$$

And from here we obtain (1.2.1) for $f(x)=e^{2\pi im\cdot x}$, since its integral in the fundamental cube is the same that we have in the limit.

The Lemma follows for f a continuous and periodic function by approximating f uniformly by trigonometric polynomials that are obtained as linear combinations of $\{e^{2\pi im\cdot x}\}_{m\in\mathbb{Z}^N}$. □

The following Lemma connects the two versions of the multiplier, we can find it in [3, Lemma 3.11, Chapter 7].

Lemma 1.2.2. *Let P and Q be trigonometric polynomials, $T_\lambda : L^p(\mathbb{R}^N)\to L^p(\mathbb{R}^N)$ a multiplier-type operator. $\tilde{T}_\lambda$ is an element of the basis of trigonometric polynomials. Let $w_\delta(y) := e^{-\pi\delta|y|^2}$, $\delta>0$, $y\in\mathbb{R}^N$. Then, for α, $\beta>0$ such that $\alpha+\beta=1$, we have that:*

$$\lim_{\epsilon\to 0}\epsilon^{N/2}\int_{\mathbb{R}^N} T_\lambda(Pw_{\epsilon\alpha})(x)\overline{Q(x)}w_{\epsilon\beta}(x)dx = \int_{Q^N}(\tilde{T}_\lambda P)(x)\overline{Q(x)}dx. \tag{1.2.2}$$

Proof. It will be enough to prove it for $P(x)=e^{2\pi ir\cdot x}$ and $Q(x)=e^{2\pi ik\cdot x}$, $r,k\in\mathbb{Z}^N$ since (1.2.2) is linear in P and in Q. Using the expressions $w_{\epsilon\alpha}(y)=e^{-\pi\epsilon\alpha|y|^2}$, $w_{\epsilon\beta}(y)=e^{-\pi\epsilon\beta|y|^2}$, Plancherel's and Fubini's theorems, it is easy to see that

$$\epsilon^{N/2}\int_{\mathbb{R}^N} T_\lambda(Pw_{\epsilon\alpha})(x)\overline{Q(x)}w_{\epsilon\beta}(x)dx = \epsilon^{N/2}\int_{\mathbb{R}^N}\lambda(x)\varphi(x)\overline{\psi(x)}\,dx, \tag{1.2.3}$$

where φ and ψ are the following Fourier transforms

$$\varphi(x) = (e^{2\pi ir\cdot x}e^{-\pi\epsilon\alpha|x|^2}\hat{)} = e^{-\pi|x-r|^2/(\alpha\epsilon)}(\alpha\epsilon)^{-N/2}$$
$$\psi(x) = (e^{2\pi ik\cdot x}e^{-\pi\epsilon\beta|x|^2}\hat{)} = e^{-\pi|x-k|^2/(\beta\epsilon)}(\beta\epsilon)^{-N/2}.$$

We distinguish now two cases

- $\underline{r \neq k}$: Since r and k are integers, we have that $|r-k| \geq 1$. Moreover, for being λ a multiplier, we can assume that it is in L^∞, i.e, $|\lambda(x)| \leq A$. With this, we can bound the left hand side of (1.2.2) by

$$\begin{aligned}
&\epsilon^{N/2} A \int_{\mathbb{R}^N} e^{-\pi|x-r|^2/(\alpha\epsilon)}(\alpha\epsilon)^{-N/2} e^{-\pi|x-k|^2/(\beta\epsilon)}(\beta\epsilon)^{-N/2}\, dx \\
&\leq \epsilon^{N/2} A \left(\int_{|x-r|\geq \frac{1}{2}} e^{-\pi|x-r|^2/(\alpha\epsilon)}(\alpha\epsilon)^{-N/2} e^{-\pi|x-k|^2/(\beta\epsilon)}(\beta\epsilon)^{-N/2}\, dx \right) + \\
&\epsilon^{N/2} A \left(\int_{|x-k|\geq \frac{1}{2}} e^{-\pi|x-r|^2/(\alpha\epsilon)}(\alpha\epsilon)^{-N/2} e^{-\pi|x-k|^2/(\beta\epsilon)}(\beta\epsilon)^{-N/2}\, dx \right),
\end{aligned}$$

  where the integration domains cover all $\mathbb{R}^N$. Observe that in the first integral, the factor $\epsilon^{N/2} e^{-\pi|x-r|^2/(\alpha\epsilon)}(\alpha\epsilon)^{-N/2}$ tends uniformly to 0 when ϵ goes to 0, while the Gaussian factor $e^{-\pi|x-k|^2/(\beta\epsilon)}(\beta\epsilon)^{-N/2}$ integrates 1 if the integral extends to the whole space $\mathbb{R}^N$. Then, the integral over $|x-r| \geq \frac{1}{2}$ tends to 0 when $\epsilon \to 0$. The same argument applies for the integral over $|x-k| \geq \frac{1}{2}$, from where we conclude that the left hand side of (1.2.2) is 0.

  Now, since P is a trigonometric polynomial, the periodic multiplier acts over P as follows:

$$(\tilde{T}_\lambda)(x) = \lambda(r) e^{2\pi i r\cdot x},$$

  for the corresponding coefficient $\lambda(r)$, $r \in \mathbb{Z}^N$. So the right hand side of (1.2.2) is

$$\int_{Q^N} (\tilde{T}_\lambda P)(x)\overline{Q(x)}dx = \int_{Q^N} \lambda(r) e^{2\pi i r\cdot x} e^{-2\pi i k\cdot x}\, dx = \int_{Q^N} \lambda(r) e^{2\pi i (r-k)\cdot x}\, dx = 0,$$

  for being the integrand of a complex exponential of period a divisor of the sidelength of the fundamental cube. We have seen that, in this case, both sides of (1.2.2) are 0.

- $\underline{r = k}$: The left hand side of (1.2.2) is

$$\lim_{\epsilon\to 0}(\epsilon\alpha\beta)^{-N/2} \int_{\mathbb{R}^N} \lambda(x) e^{-\pi(|x-r|^2/\epsilon)(\frac{1}{\alpha}+\frac{1}{\beta})}\, dx = \lim_{\epsilon\to 0}(\epsilon\alpha\beta)^{-N/2} \int_{\mathbb{R}^N} \lambda(x) e^{\pi(|x-r|^2/\epsilon)(\frac{1}{\alpha\beta})}\, dx. \tag{1.2.4}$$

The result is an integral of Gauss-Weierstrass of λ. Taking into account [3, Theorem 1.25, Chapter 1], since r is a Lebesgue point of λ, it is a continuity point by hypothesis, we conclude that the limit (1.2.4) takes the value $\lambda(r)$. This same value is the same that we obtain directly for $r = k$ in the right hand side of (1.2.2):

$$\int_{Q^N} (\tilde{T}_\lambda P)(x)\overline{Q(x)}dx = \int_{Q^N} \lambda(r) e^{2\pi i r\cdot x} e^{-2\pi i r\cdot x}\, dx = \lambda(r).$$

□

Finally, we present the Theorem mentioned above [3, Chapter 7, Theorem 3.8].

Theorem 1.2.3. *Let $1 \leq p \leq \infty$, $\lambda \in \mathcal{M}^p(\mathbb{R}^N)$ and T_λ the Fourier multiplier operator associated to the function λ that is continuous in every point of $\mathbb{Z}^N$. Then, there exists a unique periodic operator $\tilde{T}_\lambda$ defined by*

$$(\tilde{T}_\lambda f)(x) = \sum_{n\in\mathbb{Z}^N} \lambda(n) a(n) e^{2\pi i\, n\cdot x},$$

for

$$f(x) = \sum_{n\in\mathbb{Z}^N} a(n)e^{2\pi i\, n\cdot x},$$

such that $\{\lambda(n)\}_{n\in\mathbb{Z}} \in \mathcal{M}^p(\mathbb{Z}^N)$. *Moreover*

$$\|\tilde{T_\lambda}\| \leq \|T_\lambda\|.$$

Proof. We will do a distinction of cases for p.

- $\underline{1 < p < +\infty}$: Let P and Q trigonometric polynomials, and let $w_{\epsilon\alpha}$ and $w_{\epsilon\beta}$ be the weight functions defined in Lemma 1.2.2. Then, by Hölder's inequality and Lemma 1.2.2 we get that

$$\begin{aligned}\int_{Q^N} (\tilde{T_\lambda}P)(x)\overline{Q(x)}dx &\overset{\text{Lemma 1.2.2}}{=} \lim_{\epsilon\to 0} \epsilon^{N/2}\left|\int_{\mathbb{R}^N} (T_\lambda(Pw_{\epsilon\alpha}))(x)\overline{Q(x)}w_{\epsilon\beta(x)dx}\right| \\ &\leq \lim_{\epsilon\to 0} \epsilon^{N/2}\|T_\lambda\|\|Pw_{\epsilon\alpha}\|_p\|Qw_{\epsilon\beta}\|_q.\end{aligned}$$

  Take now $\alpha = \frac{1}{p}$ and $\beta = \frac{1}{q}$, that they satisfy the condition $\alpha + \beta = 1$ since they are conjugate exponents. Then

$$\begin{aligned}&\|T_\lambda\| \lim_{\epsilon\to 0} \epsilon^{N/2}\|Pw_{\frac{\epsilon}{p}}\|_p\|Qw_{\frac{\epsilon}{q}}\|_q \\ &= \|T_\lambda\| \lim_{\epsilon\to 0}\left[\epsilon^{N/2}\int_{\mathbb{R}^N} |P(x)|^p e^{-\epsilon\pi|x|^2}dx\right]^{\frac{1}{p}}\left[\epsilon^{N/2}\int_{\mathbb{R}^N} |Q(x)|^q e^{-\epsilon\pi|x|^2}dx\right]^{\frac{1}{q}} \\ &\overset{\text{Lemma 1.2.1}}{=} \|T_\lambda\|\|P\|_{L^p(Q^N)}\|Q\|_{L^q(Q^N)}.\end{aligned}$$

  Taking supremum $\|P\|_{L^p(Q^N)} \leq 1$ and $\|Q\|_{L^q(Q^N)} \leq 1$, we conclude that $\|\tilde{T_\lambda}\| \leq \|T_\lambda\|$.

- $\underline{p = 1}$: Since $T \in (L^1(\mathbb{R}^N), L^1(\mathbb{R}^N))$ ensures us that the multiplier λ is a finite Borel measure $\lambda = \mu$. Let $\hat{\lambda} = \hat{\mu}$ be its transform. Then, by $\sum_{r\in\mathbb{Z}^N} \hat{\mu}(r)e^{2\pi ir\cdot x}$ is the Fourier series of a measure $\tilde{\mu}$ in $\mathbb{T}^N$ and $\|d\tilde{\mu}\| \leq \|d\mu\|$, hence $\tilde{\mu}$ is finite. Finally, by applying [3, Theorem 3.4, Chapter 7] we obtain that $\tilde{T} \in (L^1(\mathbb{T}^N), L^1(\mathbb{T}^N))$ and $\|\tilde{T}\| = \|d\tilde{\mu}\| \leq \|d\mu\| = \|T\|$.

- $\underline{p = \infty}$: By duality, we have that

$$(L^\infty(\mathbb{R}^N), L^\infty(\mathbb{R}^N)) = (L^1(\mathbb{R}^N), L^1(\mathbb{R}^N)).$$

  Moreover, $\|T\|_{L^1(\mathbb{R}^N)} \leq \|T\|_{L^\infty(\mathbb{R}^N)}$. Then, by doing the same arguments as we did in the case $p = 1$, it is clear that $\tilde{T} \in (L^1(\mathbb{T}^N), L^1(\mathbb{T}^N))$. Now, by [3, Theorem 3.4, Chapter 7] there exists a measure μ in $\mathbb{T}^N$ such that $\|d\mu\| = \|\tilde{T}\|$ and $\tilde{T}f = f * d\mu$. For $f \in L^\infty(\mathbb{T}^N)$, $\|f * d\mu\| \leq \|f\|_{L^\infty(\mathbb{T}^N)}\|d\mu\| \implies \|\tilde{T}\|_{L^\infty(\mathbb{T}^N)} \leq \|\tilde{T}\|_{L^1(\mathbb{T}^N)} \leq \|T\|_{L^1(\mathbb{R}^N)} \leq \|T\|_{L^\infty(\mathbb{R}^N)}$.

□

Observation 1.2.4. *Observe that the hypothesis: "λ continuous in every point of $\mathbb{Z}^N$" can be relaxed to the following one: "every point of $\mathbb{Z}^N$ is a Lebesgue point of λ", that is what we are really using in the proof of Lemma 1.2.2.*

1.3 Extension of multipliers in $\mathbb{Z}^N$ to $\mathbb{R}^N$

The result that we present in this section is the reciprocal of Theorem 1.2.3, i.e., now we are going to see that if we have a periodic Fourier multiplier, then we can obtain the continuous Fourier multiplier. For the proof of such result we will make use of the following Lemma.

Lemma 1.3.1. *There exists a non negative function $\eta \in C_0(\mathbb{R}^N)$ such that*

(a) $\eta(0) = 1$,

(b) $\sum_{m\in\mathbb{Z}^N}(\eta(x+m))^p = 1$.

Proof. Consider a function $\eta_1 \in C_0(\mathbb{R}^N)$ such that $\eta_1(0) = 1$, y $\eta_1(m) = 0$ if $m \in \mathbb{Z}^N \setminus \{0\}$ and $\eta(x) > 0$ for every $x \in \overline{Q}_N$. Define now

$$\eta_2(x) = \frac{\eta_1(x)}{\sum_{m\in\mathbb{Z}^N}\eta(x+m)},$$

observe that

(a) By the definition properties of η_1 we get that

$$\eta_2(0) = \frac{\eta_1(0)}{\sum_{m\in\mathbb{Z}^N}\eta_1(m)} = \frac{1}{\eta_1(0)} = 1.$$

(b) Summing η_2 in all of the elements of the reticle $\mathbb{Z}^N$

$$\sum_{m\in\mathbb{Z}^N}\eta_2(m) = \frac{\sum_{m\in\mathbb{Z}^N}\eta_1(m)}{\sum_{m\in\mathbb{Z}^N}\eta_1(m)} = 1.$$

We conclude that the function we were in search for is $\eta = \eta_2^{\frac{1}{p}}$. □

Observation 1.3.2. *In the following result, $\Pi(\mathbb{T}^N)$ will denote the trigonometric polynomials in the torus, and $\mathcal{D}$ will denote the functions of fast decreasing.*

Let us now present the Theorem before mentioned.

Theorem 1.3.3. *Let λ be a continuous function in $\mathbb{R}^N$. Suppose that for every $\epsilon > 0$ there exists an operator $\tilde{T}_\epsilon \in \big(L^p(\mathbb{T}^N), L^p(\mathbb{T}^N)\big)$ given by*

$$(\tilde{T}_\epsilon f)(x) \sim \sum_{m\in\mathbb{Z}^N}\lambda(\epsilon m)a_m e^{2\pi imx},$$

where $\{a_m\}_{m\in\mathbb{Z}^N} = \left\{\widehat{f}(m)\right\}_{m\in\mathbb{Z}^N}$. Suppose also that $\|\tilde{T}_\epsilon\|$ is uniformly bounded. Then, λ is a multipier of type $\big(L^p(\mathbb{R}^N), L^p(\mathbb{R}^N)\big)$, T is its corresponding operator, and $\|T\| \le \sup_{\epsilon>0}\|\widehat{T}_\epsilon\|$.

Proof. Observe first that the case $p = \infty$ can be reduced to the case $p = 1$.

Let $f(x) = e^{2\pi imx}$ and $g(x) = e^{2\pi ikx}$. On the one hand

$$\int_{\mathbb{T}^N}(\tilde{T}_\epsilon f)(x)g(-x)dx = \int_{\mathbb{T}^N}\lambda(\epsilon m)e^{2\pi imx}e^{-2\pi ikx}dx =$$

$$\int_{\mathbb{T}^N}\lambda(\epsilon m)e^{2\pi i(m-k)x}dx = \begin{cases}0 & \text{si } m \neq k,\\ \lambda(\epsilon m) & \text{si } m = k.\end{cases}$$

On the other hand

$$\int_{\mathbb{T}^N} (\tilde{T}_\epsilon g)(x) f(-x) dx = \int_{\mathbb{T}^N} \lambda(\epsilon m) e^{2\pi i k x} e^{-2\pi i m x} dx =$$

$$\int_{\mathbb{T}^N} \lambda(\epsilon m) e^{2\pi i (k-m) x} dx = \begin{cases} 0 & \text{si } m \neq k, \\ \lambda(\epsilon m) & \text{si } m = k. \end{cases}$$

Obtaining, by linearity, that for $f, g \in \Pi(\mathbb{T}^N)$

$$\int_{\mathbb{T}^N} (\tilde{T}_\epsilon f)(x) g(-x) dx = \int_{\mathbb{T}^N} (\tilde{T}_\epsilon g)(x) f(-x) dx.$$

By doing a duality argument, we know that

$$\|\tilde{T}_\epsilon\|_1 \leq \|\tilde{T}_\epsilon\|_\infty.$$

Then, if $p = 1$ and we find an operator $T \in \left(L^1(\mathbb{R}^N), L^1(\mathbb{R}^N)\right)$ whose multiplier is λ and that satisfies the following inequality

$$\|T\|_1 \leq \sup_{\epsilon > 0} \|\widehat{T}_\epsilon\|_1,$$

in particular $T \in \left(L^\infty(\mathbb{R}^N), L^\infty(\mathbb{R}^N)\right)$, and

$$\|T\|_\infty \leq \sup_{\epsilon > 0} \|\widehat{T}_\epsilon\|_\infty.$$

This way, all reduces to show the case $1 \leq p < \infty$.

Suppose, by simplicity, that $\|\widehat{T}_\epsilon\|_p \leq 1$ for every $\epsilon > 0$, as a consequence of the arguments used in [3, Theorem 3.1, Chapter 7], we have that $|\lambda(\epsilon m)| \leq 1$ for every $m \in \Lambda$ and $\epsilon > 0$. Notice that the set $\left\{\epsilon m \mid \epsilon > 0,\ m \in \mathbb{Z}^N\right\}$ is dense in $\mathbb{R}^N$, therefore λ is bounded. Observe that if $f \in L^2(\mathbb{R}^N)$, then $\lambda \widehat{f} \in L^2(\mathbb{R}^N)$. In particular we can define Tf for $f \in \mathcal{D}$ as the function whose Fourier transform is $\lambda \widehat{f}$, i.e.

$$\widehat{Tf}(x) = \lambda(x) \widehat{f}(x).$$

Our goal now is to show that $\|Tf\|_p \leq \|f\|_p$. For that, take a function $f \in \mathcal{D}$ and consider the dilation

$$f_\epsilon(x) = \frac{1}{\epsilon^N} f(\frac{x}{\epsilon}),$$

and then, periodize it

$$\tilde{f}_\epsilon(x) = \epsilon^{-N} \sum_{m \in \mathbb{Z}^N} f(\frac{x+m}{\epsilon}).$$

Applying Poisson summation formula, we can express the dilated and periodized function as follows

$$\tilde{f}_\epsilon(x) = \sum_{m \in \mathbb{Z}^N} \widehat{f}(\epsilon m) e^{2\pi i m x},$$

since, $\widehat{\tilde{f}_\epsilon}(x) = \epsilon^N \widehat{f}(\epsilon x)$. Observe now that, if we apply the operator $\tilde{T}_\epsilon$ to the function we have just defined, we obtain the following function

$$\epsilon^N \left(\tilde{T}_\epsilon \tilde{f}_\epsilon\right)(\epsilon x) = \epsilon^N \sum_{m \in \mathbb{Z}^N} \lambda(\epsilon m) \widehat{f}(\epsilon m) e^{2\pi i \epsilon m x}.$$

Now, recall that λ is bounded and $\widehat{f} \in \mathcal{S}$, then the right hand expression of the previous equality is a Riemann sum. So, taking limits when $\epsilon \to 0$, we have that such expression converges to

$$\int_{\mathbb{R}^N} \lambda(t)\widehat{f}(t)e^{2\pi ixt}dt = Tf(x).$$

Obtaining that λ is a Fourier multiplier and T is its corresponding operator. Let us see now the bound on the norm. We have just obtained that

$$\lim_{\epsilon\to 0} \epsilon^N(\tilde{T}_\epsilon \tilde{f}_\epsilon)(\epsilon x) = Tf(x).$$

So, applying Lemma 1.3.1, we know that there exist a function η compactly supported such that $\eta(0) = 1$. Then, by limit properties we have that

$$\lim_{\epsilon\to 0} \epsilon^N(\tilde{T}_\epsilon \tilde{f}_\epsilon)(\epsilon x)\eta(\epsilon x) = Tf(x) \quad \forall x \in \mathbb{R}^N.$$

Hence, applying Fatou's lemma we deduce that

$$\|Tf\|_p = \int_{\mathbb{R}^N} |Tf(x)|^p dx = \int_{\mathbb{R}^N} \lim_{\epsilon\to 0} \epsilon^{Np} \left|(\tilde{T}_\epsilon \tilde{f}_\epsilon)(\epsilon x)\eta(\epsilon x)\right|^p dx \leq$$

$$\leq \liminf_{\epsilon\to 0} \epsilon^{Np} \int_{\mathbb{R}^N} \left|(\tilde{T}_\epsilon \tilde{f}_\epsilon)(\epsilon x)\eta(\epsilon x)\right|^p dx.$$

Let us bound the last integral in the previous inequality. By doing a change of variables we obtain that

$$\epsilon^{Np} \int_{\mathbb{R}^N} \left|(\tilde{T}_\epsilon \tilde{f}_\epsilon)(\epsilon x)\eta(\epsilon x)\right|^p dx = \epsilon^{N(p-1)} \int_{\mathbb{R}^N} \left|(\tilde{T}_\epsilon \tilde{f}_\epsilon)(u)\eta(u)\right|^p du =$$

$$\epsilon^{N(p-1)} \sum_{m\in\mathbb{Z}^N} \int_{Q_N} \left|(\tilde{T}_\epsilon \tilde{f}_\epsilon)(u+n)\eta(u+n)\right|^p du = \epsilon^{N(p-1)} \int_{Q_N} \left|(\tilde{T}_\epsilon \tilde{f}_\epsilon)(u+n)\right|^p \sum_{m\in\mathbb{Z}^N} |\eta(u+n)|^p \, du.$$

But, by Lemma 1.3.1 we know that $\sum_{m\in\mathbb{Z}^N} |\eta(u+n)|^p = 1$, then

$$\epsilon^{Np} \int_{\mathbb{R}^N} \left|(\tilde{T}_\epsilon \tilde{f}_\epsilon)(\epsilon x)\eta(\epsilon x)\right|^p dx = \epsilon^{N(p-1)} \int_{Q_N} \left|(\tilde{T}_\epsilon \tilde{f}_\epsilon)(u+n)\right|^p du = \epsilon^{N(p-1)} \|\tilde{T}_\epsilon \tilde{f}_\epsilon\|_p^p \leq \epsilon^{N(p-1)} \|\tilde{f}_\epsilon\|_p^p.$$

But now we know that, for ϵ sufficient small, $\epsilon^{-N} f(\frac{x}{\epsilon})$ is in the fundamental cube, Q_N. Then, for such ϵ, we can rewrite the previous inequality as follows

$$\epsilon^{Np} \int_{\mathbb{R}^N} \left|(\tilde{T}_\epsilon \tilde{f}_\epsilon)(\epsilon x)\eta(\epsilon x)\right|^p dx \leq \epsilon^{N(p-1)} \int_{Q_N} \left|\tilde{f}_\epsilon(x)\right|^p dx =$$

$$\epsilon^{N(p-1)} \int_{\mathbb{R}^N} \left|\epsilon^{-N} f\left(\frac{x}{\epsilon}\right)\right|^p dx = \int_{\mathbb{R}^N} |f(x)|^p dx = \|f\|_p^p.$$

Observe that, what we have just obtained is

$$\|Tf\|_p \leq \liminf_{\epsilon\to 0} \epsilon^{Np} \int_{\mathbb{R}^N} \left|(\tilde{T}_\epsilon \tilde{f}_\epsilon)(\epsilon x)\eta(\epsilon x)\right|^p dx \leq \int_{\mathbb{R}^N} |f(x)|^p dx.$$

Hence

$$\|Tf\|_p \leq \|f\|_p.$$

Which leads us to the equality that we were searching

$$\|T\| \leq \sup_{\epsilon>0} \|\tilde{T}_\epsilon\|.$$

So we have obtained the Theorem for a function $f \in \mathcal{D}$, which is a dense class in the space $L^p(\mathbb{R}^N)$ for $1 \leq p < \infty$. For $f \in L^p(\mathbb{R}^N)$, proceed by density. □

1.4 Restriction of multipliers in $\mathbb{R}^N$ and $\mathbb{Z}^N$

Now, we want to explore the raltionship between the continuous and the discrete case following the scheme presented in [1]. In the discrete case, the Fourier transform makes us work with periodic functions. First let us set all the notation that we are going to use in this section.

Definition 1.4.1. *Let f be a function defined in $\mathbb{R}^N$, denote by $(f)^d$ the restriction of f to $\mathbb{Z}^N$.*

Previously, we talked about the Schwartz class, now we are going to use a family of functions that form a subset of the Schwartz class: Functions of type R-exponential.

Definition 1.4.2. *We will say that a function G in $\mathbb{R}^N$ is of type R-exponential if the support of its transform forms a subset of the box $[-R, R]^N$.*

Finally, we will present the notation of the discrete convolution.

Definition 1.4.3. *Let K be a convolution kernel under $\mathbb{R}^N$, and let φ be a function in the same space. We will denote by K^φ the convolution between K and φ, i.e, $K^\varphi = K * \varphi$.*

Apart from these definitions, we will make use of the well-known interpolation theorem of Riesz-Thorin. A more general version of this result can be found in [3, Chapter 5].

Theorem 1.4.4 (Interpolation of Riesz-Thorin)**.** *Let Ω be a measure space, if we have a linear and continuous operator $T : L^p(\Omega) \to L^p(\Omega)$ for $p = p_1$ and $p = p_2$, then, it is continuous for p_t with $\frac{1}{p_t} = \frac{1-t}{p_1} + \frac{t}{p_2}$ for every $t \in [0, 1]$.*

With all the notation set, let us move to presenting the two lemmas that we are going to need for obtaining a discrete multiplier from a continuous one. The first Lemma talks about the discretization effect of a function in its norm. More specifically, it says that for a function in $L^p(\mathbb{R}^N)$, its discretization belongs to $\ell_p(\mathbb{Z}^N)$.

Lemma 1.4.5. *Let $1 \leq p \leq \infty$. Then, there exists a constant C such that:*

$$\|(g)^d\|_{\ell_p}^p \leq C^p \|g\|_{L^p}^p$$

for every g of R-exponential type.

Proof. We will show it first for $R = 1$ and for the cases $p = 1, \infty$. After that, by applying Riesz-Thorin interpolation theorem, Theorem 1.4.4, we will obtain the desired result for every p.

First, the case $p = \infty$ is clear, since

$$\sup_{m \in \mathbb{Z}^N} |g(m)| \leq \sup_{x \in \mathbb{R}^N} |g(x)|.$$

For the case $p = 1$ we will assume that $R = 1$. Take a function ψ such that $\text{supp}\psi \subseteq [-2, 2]^N$ and $\hat{\psi}(\xi) = 1$ when $\xi \in [-1, 1]^N$. Then $(g * \psi)\widehat{a}(\xi) = \hat{g}(\xi)\hat{\psi}(\xi) = \hat{g}(\xi)$ since for $\xi \in [-1, 1]^N$ $\hat{\psi}(\xi) = 1$, and in the other case $\hat{g}(\xi) = 0$. By the Inversion theorem we have that $(\psi * g)(x) = g(x)$ and:

$$\sum_{m \in \mathbb{Z}^N} |g(m)| \leq \sum_{m \in \mathbb{Z}^N} \int_{\mathbb{R}^N} |g(x)||\psi(m - x)|dx = \int_{\mathbb{R}^N} |g(x)| \sum_{m \in \mathbb{Z}^N} |\psi(m - x)|dx \leq C\|g\|_{L^1}.$$

Since, for being ψ a function in the Schwartz class, there exists C a bound of $\sum_{m \in \mathbb{Z}^N} |\psi(m-x)|$. Applying Riesz-Thorin interpolation theorem, we obtain the result for $R = 1$.

If $R \neq 1$, define $\alpha = \lfloor R \rfloor + 1$, therefore g is of α-exponential type, and the function $g\left(\frac{\cdot}{\alpha}\right)$ is of 1-exponential type and

$$\sum_{m\in\mathbb{Z}^N} |g(m)|^p \leq \sum_{m\in\mathbb{Z}^N} |g(\frac{m}{\alpha})|^p \leq C^p \int_{\mathbb{R}^N} \left|g\left(\frac{x}{\alpha}\right)\right|^p dx = C^p \alpha^N \|g\|_{L^p}^p.$$

□

On the other hand, for the proof of the main theorem of this section, we will define a function from a sequence in $\ell_p(\mathbb{Z}^N)$. For that, we will need to relate the norm of this function with the norm of the sequence. This Lemma presents the same ideas of the previous one, we present its proof for the sake of completeness.

Lemma 1.4.6. *Let $1 \leq p \leq \infty$, $\hat{\varphi} \in L^\infty(\mathbb{R}^N) \cap \mathcal{M}_p(\mathbb{R}^N)$ be such that supp$\hat{\varphi} \subseteq [-R, R]^N$. Then $\varphi \in L^p$ and there exists a constant C such that:*

$$\| \sum_{m\in\mathbb{Z}^N} a(m)\varphi(\cdot - m)\|_{L^p} \leq C\|\hat{\varphi}\|_{\mathcal{M}_p}\|a\|_{\ell^p},$$

for every sequence $\{a(m)\}_{m\in\mathbb{Z}^N}$ in ℓ^p.

Proof. For $R = 1$ consider the function $\psi \in L^p(\mathbb{R}^N)$ such that supp$\hat{\psi} \subseteq [-2, 2]^N$, $\hat{\psi} \in C^\infty$, and $\hat{\psi}(\xi) = 1$ when $\xi \in [-1, 1]^N$. We know that φ is a function of 1-exponential type and $\varphi = \psi * \varphi$. Since $\psi \in L^p(\mathbb{R}^N)$ and φ is a kernel convolution in $L^p(\mathbb{R}^N)$, we have that the convolution also belongs to $L^p(\mathbb{R}^N)$. Now, since $\varphi \in \mathcal{M}_p(\mathbb{R}^N)$, we have that

$$\left\| \sum_{m\in\mathbb{Z}^N} a(m)\varphi(\cdot - m) \right\|_{L^p} = \left\| \sum_{m\in\mathbb{Z}^N} a(m)\psi(\cdot - m) * \varphi \right\|_{L^p} \leq \|\hat{\varphi}\|_{\mathcal{M}_p(\mathbb{R}^N)} \left\| \sum_{m\in\mathbb{Z}^N} a(m)\psi(\cdot - m) \right\|_{L^p}.$$

If $p = 1$, it is clear that the right hand side of the last inequality is less or equal to $\|a\|_{\ell^1}\|\psi\|_{L^1}$. If $p = \infty$, it is less or equal to $\|a\|_{\ell^\infty} \sum_{m\in\mathbb{Z}^N} |\psi(x - n)|$, and since ψ is in the Schwartz class we have that it is bounded. By Theorem 1.4.4, we obtain the result for $1 \leq p \leq \infty$ and $R = 1$.

If $R < 1$, repeat the process with $\psi_{\frac{1}{R}}(x) = \psi(Rx)R^N$, which gives similar results. □

With both lemmas in mind, we tackle now the problem of this section, how to construct a discrete Fourier multiplier from a continuous one.

Theorem 1.4.7. *Let $1 \leq p \leq \infty$, $\hat{\varphi} \in L^\infty(\mathbb{R}^N) \cap \mathcal{M}_p(\mathbb{R}^N)$ with supp$\hat{\varphi} \subseteq [-R, R]^N$. Let K be a convolution kernel such that $\|K * f\|_{L^p} \leq \|f\|_{L^p}$ for every $f \in L^p(\mathbb{R}^N)$. We have that*

$$\|K^\varphi \star a\|_{\ell^p} \leq A\|a\|_{\ell^p},$$

for every sequence $\{a(m)\}_{m\in\mathbb{Z}^N} \in \ell^p$.

Proof. Let the sequence $\{a(m)\}_{m\in\mathbb{Z}^N}$ be in ℓ^p, and define $f(x) = \sum_{m\in\mathbb{Z}^N} a(m)\varphi(x - m)$. Observe that the support of $\hat{f}$ is contained in the support of $\hat{\varphi}$, therefore f is of R-exponential type. By Lemma 1.4.6, we have that $f \in L^p(\mathbb{R}^N)$. On the one hand

$$(K^\varphi \star a)(l) = \sum_{m\in\mathbb{Z}^N} K^\varphi(l - m)a(m) = \sum_{m\in\mathbb{Z}^N} (K * \varphi)(l - m)a(m) \quad l \in \mathbb{Z}^N.$$

On the other hand

$$(K * f)(x) = (K * \sum_{m\in\mathbb{Z}^N} a(m)\varphi(\cdot - m)) = \sum_{m\in\mathbb{Z}^N} a(m)(K * \varphi)(x - m) \quad x \in \mathbb{R}^N.$$

Then $(K^\varphi \star a)(l) = (K * f)(l)$ if $l \in \mathbb{Z}^N$ and:

$$\|K^\varphi \star a\|_{\ell^p} = \|(K * f)^d\|_{\ell^p} \leq C\|K * f\|_{L^p} \leq C\|f\|_{L^p} \leq A\|a\|_{\ell^p}.$$

Where the first inequality is due to Lemma 1.4.5 and the last one due to Lemma 1.4.6. □

1.5 Extension of multipliers in $\mathbb{T}^N$ to $\mathbb{R}^N$

Carrying on with the relationship between continuous and discrete multipliers, in this section we will prove a theorem that we can think of as the reciprocal to the previous one. However, the hypotheses are a bit different. We must demand some different conditions in order to obtain the boundedness of the continuous convolution operator from the discrete one.

Theorem 1.5.1. *Let $1 \leq p < \infty$. Suppose that $\hat{\varphi}$ satisfies the following conditions:*

(i) supp$\hat{\varphi} \subseteq [-R, R]^N$, where $R < 1$.

(ii) There exists $\varepsilon > 0$ y $h \in C^\infty((-\varepsilon, \varepsilon)^N)$, $h \equiv 1$ in $[-\varepsilon/2, \varepsilon/2]^N$, such that $h/\hat{\varphi} \in M_p(\mathbb{R}^N)$.

Consider the dilated kernel $K_t(x) = t^{-N}K(t^{-1}x)$, where $t > 0$ and K is a convolution kernel. Then, inequality

$$\|K_t^\varphi \star a\|_{\ell^p} \leq \|a\|_{\ell^p},$$

for every $t > 0$ and $\{a(m)\}_{m\in\mathbb{Z}^N} \in \ell^p$, implies that

$$\|K * f\|_{L^p} \leq A\|f\|_{L^p},$$

where $f \in L^p(\mathbb{R}^N)$ and $A \leq M_p(h/\hat{\varphi})$.

Proof. Suppose that $f \in \mathcal{S}(\mathbb{R}^N)$ satisfies supp$\hat{f} \subseteq [-\delta, \delta]^N$ where $\delta < \varepsilon/2$ and $\delta < 1 - R$. By density of the Schwartz class in $L^p(\mathbb{R}^N)$, and by the dilatation property for $\mathbb{R}^N$ we can assume that supp$\hat{f} \subseteq [-\delta, \delta]^N$. So, if $f \in L^p(\mathbb{R}^N)$ and $g(x) = r^N f(rx)$ for $r > 0$, then

$$|(K_t * g)(x)| = r^N|(K_t * f)(rx)|.$$

Observe that $\hat{f} = \hat{f}h = (\hat{f}h/\hat{\varphi})\hat{\varphi}$; then for $x = n + u$, $n \in \mathbb{Z}^N$, $u \in [0, 1)^N$ we get that

$$\begin{aligned}\hat{f}(\xi)e^{2\pi ix\xi} &= ((\hat{f}h/\hat{\varphi})(\xi)e^{2\pi iu\xi})\hat{\varphi}(\xi)e^{2\pi in\xi} \\ &= \left(\sum_{k\in\mathbb{Z}^N} \left(\frac{\hat{f}h}{\hat{\varphi}}e^{2\pi iu\cdot}\right)(\xi + k)\right)\hat{\varphi}(\xi)e^{2\pi in\xi}.\end{aligned}$$

Where the last equality is obtained for our choice of δ and its relationship with R (the terms vanish when $k \neq 0$). Moreover, the previous series defines a function $P_u(\xi)$ whose Fourier coefficients are

$$a^u(m) = \int_{\mathbb{R}^N} (\hat{f}h/\hat{\varphi})(\xi)e^{2\pi iu\xi}e^{2\pi im\xi}d\xi.$$

From where we obtain

$$(C_t f)(x) = (D_t^{\varphi} a^u)(n), \quad \text{where } x = n + u.$$

This is,

$$(K_t * f)(x) = (K_t^{\varphi} \star a^u)(n), \quad \text{where } x = n + u.$$

Hence

$$\begin{aligned}\|K_t * f\|_{L^p}^p &= \int_{[0,1)^N} \sum_{n \in Z^N} |(K_t^{\varphi} \star a^u)(n)|^p du \leq \int_{[0,1)^N} \|a^u\|_{l^p}^p du \\ &= \int_{[0,1)^N} \sum_{m \in Z^N} \left| \int_{\mathbb{R}^N} (\hat{f} h/\hat{\varphi})(\xi) e^{2\pi i u \xi} e^{2\pi i m \xi} d\xi \right|^p du \\ &= \int_{\mathbb{R}^N} \left| \int_{\mathbb{R}^N} \hat{f}(\xi)(h/\hat{\varphi})(\xi) e^{2\pi i x \xi} d\xi \right|^p dx \leq M_p (h/\hat{\varphi})^p \|f\|_{L^p}^p.\end{aligned}$$

□

1.6 Application: boundedness of the discrete Hilbert transform

As a consequence of the last result, let us see that the discrete Hilbert transform is a bounded operator in $\ell^p(\mathbb{Z})$.

Consider the kernel associated to the Hilbert transform $K = PV(\frac{1}{\pi x})$, its Fourier multiplier is $m(\xi) = -\mathrm{i}\ \mathrm{sign}(\xi)$. Let φ be an even function with $\mathrm{supp}\widehat{\varphi} \subseteq [-R, R]$, $\widehat{\varphi} \in C^{\infty}(\mathbb{R})$, and $\widehat{\varphi}(0) = 1$. Then

$$K_t^{\varphi}(m) = (K_t * \varphi)(m) = \int_{\mathbb{R}} -\mathrm{i}\ \mathrm{sign}(\xi) \widehat{\varphi}(\xi) e^{2\pi i m \xi} d\xi = 2 \int_0^{+\infty} \widehat{\varphi}(\xi) \sin(2\pi m \xi) d\xi.$$

Applying integration by parts, we get that

$$K_t^{\varphi}(m) = \frac{1}{\pi m} + O\left(\frac{1}{m^2}\right).$$

From where we deduce that

$$(K_t^{\varphi} \star a)(n) = H^d a(n) + C a(n),$$

where H^d is the discrete Hilbert transform, defined as follows

$$H^d a(n) = \sum_{m \neq n} \frac{a(m)}{\pi(n - m)}.$$

So, if the Hilbert transform is bounded in $L^p(\mathbb{R})$ for $1 < p < \infty$, then, by the previous theorem, we deduce that the discrete operator associated to the Hilbert transform is bounded in $\ell^p(\mathbb{Z})$ for $1 < p < \infty$.

Acknowledgements

We want to thank Escuela-Taller Bernardo Cascales for providing us the great opportunity of introducing ourselves to the world of mathematical research. In particular, we thank professor Santiago Boza for his priceless help, firstly mentoring us during the process of the Escuela-Taller, and secondly in the revisions of this article. We also want to thank the organizers of XII Congreso del máster en investigación matemática y del Doctorado en matemáticas for bringing us the opportunity to publish this article.

References

[1] P. Auscher and M. J. Carro, *On relations between operators on $\mathbb{R}^N$, $\mathbb{T}^N$ and $\mathbb{Z}^N$.*, Studia. Math., **5(2)**: 27, 2021.

[2] Francisco Javier Duoandikoetxea Zuazo, *Análisis de Fourier*, Addison-Wesley, 1991.

[3] E. Stein and G. Weiss, *Introduction to Fourier analysis on euclidean spaces.*, Princeton University Press (1971).

The Analogy between Electromagnetic and Acoustic Waves

Guillem Fernández Rodríguez♯

(♯) Faculty of Mathematics, University of Valencia, guillem.fernandez@uv.es

1.1 Introduction

Electromagnetic and acoustic waves, while fundamentally different in their physical nature, share deep mathematical similarities under specific assumptions. Both are governed by hyperbolic partial differential equations that describe wave propagation through space and time. This analogy not only facilitates the transfer of intuition between fields but also enables shared numerical techniques for simulation.

Electromagnetic theory is classically described by Maxwell's equations, a set of coupled first-order differential equations introduced in the 19th century and now foundational to modern physics and engineering. These equations, discussed in detail in [1] and [3], provide a complete description of how electric and magnetic fields evolve and interact with matter. Their wave-like behavior in the absence of sources was a key step in identifying light as an electromagnetic phenomenon. The mathematical structure behind these laws is also explored from a theoretical physics and mathematical perspective in [2].

The parallel with acoustics becomes especially apparent in linear, isotropic, non-dispersive media, where both fields satisfy similar wave equations. The analogy has been explored in works such as [7] and further extended to elastic wave modeling as discussed in [6] and [4]. These analogies become particularly fruitful when developing numerical algorithms, as techniques like the Finite-Difference Time-Domain (FDTD) method, originally developed for electromagnetics by [5], can be adapted with minor modifications to simulate acoustic phenomena.

This work presents a pedagogical and numerical study of this analogy. We begin with a theoretical introduction to electromagnetic waves via Maxwell's equations, emphasizing the derivation of the wave equation and key concepts like transversality and the dispersion relation. Following this, we construct a numerical framework based on the FDTD method. We derive stability conditions, analyze numerical dispersion and phase velocity, and implement absorbing boundary conditions. Emphasis is placed on the one-dimensional case to explore essential concepts, then extended to two dimensions via Yee's scheme.

All simulations were developed using a custom Python package [8], which serves both as a learning tool and a computational platform. The ultimate goal is to illuminate how analogous mathematical structures enable cross-domain simulations and to encourage broader use of shared numerical methods in physics and engineering contexts.

1.2 Electromagnetic waves

1.2.1 Maxwell's equations

Maxwell's equations form the cornerstone of classical electromagnetism, encompassing the relationships between electric fields $\boldsymbol{E}$, magnetic fields $\boldsymbol{H}$, electric currents $\boldsymbol{J}$, and charge densities ρ. It is often also introduced a hypothetical magnetic current $\boldsymbol{M}$ useful for developing numerical methods. These equations describe how these quantities interact within space-time ($\mathbb{R}^4$) and are pivotal for understanding electromagnetic waves.

Maxwell's equations relating $\boldsymbol{E}$ and $\boldsymbol{H}$ in linear, isotropic, nondispersive, lossy materials are:

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon} \tag{1.1}$$

$$\nabla \cdot \boldsymbol{H} = 0 \tag{1.2}$$

$$\nabla \times \boldsymbol{E} = -\mu \frac{\partial \boldsymbol{H}}{\partial t} - M \tag{1.3}$$

$$\nabla \times \boldsymbol{H} = \epsilon \frac{\partial \boldsymbol{E}}{\partial t} + \boldsymbol{J} \tag{1.4}$$

where the currents can be decomposed into an independent source ($\boldsymbol{J}_{source}$, $\boldsymbol{M}_{source}$) and a loss due to the material conductivity (σ, σ^*)

$$\boldsymbol{J} = \boldsymbol{J}_{source} + \sigma \boldsymbol{E}, \quad \boldsymbol{M} = \boldsymbol{M}_{source} + \sigma^* \boldsymbol{H}$$

and the fields can be related to their corresponding flux densities with the proportions

$$\boldsymbol{D} = \epsilon \boldsymbol{E} = \epsilon_r \epsilon_0 \boldsymbol{E}, \quad \boldsymbol{B} = \mu \boldsymbol{H} = \mu_r \mu_0 \boldsymbol{H}$$

in linear, isotropic, nondispersive media. When not in isotropic media, we allow the electrical permittivity ϵ and magnetic permeability μ to be matrices indicating a different material behavior depending on the direction[1].

The constants ϵ_0, μ_0 are the respective properties in the void, which take the following values:

$$\epsilon_0 \approx 8.85 \times 10^{-12}, \quad \mu_0 = 4\pi \times 10^{-7}.$$

They are related to the speed of light c in the following way

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}}.$$

1.3 Numerical methods

The electromagnetic simulations showcased in this and the following sections were performed using a custom Python library, which is provided as supplementary material [8].

1.3.1 One-dimensional FDTD

The finite-difference time-domain method is used in computational electrodynamics in which finite differences are used to approximate Maxwell's differential equation and a discrete grid of the domain is used to solve for the electromagnetic field forward in time. To understand the different issues that arise with this problem we will consider the simplest wave equation, the one dimensional scalar wave equation traveling at the speed of light c

$$\frac{\partial^2 u}{\partial t^2} = \quad c^2 \frac{\partial^2 u}{\partial x^2}. \tag{1.5}$$

[1]This is because, in isotropic materials, there is no preferred direction.

We define a discrete grid of points $(x_i, t_n) = (i\Delta x, n\Delta t)$ for $i = 1, \dots N_x$and $n = 1, \dots, N_t$ where $\Delta, x\Delta t > 0$ define the size of the steps. To approximate the second order derivatives at each point, we use the second order central approximation

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{x_i,t_n} = \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{(\Delta x)^2}, \quad \frac{\partial^2 u}{\partial t^2}|_{x_i,t_n} = \frac{u_i^{n-1} - 2u_i^n + u_i^{n-1}}{(\Delta t)^2}$$

When introducing this approximations into the wave equation (1.5) we obtain the following scheme

$$u_i^{n-1} = \left(\frac{c\Delta t}{\Delta x}\right)^2 \left[u_{i+1}^n - 2u_i^n + u_{i-1}^n\right] + 2u_i^n - u_i^{n-1}$$

It is remarkable the case $c\Delta t = \Delta x$ in which the scheme is true for the exact solution at the sampled values.

When not under these circumstances, however, we can characterize the numerical error by computing the numerical dispersion relation.

Let us consider a wave solution in the exponential form $e^{j(wx-kx)}$, where we will denote the imaginary number $j := \sqrt{-1}$ for simplicity in this section and to avoid confusion with the index i. Similar to how the dispersion relation for this wave is $w^2 = c^2k^2$, we can try to obtain the relationship for the numerical wave that is obtained with the scheme. Let us consider a wave of this form with angular frequency w, and let us call the numerical wavenumber a possibly complex value $\overline{k}$, the wavenumber of the wave sampled at the discrete grid

$$u_i^n = e^{j(wn\Delta t - i\overline{k}\Delta x)} = e^{\overline{k}_{imag} i\Delta x} e^{j(wn\Delta t - i\overline{k}_{real}\Delta x)} \tag{1.6}$$

Introducing the numerical wave into the scheme, we obtain the we get the following relationship which we call the numerical dispersion relation

$$\tilde{k} = \frac{1}{\Delta x}\cos^{-1}\left\{1 + \left(\frac{\Delta x}{c\Delta t}\right)^2 [\cos(\omega\Delta t) - 1]\right\} = \frac{1}{\Delta x}\cos^{-1}\left\{1 + \left(\frac{1}{S}\right)^2 \left[\cos\left(\frac{2\pi S}{N_\lambda}\right) - 1\right]\right\} = \frac{1}{\Delta x} cos^{-1}(\xi)$$

where we define the Courant number $S = \frac{c\Delta y}{\Delta x}$ and the grid sampling resolution in space cells per free-space wavelength $N_{\lambda_0} = \frac{\lambda_0}{\Delta x}$. Using the magic timestep we obtain the dispersion relation, but in general the value of $\overline{k}$ differs from k for different resolutions. We can see how the numerical phase velocity $\overline{v}_p = \frac{w}{\overline{k}_{real}}$ and the amplitude multiplier $e^{\overline{k}_{imag}\Delta x}$ behave for different values of N_{λ_0} and $S = 0.5$ in Fig. 1.1 and Fig. 1.2 respectively.

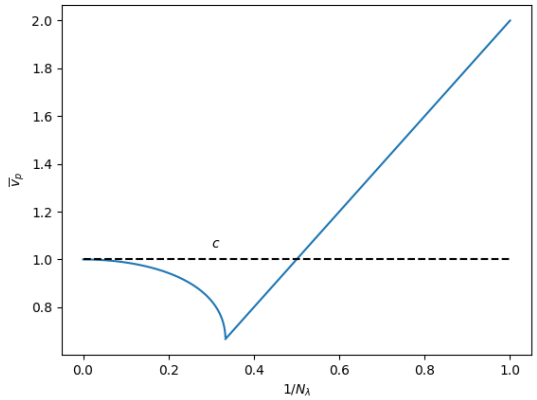


Figure 1.1: Numerical phase velocity for $S = 0.7$depending on $1/N_\lambda$.

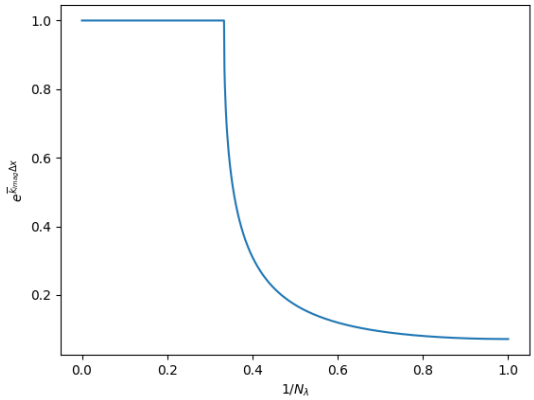


Figure 1.2: Amplitude multiplier for $S = 0.7$ depending on $1/N_\lambda$.

Evaluating different values of S, N can result in $\overline{k}$ being real or complex. It can be seen that the values of N_λ such that $\xi = -1$ gives us a series of thresholds between real and complex wavenumbers. If $S \leq 1$ (a condition for stability seen later) then for a big enough N_λ we have $\xi > -1$ such that we always obtain a real wavenumber.

In particular, for the minimum free-space wavelength $\lambda_{0,min}$ that can be sampled, its phase velocity is $\overline{v}_{p,max} = \frac{1}{S}c = \frac{\Delta x}{\Delta t}$. Again assuming $S \leq 1$, we have that the frequency components corresponding to this wavenumber might travel faster than c, and the maximum speed will be a spatial cell Δx per temporal cell Δt. In Fig. 1.3 due to a sparsely sampled discontinuity, we can see some ringing due to retarded propagation ($\overline{v}_p < c$) and a super-luminal component ahead of the right discontinuity ($\overline{v}_p > c$) for the cases $S = 0.99, 0.5$.

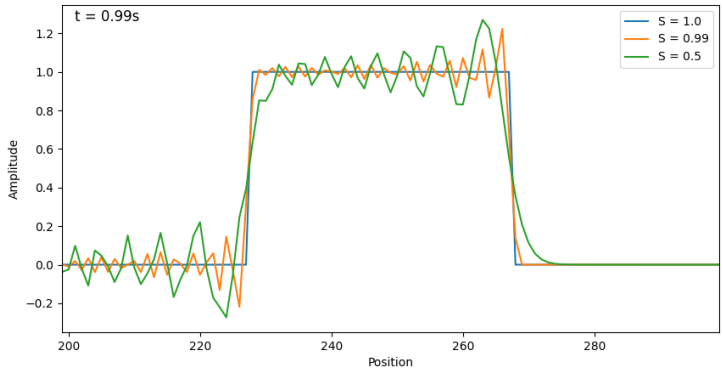


Figure 1.3: Comparison of rectangular pulse propagation using different Courant numbers shown for the same real time position.

Now we will study how S should be chosen for stability. We can follow a similar complex-frequency analysis for a numerical wave with numerical angular frequency $\overline{w} = \overline{w}_{real} + j\overline{w}_{imag}$ and numerical wavenumber $\overline{k}$.

$$u_i^n = e^{-\overline{w}_{imag} n\Delta t} e^{j(\overline{w}_{real} n\Delta t - i\overline{k}\Delta x)} \tag{1.7}$$

Then we obtain a similar relation

$$\bar{\omega} = \frac{1}{\Delta t}\cos^{-1}\{S^2[\cos(\tilde{k}\Delta x) - 1] + 1\}.$$

In this equation we see that $\xi < -1$ gives us a complex value, and that occurs if $S > 1$. In this case we see that the amplitude multiplier is

$$e^{-\overline{w}_{imag} n\Delta t} = (-\xi + \sqrt{\xi^2 - 1})^n$$

which will grow exponentially in time. In particular the maximum growth is achieved for the wavelength $\overline{\lambda} = 2\Delta x$. We can see in Figs 1.4-1.5 that a wave with this wavelength dominates the simulation after around 200 timesteps.

1.3.2 ABC for one dimension

It is often important to isolate a region of interest where some fenomena are happening without the interference from reflection at boundaries. To solve this numerous techniques have been devised like absorbing boundary conditions or the use of added lossy layers. We explore the case of a first order absorbing boundary condition for the one-dimensional wave.

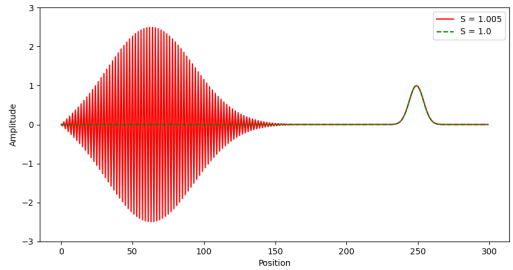


Figure 1.4: Evolution of right propagating Gaus-sian pulse at time $t = 2s$ for $S = 1, 1.005$ whichcorresponds to $nt = 199, 200$.

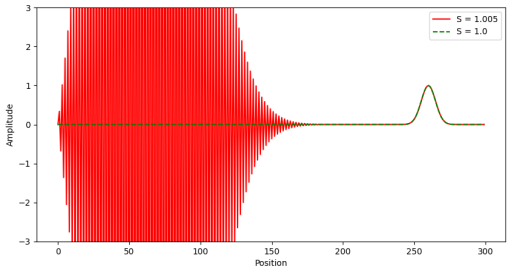


Figure 1.5: Evolution of right propagating Gaussian pulse at time $t = 2.11s$ for $S = 1, 1.005$ which corresponds to $nt = 210, 211$.

Solutions to the Eq. are of the form $u(x,t) = F(x - vt) + G(x + vt)$, in one dimension these are a sum of waves traveling to the right or to the left. Equivalently, the wave equation can be written as the product of advection equations

$$\frac{\partial^2 u}{\partial x^2} - \frac{1}{\mu\epsilon}\frac{\partial^2 u}{\partial t^2} = \left(\frac{\partial}{\partial x} - \frac{1}{\sqrt{\mu\epsilon}}\frac{\partial}{\partial t}\right)\left(\frac{\partial}{\partial x} + \frac{1}{\sqrt{\mu\epsilon}}\frac{\partial}{\partial t}\right) u = 0.$$

A wave traveling to the left $F(x + vt)$ is a solution of

$$\frac{\partial u}{\partial x} - \frac{1}{\sqrt{\mu\epsilon}}\frac{\partial u}{\partial t}.$$

We can then numerically solve this equation for the left boundary cell u_0^{n+1}, that way there is no reflection for left traveling waves at speed $\frac{1}{\sqrt{\mu\epsilon}}$.

A stable scheme can be found expanding this advection equation at $(x_{1/2}, t_{1/2})$. We use first order central differences for each partial derivative and approximate the terms at half steps that appear taking averages:

$$\sqrt{\mu\epsilon}\frac{\partial u}{\partial t}|_{1/2,n+1/2} \approx \sqrt{\mu\epsilon}\frac{\frac{u_0^{n+1}+u_1^{n+1}}{2} - \frac{u_0^n+u_1^n}{2}}{\Delta_t}, \quad \frac{\partial u}{\partial x}|_{1/2,n+1/2} \approx \sqrt{\mu\epsilon}\frac{\frac{u_1^{n+1}+u_1^{n+1}}{2} - \frac{u_0^n+u_0^n}{2}}{\Delta_t}$$

Then we obtain the scheme

$$u_0^{n+1} = u_1^n + \frac{\frac{S_c}{\sqrt{\mu_r\epsilon_r}} - 1}{\frac{S_c}{\sqrt{\mu_r\epsilon_r}} + 1}\left(u_1^{n+1} - u_0^n\right).$$

Similarly, we obtain an analogous scheme for the right boundary cell

$$u_N^{n+1} = u_{N-1}^n + \frac{\frac{S_c}{\sqrt{\mu_r\epsilon_r}} - 1}{\frac{S_c}{\sqrt{\mu_r\epsilon_r}} + 1}\left(u_{N-1}^{n+1} - u_N^n\right).$$

A wave can be traveling at a different speed due to numerical dispersion and fail to be absorbed, if this error is noteworthy higher order ABC conditions may be used. A simple second order ABC boundary condition can be found by applying the corresponding advection equation operator two times. In general, a second order ABC con be constructed in three dimensions [9] for the FDTD method. We see an example in Fig. 1.6 of using a first and second order ABC condition on a Gaussian pulse where a small reflection can be mitigated in the second case.

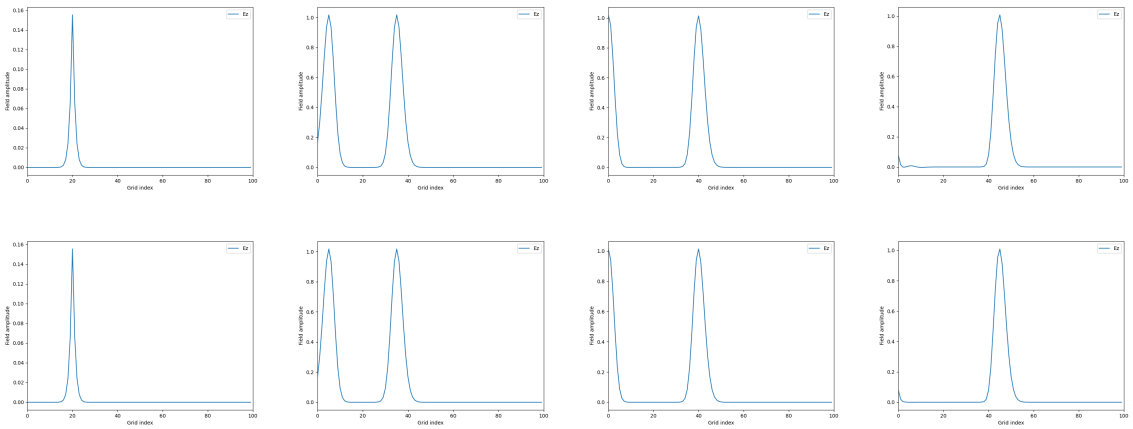


Figure 1.6: Evolution of the electromagnetic wave under two different boundary conditions. The first row corresponds to a first order ABC while the second one to a second order ABC, using a courant number $S = 0.5$ to produce a small numerical velocity error.

1.3.3 Yee's algorithm

We have given a simple scheme for solving the one-dimensional wave equation, to consider Maxwell's curl equations in three dimensions using FDTD, Yee's algorithm [5] was proposed. This algorithm consisted in developing a scheme for solving the curls equations (1.3), (1.4) using a staggered grid that automatically checks the divergence equations (1.1), (1.2).

Let us consider an equidistant but staggered in space and time grid as shown in Fig 1.7. We write the indices $(i, j, k) := (i\Delta x, j\Delta y, k\Delta z)$ ($\Delta := \Delta_x = \Delta_y = \Delta_z$ when all space steps coincide) and $n := n\Delta t$ so that $u^n_{i,j,k} = u(i\Delta x, j\Delta y, k\Delta z, n\Delta t)$.

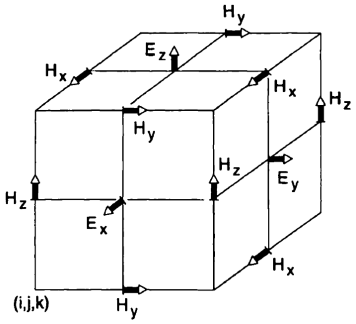


Figure 1.7: Yee's algorithm individual discretization cell.

The scheme for all field components is obtained analogously by using central differences and a semi-implicit approximation to fit time-steps into the correct grid.

We will consider the 2-dimensional TE_2 mode in which the fields are constant in the y direction and (H_2, E_1, E_3) form an independent set of equations. Let us obtain Yee's algorithm for this case, Maxwell's curl equations are:

$$\frac{\partial H_2}{\partial t} = \frac{1}{\mu}\left(\frac{\partial E_3}{\partial x} - \frac{\partial E_1}{\partial z} - (M_y + \sigma_m H_2)\right)$$
$$\frac{\partial E_1}{\partial t} = \frac{1}{\epsilon}\left(-\frac{\partial H_2}{\partial z} - (J_{source_1} + \sigma E_1)\right)$$
$$\frac{\partial E_3}{\partial t} = \frac{1}{\epsilon}\left(\frac{\partial H_2}{\partial x} - (J_{source_3} + \sigma E_3)\right)$$

Then simplifying Yee's algorithm in three dimensions:

$$
\begin{aligned}
H_2|_{i,k+1}^{n+1} &= D_a|_{i,k+1} H_2|_{i,k+1}^{n} \\
&\quad + D_b|_{i,k+1} \cdot \left(E_3|_{i+1/2,k+1}^{n+1/2} - E_3|_{i-1/2,k+1}^{n+1/2} + E_1|_{i,k+1/2}^{n+1/2} - E_1|_{i,k+3/2}^{n+1/2} - M_2|_{i-j+1/2,k+1}^{n+1/2} \Delta \right) \\
E_1|_{i,k+1/2}^{n+1/2} &= C_a|_{i,k+1/2} E_1|_{i,k+1/2}^{n-1/2} \\
&\quad + C_b|_{i,k+1/2} \cdot \left(H_2|_{i,k}^{n} - H_2|_{i,k+1}^{n} - J_{source_1}|_{i,k+1/2}^{n} \Delta \right) \\
E_3|_{i-1/2,k+1}^{n+1/2} &= C_a|_{i-1/2,k+1} E_z|_{i-1/2,k+1}^{n-1/2} \\
&\quad + C_b|_{i-1/2,k+1} \cdot \left(H_2|_{i,k+1}^{n} - H_2|_{i-1,k+1}^{n} - J_{source_3}|_{i-1/2,k+1}^{n} \Delta \right)
\end{aligned}
$$

where we define the medium coefficients

$$
\begin{aligned}
C_a|_{i,j,k} &= \left(1 - \frac{\sigma_{i,j,k} \Delta t}{2\epsilon_{i,j,k}} \right) / \left(1 + \frac{\sigma_{i,j,k} \Delta t}{2\epsilon_{i,j,k}} \right) \\
C_b|_{i,j,k} &= \left(\frac{\Delta t}{\epsilon_{i,j,k} \Delta} \right) / \left(1 + \frac{\sigma_{i,j,k} \Delta t}{2\epsilon_{i,j,k}} \right) \\
D_a|_{i,j,k} &= \left(1 - \frac{\sigma^*_{i,j,k} \Delta t}{2\mu_{i,j,k}} \right) / \left(1 + \frac{\sigma^*_{i,j,k} \Delta t}{2\mu_{i,j,k}} \right) \\
D_b|_{i,j,k} &= \left(\frac{\Delta t}{\mu_{i,j,k} \Delta} \right) / \left(1 + \frac{\sigma^*_{i,j,k} \Delta t}{2\mu_{i,j,k}} \right).
\end{aligned}
$$

A suitable grid for this case is shown in **??**. This scheme is then enough to make simple simulations efficiently, as an example we simulate the TE_2 mode affected by a Sinusoidal source in Fig. 1.8.

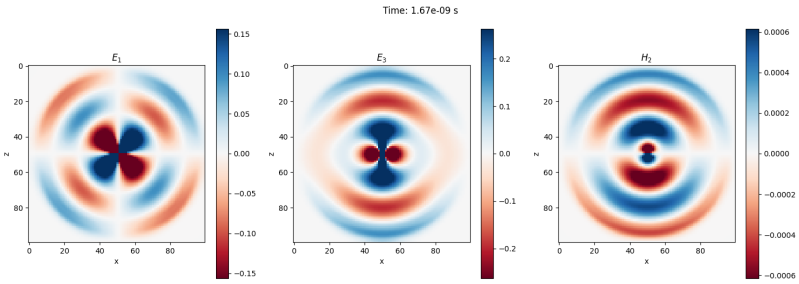


Figure 1.8: Two-dimensional simulation of TE_2 mode with a sinusoidal source at the center of the grid.

1.4 Conclusions

This work has explored the deep mathematical analogy between electromagnetic and acoustic wave phenomena, highlighting how similar underlying structures allow for shared analytical insights and numerical methods. Beginning with Maxwell's equations, we derived the corresponding wave equations and examined their numerical solution using the Finite-Difference Time-Domain (FDTD) method. A detailed analysis of numerical dispersion and stability conditions revealed the influence of discretization parameters on the propagation characteristics and the emergence of spurious modes.

We implemented absorbing boundary conditions (ABC) to mitigate reflections in finite domains and compared first- and second-order schemes. Extending these ideas, we introduced Yee's

algorithm as a generalization of FDTD in higher dimensions, culminating in a simulation of the TE_2 mode using a custom Python package.

By focusing on a pedagogical development of the theory and its numerical implementation, this work emphasizes how analogies between physical systems can inspire efficient cross-disciplinary modeling strategies. Such insights are valuable not only in theoretical investigations but also in practical applications across physics and engineering.

Financial disclosure

This research was partially supported by Spanish Ministerio de Ciencia, Innovación y Universidades PGC2018-095896-B-C22, by the internal research project ADMIREN of Universidad Internacional de La Rioja (UNIR) and was partially supported by Universitat Politècnica de València Contrato Predoctoral PAID-01-20-17 (UPV).

Bibliography

[1] D. J. Griffiths, *Introduction to Electrodynamics*, 4th ed. Cambridge: Cambridge University Press, 2013.

[2] T. A. Garrity, *Electricity and Magnetism for Mathematicians: A Guided Path from Maxwell's Equations to Yang–Mills*. Cambridge: Cambridge University Press, 2015. doi: 10.1017/CBO9781139939683.

[3] J. A. Kong, *Electromagnetic Wave Theory*. Cambridge: EMW Publishing, 1986.

[4] B. A. Auld, *Acoustic Fields and Waves in Solids*, vol. 1. Stanford University: John Wiley & Son, 1973.

[5] K. Yee, "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media," *IEEE Trans. Antennas Propag.*, vol. 14, no. 3, pp. 302–307, May 1966. doi: 10.1109/TAP.1966.1138693.

[6] W. C. Chew, M. S. Tong, and B. Hu, *Integral Equation Methods for Electromagnetic and Elastic Waves*, Synthesis Lectures on Computational Electromagnetics. Cham: Springer International Publishing, 2009. doi: 10.1007/978-3-031-01707-0.

[7] J. M. Carcione and F. Cavallini, "On the acoustic-electromagnetic analogy." [Online]. Available: `https://doi.org/10.1016/0165-2125(94)00047-9`

[8] G. Fernández-Rodríguez, "Electrology," GitHub repository, 2025. [Online]. Available: `https://github.com/introspective-swallow/Electromagnetic-Wave-Simulation`

[9] G. Mur, "Absorbing Boundary Conditions for the Finite-Difference Approximation of the Time-Domain Electromagnetic-Field Equations," *IEEE Trans. Electromagn. Compat.*, vol. EMC-23, no. 4, pp. 377–382, Nov. 1981. doi: 10.1109/TEMC.1981.303970.

Cómo Manipular el Comportamiento de las Masas en la Era de los Datos Masivos: El Algoritmo de Page Ranking de Google

Yu Zhang 1 ♭, Haijiao Kong 2 ♮, Sijia Guan 3 ♮

(♭) yzhang2@upv.edu.es
(♮) HKONG@posgrado.upv.es
(♮) siguan@alumni.uv.es

1.1 introducción

En las últimas décadas, la organización de la información en la web se ha convertido en un desafío esencial en la ciencia de datos, debido al crecimiento acelerado del volumen de contenido digital y a la necesidad de acceder a información pertinente de forma rápida y precisa. En este contexto, el algoritmo PageRank ha sido una de las herramientas más influyentes en el campo de la recuperación de información, al establecer un criterio cuantitativo de relevancia basado en la estructura de enlaces entre páginas web. Su formulación original, fundamentada en cadenas de Markov y navegación aleatoria, permitió transformar la web en un grafo dirigido donde cada nodo representa una página y cada enlace actúa como una votación implícita de importancia.
Sin embargo, el modelo clásico presenta limitaciones cuando se aplica a redes dinámicas, como las redes sociales o los sistemas de publicación científica, donde los enlaces evolucionan con el tiempo y la actualidad del contenido influye directamente en su relevancia. En tales entornos, un modelo estático tiende a privilegiar páginas históricamente consolidadas, lo que reduce su capacidad para captar información emergente o contenidos recientemente generados pero de alto impacto.
Ante esta problemática, el presente trabajo propone una extensión del algoritmo PageRank mediante la incorporación explícita de un factor temporal en la matriz de transición. En particular, se introduce un coeficiente de ponderación basado en un decaimiento exponencial en función del tiempo de creación del enlace:

$$T_{ij}(t) = e^{-\lambda(t_{\text{actual}} - t_{ij})},$$

donde t_{ij} representa la fecha en que se estableció el enlace entre las páginas i y j, y λ es un parámetro de sensibilidad temporal. Esta modificación busca reforzar la visibilidad de contenidos recientes sin ignorar la estructura global del grafo, mejorando la detección de nodos emergentes en entornos donde la actualidad es un atributo relevante. Además de presentar esta mejora metodológica, se analizan de forma crítica los efectos positivos y negativos del algoritmo PageRank. Por un lado, se estudian sus beneficios en términos de eficiencia en la búsqueda, incentivo a la producción de contenido de calidad y democratización del acceso al conocimiento. Por otro, se discuten sus efectos colaterales, como la formación de burbujas informativas, la monopolización del posicionamiento digital y la propagación de desinformación. Finalmente, se

exploran alternativas híbridas como SALSA y estrategias basadas en aprendizaje automático, con el fin de mitigar las limitaciones observadas y avanzar hacia sistemas de clasificación más justos, adaptativos y robustos.

1.2 Fundamentos teóricos del algoritmo PageRank

El algoritmo PageRank se fundamenta en la teoría de grafos, ya que modela la Web como un grafo dirigido donde las páginas web son nodos y los hipervínculos entre ellas son aristas dirigidas. Esta sección introduce los conceptos fundamentales que permiten comprender la lógica estructural del algoritmo, tales como los grafos, los tipos de grafos, matrices asociadas, caminos, ciclos y componentes conexos.

1.2.1 Conceptos básicos de grafos

Un grafo dirigido (o dígrafo) se define como un par $G = (V, E)$, donde V es un conjunto finito de vértices (nodos) y E es un conjunto de aristas dirigidas, es decir, pares ordenados (v_i, v_j) que representan una conexión dirigida desde el nodo v_i al nodo v_j.
En el contexto de la Web, cada página se modela como un nodo $v_i \in V$, y un enlace de una página i hacia una página j se modela como una arista $(v_i, v_j) \in E$.
El *grado de salida* (outdegree) de un nodo v_i es el número de aristas que salen de él. El *grado de entrada* (indegree) es el número de aristas que apuntan hacia él.

1.2.2 Matriz de adyacencia y matriz de transición

La estructura de enlace entre las páginas se representa formalmente mediante una matriz de adyacencia $A = [a_{ij}]$, donde:

$$a_{ij} = \begin{cases} 1 & \text{si hay un enlace de la página } i \text{ a la página } j \\ 0 & \text{en otro caso} \end{cases}$$

A partir de esta matriz, se construye la **matriz de transición** $H = [h_{ij}]$, donde cada elemento representa la probabilidad de transición de un "navegante aleatorio" de una página a otra:

$$h_{ij} = \begin{cases} \frac{1}{d_i} & \text{si } a_{ij} = 1 \text{ y } d_i > 0 \\ 0 & \text{en otro caso} \end{cases}$$

donde d_i es el número de enlaces salientes de la página i (su grado de salida).

1.2.3 Conectividad y componentes conexos

Un concepto importante en el análisis de grafos es el de conectividad. Un grafo es *fuertemente conexo* si existe un camino dirigido desde cualquier nodo a cualquier otro nodo. En la Web real, esto no siempre se cumple, por lo que es necesario considerar la existencia de nodos colgantes y componentes de sumidero, lo que más adelante se aborda en la construcción de la matriz de Google.

1.2.4 Ejemplo ilustrativo

Considérese un pequeño grafo dirigido con tres páginas A, B y C, y los siguientes enlaces: - $A \to B$, $A \to C$, $B \to C$, C no enlaza a ninguna página.

Esto genera una matriz de adyacencia:

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

y una matriz de transición H:

$$H = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Obsérvese que la fila correspondiente al nodo C es cero: es un *nodo colgante*, lo que rompe la propiedad de estocasticidad (las filas deben sumar 1). Para solucionar este problema, se redefine la matriz reemplazando las filas vacías con una distribución uniforme, lo que será tratado en la siguiente sección.

1.3 Modelo de PageRank

El modelo de PageRank se basa en una analogía con el comportamiento de un navegante aleatorio que recorre la Web a través de los hiperenlaces. Este comportamiento se modela mediante una caminata aleatoria sobre un grafo dirigido que representa la estructura de Internet.

1.3.1 Modelo de caminata aleatoria

Supongamos que un usuario comienza en una página web aleatoria y, en cada paso, hace clic en uno de los enlaces salientes de la página actual, elegidos con igual probabilidad. Este proceso se puede modelar como una cadena de Markov homogénea sobre el grafo de la Web, donde el estado del sistema es la página en la que se encuentra el usuario en un momento dado.
Formalmente, sea n el número total de páginas, y definamos una matriz de transición $H = [h_{ij}]$ tal que:

$$h_{ij} = \begin{cases} \frac{1}{d_i} & \text{si existe un enlace de la página } i \text{ a la página } j \\ 0 & \text{en otro caso} \end{cases}$$

donde d_i es el número de enlaces salientes de la página i.

1.3.2 Definición del vector PageRank

Sea $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ un vector de distribución de probabilidad sobre las páginas web. La ecuación fundamental de PageRank es:

$$\pi = \pi H$$

es decir, π es el vector propio izquierdo de la matriz H asociado al valor propio 1. Este vector representa la distribución estacionaria de la cadena de Markov, es decir, la probabilidad a largo plazo de que el navegante se encuentre en cada página.

1.3.3 Problemas prácticos del modelo simple

El modelo descrito anteriormente enfrenta dos problemas fundamentales cuando se aplica a la Web real:

- **Nodos colgantes (Dangling nodes)**: Son páginas sin enlaces salientes, es decir, nodos con $d_i = 0$. En ese caso, la fila i de H es un vector nulo, lo cual impide que H sea estocástica.

- **Componentes de sumidero (Sink components)**: Son subconjuntos del grafo donde todos los enlaces apuntan dentro del propio componente, y no hay enlaces salientes hacia el resto del grafo. Esto provoca que una caminata aleatoria pueda quedar atrapada permanentemente en dichos componentes.

1.3.4 Solución: matriz de Google

Para resolver estos problemas, se define la **matriz de Google** G como una combinación convexa entre la matriz H corregida (denotada S) y una matriz de "teletransporte" E:

$$G = \alpha S + (1 - \alpha)E$$

donde:

- S es la matriz de transición con filas correspondientes a nodos colgantes reemplazadas por el vector uniforme $u = \frac{1}{n}(1, 1, \ldots, 1)$.
- E es una matriz donde todas las filas son iguales a u, es decir, $E = ue^T$.
- $\alpha \in (0, 1)$ es el **factor de amortiguamiento** o *damping factor*, usualmente $\alpha = 0.85$.

1.3.5 Interpretación probabilística

La interpretación de esta fórmula es que, con probabilidad α, el usuario sigue un enlace de la página actual (modelo estructural), y con probabilidad $1 - \alpha$ elige aleatoriamente cualquier otra página (teletransporte). Esto garantiza que el sistema sea irreducible y aperiódico, lo que asegura la existencia y unicidad de una distribución estacionaria π.

1.3.6 Cálculo iterativo del PageRank

Dado que el número de páginas web es enorme, el vector π se calcula mediante un algoritmo iterativo conocido como el *método de potencias*:

$$\pi^{(k+1)} = \pi^{(k)} G$$

Se parte de un vector inicial (por ejemplo, $\pi^{(0)} = \frac{1}{n}(1, 1, \ldots, 1)$), y se itera hasta que se cumpla un criterio de convergencia:

$$\|\pi^{(k+1)} - \pi^{(k)}\| < \varepsilon$$

1.3.7 Ejemplo numérico simple

Dado el grafo dirigido con los nodos a, b, c y las siguientes aristas:

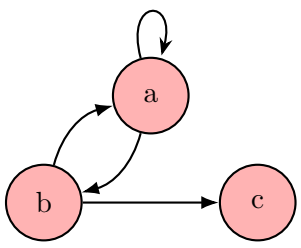


Figure 1.1: ejemplo

Paso 1: Construcción de la matriz de adyacencia
La matriz de adyacencia A para el grafo dado es:

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

Paso 2: Cálculo de la probabilidad uniforme y obtención de la nueva matriz S
Calculamos el vector de probabilidad uniforme u. Dado que $n = 3$ (el número de nodos en el grafo), entonces:

$$u = \frac{1}{n} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}$$

Reemplazando la fila correspondiente al nodo colgante, obtenemos la matriz S:

$$S = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

Paso 3: Obtención de la nueva matriz G
Tomemos el factor de amortiguamiento $a = 0.85$, entonces $1 - a = 0.15$. La matriz de teletransporte E, dado que $n = 3$, es:

$$E = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

Según la fórmula $G = aS + (1 - a)E$, calculamos la matriz Google G:

$$\begin{aligned} G &= 0.85S + 0.15E \\ &= \begin{bmatrix} 0.9 & 0.9 & 0.05 \\ 0.9 & 0.05 & 0.9 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \end{aligned}$$

Paso 4: Iteración del cálculo del PageRank

Supongamos que el vector inicial de valores de PageRank $\pi_0 = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}$. En la primera iteración, calculamos π_1 como sigue:

$$\pi_1 = \pi_0 G = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 0.9 & 0.9 & 0.05 \\ 0.9 & 0.05 & 0.9 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

Se deben continuar con múltiples iteraciones hasta que el vector de valores de PageRank converja (es decir, la diferencia entre los resultados de dos iteraciones consecutivas sea muy pequeña).

1.4 Impacto Socioeconómico y Crítico del Algoritmo PageRank

1.4.1 Influencia en el Comercio Electrónico y la Dinámica de Mercado

Correlación directa entre tráfico y tasa de conversión

Los datos empíricos muestran que un PageRank alto incrementa significativamente la visibilidad de las páginas de productos. Por ejemplo:

- Productos en el **top 10**: 5,000–10,000 visitas/mes, tasa de conversión del 10
- Productos en posiciones **51–100**: 500–1,000 visitas/mes, tasa de conversión 5

Table 1.1: Volumen de visitas y tasa de conversión según ranking

Ranking	Visitas/mes	Tasa de conversión
1–10	5,000–10,000	10%–20%
11–50	1,000–3,000	5%–10%
51–100	500–1,000	2%–5%

Desigualdad de recursos y competencia desleal

Las grandes empresas dominan los primeros lugares en los motores de búsqueda mediante el uso intensivo de técnicas de SEO, marginando a pequeños comerciantes. Esta dinámica consolida monopolios digitales y restringe la diversidad de opciones disponibles para los consumidores.

1.4.2 Difusión de Información y Formación de Tendencias Sociales

Aceleración de la propagación de contenido

El PageRank también desempeña un rol clave en la velocidad de difusión de noticias e información en redes sociales. Contenido con PageRank alto (por ejemplo, **top 10**) puede alcanzar hasta **500 compartidos/hora**, comparado con solo 50 en rankings bajos (ver Tabla 1.2).

Table 1.2: Velocidad de difusión según PageRank

PageRank	**Compartidos/hora**
Alto (1–10)	500
Medio (4–7)	200
Bajo (1–3)	50

Configuración de tendencias y efecto burbuja

PageRank amplifica narrativas dominantes al priorizar el contenido más enlazado. Estudios demuestran que el 90

1.4.3 Ventajas Técnicas y Sociales del Algoritmo PageRank

Eficiencia en la búsqueda de información

PageRank introdujo un modelo robusto y eficiente para ordenar páginas web combinando cadenas de Markov con un modelo de navegación aleatoria. La fórmula generalizada es:

$$P = (1 - d)E + dMTP, \tag{1.1}$$

donde:

- P es el vector de importancia de las páginas.
- d es el factor de amortiguación.

- M es la matriz de transición.
- T pondera la antigüedad de los enlaces.

Un refinamiento incluye un factor temporal dinámico:

Impulso a la creación de contenido de calidad

El algoritmo recompensa contenidos con enlaces entrantes relevantes y confiables. Esto ha incentivado la producción de contenidos bien referenciados, especialmente en entornos académicos, donde también se utilizan variantes como el *Weighted PageRank*, que incorpora la autoridad de la fuente y la interacción de los usuarios.

Democratización del acceso al conocimiento

En contextos educativos, PageRank y sus derivados como SALSA han facilitado la redistribución del acceso a información en regiones menos desarrolladas, permitiendo que recursos locales tengan visibilidad en redes más amplias.

1.4.4 Impactos Negativos y Alternativas Algorítmicas

Formación de burbujas informativas

La estructura del algoritmo tiende a reforzar comunidades ideológicas cerradas. Esto puede medirse con la modularidad Q:
SALSA y otros algoritmos de redistribución buscan aumentar la diversidad intercomunitaria penalizando nodos sobreconectados en clústeres cerrados.

Desigualdad de acceso y competencia desleal

El diseño original favorece a grandes actores con recursos para optimizar su SEO. Esto limita la entrada de nuevos competidores. Aunque existen intentos de incluir pesos personalizados, el sesgo estructural persiste.

Propagación de información falsa

PageRank no evalúa la veracidad de los contenidos, lo que ha permitido la difusión de noticias falsas. Algunas soluciones incluyen el uso de redes neuronales para detectar patrones semánticos no confiables. No obstante, estas técnicas enfrentan retos éticos y técnicos.

1.4.5 Propuestas de mejora y futuras líneas de investigación

- **Modelos híbridos**: Algoritmos como SALSA o variantes con retroalimentación comunitaria aumentan la diversidad informativa.
- **Ponderación temporal**: Ajustar los pesos de los enlaces en función de su antigüedad para reducir el sesgo hacia nodos históricos.
- **Integración de confiabilidad**: Incorporar medidas de credibilidad y verificación algorítmica antes de la propagación.

1.5 Conclusion

El algoritmo PageRank de Google ha tenido un impacto profundo en la era de la información, mejorando la eficiencia en la recuperación de datos y fomentando el desarrollo del comercio electrónico. Sin embargo, también ha generado una serie de problemas relacionados con la equidad en el acceso a la información, la concentración de poder informativo y la ética en la difusión de noticias. A medida que la tecnología avanza, es fundamental continuar investigando y perfeccionando el algoritmo para lograr un equilibrio entre precisión, diversidad y equidad en la información. Además, se deben considerar los aspectos éticos en el desarrollo de estos sistemas para garantizar un entorno de información saludable y justo.

References

[1] S. Brin and L. Page, *The anatomy of a large-scale hypertextual Web search engine*, Computer Networks and ISDN Systems, **30(1–7)**: 107–117, 1998.

[2] T. H. Haveliwala, *Topic-sensitive PageRank*, Proceedings of the 11th International Conference on World Wide Web, pp. 517–526, 2002.

[3] J. M. Kleinberg, *Authoritative sources in a hyperlinked environment*, Journal of the ACM, **46(5)**: 604–632, 1999.

[4] E. Pariser, *The Filter Bubble: What the Internet is Hiding from You*, Penguin UK, 2011.

[5] H. Allcott and M. Gentzkow, *Social media and fake news in the 2016 election*, Journal of Economic Perspectives, **31(2)**: 211–236, 2017.

Iterates of composition operators on global spaces of ultradifferentiable functions

Héctor Ariza Remacha ♭, (♭) harirem@upvnet.upv.es

Joint work with:
Carmen Fernández[1], Department of Mathematical Analysis, University of Valencia.
Antonio Galbis[2], Department of Mathematical Analysis, University of Valencia.

1.1 Introduction

This a continuation of the talk given in the XI Congreso del Máster en Investigación Matemática y Doctorado en Matemáticas, entitled Composition operators on Gelfand-Shilov classes. For the sake of self-containment, we are going to repeat some parts of the previous talk.

Given a function $\psi : \mathbb{K}^N \to \mathbb{K}^N$ and a suitable family of functions X defined on $\mathbb{K}^N$, the composition operator associated with ψ on X is $C_\psi f = f \circ \psi$, for every $f \in X$. Given a topological vector space X, a relevant and not always obvious problem is to find necessary and sufficient conditions on ψ for $C_\psi(X) \subset X$ and $C_\psi : X \to X$ to hold some property such as continuity, power boundedness or mean ergodicity.

The composition operators and their properties have been studied in several function spaces such as in the space of holomorphic functions, of real analytic functions (see for instance [11, 20, 21] and the references therein), of smooth functions (see for instance, [19] and the references therein) and also in the Schwartz class (see for instance [13, 14, 15]). There are many classical problems related to this operator (see for instance [9] and the references therein).

It is well-known the following classical result:

Theorem 1.1.1 (Borel's theorem) *Any formal series $\sum_{j=0}^{\infty} c_j x^j$ is the Taylor series of a smooth function defined in an open neighborhood of the origin. In other words, the Borel map $B : C^\infty(\mathbb{R}) \to \mathbb{R}^{\mathbb{N}}$ defined by $B(f) = (f^{(j)}(0))_j$ is surjective.*

From this, we see at once that the space of smooth functions is much "bigger" than the space of real analytic functions. It would be interesting to find intermediate families of functions to parametrize the gap existing between both. Are there spaces between one and the other that have "nice" properties and for which the composition operator is worth studying? It turns out that there is a family of classes that gives an affirmative answer to the previous question:

[1]e-mail: carmen.fdez-rosell@uv.es
[2]e-mail: antonio.galbis@uv.es

Definition 1.1.1 *The Gevrey class (of index $s \geq 1$) $G^s(\mathbb{R})$ is defined as the set of smooth functions f such that for every compact subset K there exists a $C = C_{K,f} > 0$ satisfying that*

$$\sup_{x \in K} |f^{(j)}(x)| \leq C^{j+1} (j!)^s$$

for all $j \in \mathbb{N}_0$.

On the one hand, if f is a real analytic function it is easy to see using Cauchy's integral formula that for every compact subset K there exists a $C = C_{K,f} > 0$ satisfying that

$$\sup_{x \in K} |f^{(j)}(x)| \leq C^{j+1} j!$$

for all $j \in \mathbb{N}_0$. So $f \in G^1(\mathbb{R})$. On the other hand, if $f \in G^1(\mathbb{R})$ then, it holds by Cauchy–Hadamard theorem that f is real analytic. So $G^1(\mathbb{R}) = \mathcal{A}(\mathbb{R})$. It is trivial that $G^s(\mathbb{R}) \subset G^{s+h}(\mathbb{R}) \subset C^\infty(\mathbb{R})$, for all $s \geq 1, h > 0$. However, the following inclusions ae strict: $\bigcup_{s \geq 1} G^s(\mathbb{R}) \subsetneqq C^\infty(\mathbb{R})$ (this is an easy consequence of Borel's theorem) and $\mathcal{A}(\mathbb{R}) \subsetneqq \bigcap_{s>1} G^s(\mathbb{R})$ (this result is not easy to obtain, see, for instance, [7]).

These classes appeared for the first time in the work of Gevrey, who measured the growth behaviour of such functions in terms of a weight sequence $(M_p)_p$, which is $((p!)^s)_p$, $s \geq 1$, in the Gevrey case and which satisfies certain technical conditions in the general case of $(M_p)-$ultra-differentiable classes. Later Beurling [5] pointed out that one can also use weight functions ω to measure the smoothness of functions with compact support by the decay properties of their Fourier transform. This method was modified by Braun, Meise, and Taylor in [10], who showed that also these classes can be defined by the decay behaviour of their derivatives, if one uses the Young conjugate of the function $t \to \omega(e^t)$. Meise and Taylor in [10] showed that under rather strong conditions both ways lead to the same class. But in general there are classes defined in one way which cannot be defined in the other way. For more details on the exact relationship between both approaches see [8]. The composition operator on the case of $\omega-$ultradifferentiable functions has been studied in [12].

Recall that the Schwartz class $\mathcal{S}(\mathbb{R})$ consists of those smooth functions $f : \mathbb{R} \to \mathbb{R}$ with the property that

$$p_n(f) := \sup_{x \in \mathbb{R}} \sup_{1 \leq j \leq n} (1 + x^2)^n |f^{(j)}(x)| < \infty$$

for each $n \in \mathbb{R}$. It turns out that $\mathcal{S}(\mathbb{R})$ is a Fréchet space when it is endowed with the topology generated by the sequence of seminorms $(p_n)_{n \in \mathbb{N}}$.

The Gevrey classes are made of functions whose derivatives verify certain local estimations, whereas the Schwartz class is made of functions whose derivatives asymptotically decrease fast "enough". Combining both the Gevrey classes and the Schwartz class, we define the following well-known family of smooth functions (originally introduced in [16], see [18] and the references therein for further information):

Definition 1.1.2 *The Gelfand-Shilov space $\Sigma_d(\mathbb{R})$, with $d > 1$, consists of those functions $f \in C^\infty(\mathbb{R})$ such that, for each $h > 0$:*

$$\sup_{x \in \mathbb{R}} \sup_{j, \ell \in \mathbb{N}_0} \frac{|x^\ell f^{(j)}(x)|}{h^{j+\ell} [(j+\ell)!]^d} < +\infty.$$

We can define more general families of ultra-differentiable functions by changing the sequence $([(j+\ell)!]^d)_{j,\ell}$ above for a suitable sequence $(M_{j+\ell})_{j,\ell}$, called weight sequence.

Definition 1.1.3 *The space $\mathcal{S}_{(M_p)}(\mathbb{R})$ associated to the weight sequence $(M_p)_{p\in\mathbb{N}_0}$ consists of those functions $f \in C^\infty(\mathbb{R})$ such that, for each $h > 0$:*

$$\sup_{x\in\mathbb{R}} \sup_{j,\ell\in\mathbb{N}_0} \frac{|x^\ell f^{(j)}(x)|}{h^{j+\ell} M_{j+\ell}} < +\infty.$$

As we have hinted above, we can define the following global class of smooth functions using weight functions instead of weight sequences in the following way:

Definition 1.1.4 *A continuous increasing function $\omega : [0,\infty[\longrightarrow [0,\infty[$ is called a weight if it satisfies:*

(α) *there exists $K \geq 1$ with $\omega(2t) \leq K(\omega(t)+1)$ for all $t \geq 0$,*

(β) $\displaystyle\int_0^\infty \frac{\omega(t)}{1+t^2}\, dt < \infty$,

(γ) $\log(1+t^2) = o(\omega(t))$ *as t tends to ∞,*

(δ) *$\varphi_\omega : t \to \omega(e^t)$ is convex.*

The function ω is extended to $\mathbb{R}$ as $\omega(x) = \omega(|x|)$. The *Young conjugate* $\varphi_\omega^* : [0,\infty[\longrightarrow \mathbb{R}$ of φ_ω is defined by

$$\varphi_\omega^*(s) := \sup\{st - \varphi_\omega(t) :\ t \geq 0\},\ s \geq 0.$$

Then φ_ω^* is convex, $\varphi_\omega^*(s)/s$ is increasing and $\lim_{s\to\infty} \dfrac{\varphi_\omega^*(s)}{s} = +\infty$. Moreover, for every $A > 0$, $\lambda > 0$ there is $C > 0$ such that

$$A^j j! \leq C e^{\lambda\varphi_\omega^*(\frac{j}{\lambda})}$$

for each $j \in \mathbb{N}_0$. The weight function ω is said to be a *strong weight* if

(ε) there exists a constant $C \geq 1$ such that for all $y > 0$ the following inequality holds

$$\int_1^\infty \frac{\omega(yt)}{t^2}\, dt \leq C\omega(y) + C. \tag{1.1}$$

Definition 1.1.5 *Let ω be a weight function. The Gelfand-Shilov space of Beurling type $\mathcal{S}_{(\omega)}(\mathbb{R})$ consists of those functions $f \in L^1(\mathbb{R})$ with the property that $f, \widehat{f} \in C^\infty(\mathbb{R})$ and*

$$q_{\lambda,j}(f) := \max\big(\sup_{x\in\mathbb{R}} |f^{(j)}(x)|e^{\lambda\omega(x)},\ \sup_{\xi\in\mathbb{R}} |\widehat{f}^{(j)}(\xi)|e^{\lambda\omega(\xi)}\big) < +\infty$$

for every $\lambda > 0, j \in \mathbb{N}_0$.

$\mathcal{S}_{(\omega)}(\mathbb{R})$ is a Fréchet space with different equivalent systems of seminorms. In particular we shall use the families of seminorms (see for instance [2, 6])

$$p_\lambda(f) := \sup_{j,k\in\mathbb{N}_0} \sup_{x\in\mathbb{R}} |x^k f^{(j)}(x)|e^{-\lambda\varphi_\omega^*(\frac{j+k}{\lambda})},\quad \lambda > 0$$

or

$$\pi_{\lambda,\mu}(f) := \sup_{j\in\mathbb{N}_0} \sup_{x\in\mathbb{R}} |f^{(j)}(x)|e^{-\lambda\varphi_\omega^*(\frac{j}{\lambda})+\mu\omega(x)},\quad \lambda > 0, \mu > 0.$$

Let $d > 1$ be given. The Gelfand-Shilov space $\Sigma_d(\mathbb{R})$ is

$$\Sigma_d(\mathbb{R}) = \mathcal{S}_{(M_p)}(\mathbb{R}) = \mathcal{S}_{(\omega)}(\mathbb{R}),$$

where

$$M_p = p!^d, \quad \omega(t) = t^{\frac{1}{d}}.$$

The two approaches described are nonequivalent, it is strictly more general the one that uses weight functions. This is one of the reasons we favor working with weight functions ω instead of weight sequences $(M_p)_p$. In [3] we started the investigation of composition operators on the Gelfand-Shilov space of Beurling type $\mathcal{S}_{(\omega)}(\mathbb{R})$. We proved in particular the following result:

Theorem 1.1.2 *Let ψ be a polynomial of degree $N > 1$. Then for every $d \leq d' < \frac{2N-1}{N} d$ there is $f \in \Sigma_d(\mathbb{R})$ such that $f \circ \psi \notin \Sigma_{d'}(\mathbb{R})$. In particular, for any polynomial ψ of degree greater than one and $d \leq d' < \frac{3}{2} d$ there is $f \in \Sigma_d(\mathbb{R})$ such that $f \circ \psi \notin \Sigma_{d'}(\mathbb{R})$.*

Observe that $\frac{3}{2} \leq \frac{2N-1}{N} < 2.$ for all $N \in \mathbb{N}$. In particular, we have the following result:

Corollary 1.1.3 *Let ψ be a polynomial of degree $N > 1$ and $d > 1$. Then, $C_\psi(\Sigma_d(\mathbb{R})) \not\subset \Sigma_d(\mathbb{R})$.*

So we cannot iterate the composition operator associated with a polynomial of degree greater than one in the usual sense. However, we have the following positive result:

Theorem 1.1.4 *Let ω be a subadditive weight, $\sigma(t) = \omega(t^{\frac{1}{2}})$ and ψ a non constant polynomial. Then $f \circ \psi \in \mathcal{S}_{(\sigma)}(\mathbb{R})$ for every $f \in \mathcal{S}_{(\omega)}(\mathbb{R})$. In particular, if $d > 1$ and ψ is a non constant polynomial then $C_\psi : \Sigma_d(\mathbb{R}) \to \Sigma_{2d}(\mathbb{R})$ is continuous.*

Noticing that $C_\psi^2 = C_{\psi \circ \psi}$ and that $\psi \circ \psi$ is also a polynomial, we can still iterate the composition operator associated with polynomials and study its dynamics. More explicitly, if we denote $\psi_m = \overbrace{\psi \circ ... \circ \psi}^{\text{m times}}$, for all $m \in \mathbb{N}$ then we have that $C_\psi^m \equiv C_{\psi_m} : \Sigma_d(\mathbb{R}) \to \Sigma_{2d}(\mathbb{R})$ is continuous for all $m \in \mathbb{N}$. And it turns out that $2d$ is the optimal index for studying the dynamics of composition operators associated with polynomials of degree greater than one acting on the Gelfand-Shilov space $\Sigma_d(\mathbb{R})$ with an index $d > 1$.

Let us recall some concepts related to the study of the dynamics of continuous operators. In particular, the concept of power boundedness and mean ergodicity of operators. An operator $T : X \to X$ is said to be power bounded if $\{T^n : n \in \mathbb{N}\}$ is an equicontinuous set. If X is a Fréchet space then T is power bounded if and only if $\{T^n(x) : n \in \mathbb{N}\}$ is bounded for each $x \in X$. A closely related concept to power boundedness is that of *mean ergodicity*. Given $T \in L(X)$, the Cesàro means of T are defined as $T_{[n]} = \sum_{k=1}^{n} T^k/n$. T is said to be mean ergodic when $T_{[n]}$ converges to an operator P. Clearly, if T is mean ergodic then $\lim_{n\to\infty} \frac{T^n(x)}{n} = 0$ for each $x \in E$.

In [14], it was proved the following result for the Schwartz space $\mathcal{S}(\mathbb{R})$:

Theorem 1.1.5 Let φ be a polynomial with degree greater than or equal to two. Then, the following are equivalent:

(1) $C_\varphi : \mathcal{S}(\mathbb{R}) \to \mathcal{S}(\mathbb{R})$ is power bounded.

(2) $C_\varphi : \mathcal{S}(\mathbb{R}) \to \mathcal{S}(\mathbb{R})$ is mean ergodic.

(3) The degree of φ is even and it has no fixed points.

We will see that the same result holds in the setting of Gelfand-Shilov classes.

1.2 Main results

Let us begin considering the case of ψ being a polynomial of degree 1. In this case, we have that $C_\psi(\mathcal{S}_\omega(\mathbb{R})) \subset \mathcal{S}_\omega(\mathbb{R})$.

Proposition 1.2.1 *Let $\psi(x) = ax + b$, with $a \neq 0$. The following are equivalent:*

1. *$C_\psi : \mathcal{S}_\omega(\mathbb{R}) \to \mathcal{S}_\omega(\mathbb{R})$ is power bounded.*
2. *$C_\psi : \mathcal{S}_\omega(\mathbb{R}) \to \mathcal{S}_\omega(\mathbb{R})$ is mean ergodic.*
3. *$(C_{\psi_m})_m$ is equicontinuous in $\mathcal{L}(\mathcal{S}_\omega(\mathbb{R}), \mathcal{S}(\mathbb{R}))$.*
4. *$\psi(x) = x$ or $\psi(x) = -x + b$.*

The key estimate to prove Theorem 1.1.5 in our setting is the following rather technical one:

Lemma 1.2.2 *Let ψ be a polynomial of degree greater than 1 without fixed points. For every $\alpha > 1$ there exist $C > 0$ and $r > 1$ such that*

$$|\psi_m^{(n)}(x)| \leq C r^n n!^2 \left(1 + |\psi_m(x)|\right)^\alpha$$

for all $x \in \mathbb{R}$, $n \in \mathbb{N}, m \in \mathbb{N}$.

Once we have proved Lemma 1.2.2, it is possible to show the following result:

Theorem 1.2.3 *Let ω be any subadditive weight and $\sigma(t) = \omega(t^{\frac{1}{a}})$ for $a > 2$. Given a polynomial ψ of degree greater than one, the following statements are equivalent:*

1. *ψ lacks fixed points.*
2. *$\lim_m f \circ \psi_m = 0$ in $\mathcal{S}_\sigma(\mathbb{R})$ for every $f \in \mathcal{S}_\omega(\mathbb{R})$.*
3. *$\lim_n \frac{1}{n} \sum_{m=1}^{n} f \circ \psi_m$ exists in $\mathcal{S}_\sigma(\mathbb{R})$ for every $f \in \mathcal{S}_\omega(\mathbb{R})$.*
4. *$\lim_n \frac{1}{n} \sum_{m=1}^{n} f \circ \psi_m$ exists in $\mathcal{S}(\mathbb{R})$ for every $f \in \mathcal{S}_\omega(\mathbb{R})$.*

We do not know whether the above results are also true for $a = 2$.

It may be worth making the above results explicit in the case where the weight ω is a power of the logarithm, rather than a Gevrey weight. In this case, keeping the notation of the previous result, $\mathcal{S}_\omega(\mathbb{R}) = \mathcal{S}_\sigma(\mathbb{R})$. The limit case $p = 1$ corresponds to [13, Theorem 3.11], since in this case $\mathcal{S}_\omega(\mathbb{R}) = \mathcal{S}(\mathbb{R})$, despite of the fact that ω would not be strictly speaking a weight function (Definition 1.1.4(γ) does not hold).

Corollary 1.2.4 *Let $\omega(x) = \max\{0, \log^p(x)\}$ with $p > 1$. Given a polynomial ψ of degree greater than one, the following statements are equivalent:*

1. *$C_\psi : \mathcal{S}_\omega(\mathbb{R}) \to \mathcal{S}_\omega(\mathbb{R})$ is power bonded.*
2. *ψ lacks fixed points.*
3. *$C_\psi : \mathcal{S}_\omega(\mathbb{R}) \to \mathcal{S}_\omega(\mathbb{R})$ is mean ergodic.*

Corollary 1.2.4 holds for every weight ω satisfying the following condition:

$$\exists \gamma > 1 \; \exists C \geq 1 \; \forall t \geq 0 : \; \omega(t^\gamma) \leq C\omega(t) + C.$$

References

[1] A.A. Albanese, J. Bonet, W.J. Ricker, *Mean ergodic operators in Fréchet spaces*, Ann. Acad. Sci. Fenn. Math. **34** (2009), 401–436.

[2] Asensio, V., Jornet, D.: *Global pseudodifferential operators of infinite order in classes of ultradifferentiable functions.* Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. RACSAM **113**, no. 4, 3477–3512 (2019)

[3] Ariza, H; Fernández, C; Galbis, A; *Composition operators on Gelfand-Shilov classes.* (2024). *Journal of Mathematical Analysis and Applications*, Volume 531(Issue 2, Part 2), 127869.

[4] Ariza, H., Fernández, C., Galbis, A.: *Iterates of composition operators on global spaces of ultradifferentiable functions.* Rev. Real Acad. Cienc. Exactas Fis. Nat. Ser. A-Mat. 119, 9 (2025).

[5] Beurling, A; *Quasi-analyticity and general distributions.* Lecture 4 and 5, AMS Summer Institute, Standford, 1961.

[6] Boiti, C., Jornet, D., Oliaro, A.: *Regularity of partial differential operators in ultradifferentiable spaces and Wigner type transforms.* J. Math. Anal. Appl., **446**, 920–944 (2017)

[7] Boman, J.; *On the intersection of classes of infinitely differentiable functions.* Ark. Mat. 5, 301–309 (1964).

[8] Bonet, J.; Meise, R.; Melikhov, S.N.; *A comparison of two different ways to define classes of ultradifferentiable functions.* Bull. Belg. Math. Soc. Simon Stevin 14, 424–444 (2007)

[9] Bonet, J.; Domanski, P; *Abel's functional equation and eigenvalues of composition operators on spaces of real analytic functions.* Integral Equations Operator Theory 81(4), 455–482 (2015)

[10] Braun, R; Meise, R; Taylor, B.A.; *Ultradifferentiable functions and Fourier analysis*, Result. Math. 17 (1990), 206–237.

[11] Cowen, C. C.; MacCluer, B. D.; *Composition operators on spaces of analytic functions.* Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1995.

[12] Fernández, C.; Galbis, A; *Superposition in Classes of Ultradifferentiable Functions.* Publ. Res. Inst. Math. Sci. 42 (2006), no. 2, pp. 399–419

[13] Fernández, C; Galbis, A; Jordá, E; *Spectrum of composition operators on $\mathcal{S}(\mathbb{R})$ with polynomial symbols.* Adv. Math. **365** (2020), 107052, 24 pp.

[14] Fernández, C; Galbis, A; Jordá, E; *Spectrum of composition operators on $\mathcal{S}(\mathbb{R})$ with polynomial symbols.* Adv. Math. **365** (2020), 107052, 24 pp.

[15] Galbis, A; Jordá, E; *Composition operators on the Schwartz space.* Rev. Mat. Iberoam. 34 (2018), no. 1, 397–412.

[16] Gelfand, I. M.; Shilov, G. E. (1968) [1958], *Generalized functions. Vol. 2. Spaces of fundamental and generalized functions*, Boston, MA

[17] Meise, R; Taylor, B. A.; *Whitney's extension theorem for ultradifferentiable functions of Beurling type*, Ark. Mat. 26 (1988), 265–287.

[18] Nicola, F; Rodino, L; *Global Pseudo-Differential Calculus on Euclidean Spaces*, volume 4 of Pseudo-Differential Operators. Theory and Applications. Birkhäuser Verlag, Basel, 2010.

[19] Przestacki, A.; *Composition operators with closed range for one-dimensional smooth symbols.* J. Math. Anal. Appl. 399 (2013), no. 1, 225–228.

[20] Shapiro, J. H.; *Composition operators and classical function theory.* Universitext, Tracts in Mathematics, Springer-Verlag, New York, 1993.

[21] Shapiro, J. H.; *Composition operators and Schroder functional equation.* Studies on composition operators (Laramie, WY, 1996), Contemp. Math. 213, 213–228. Amer. Math. Soc., Providence, 1998.

Some new subdivision schemes in the context of cell-averages

Inmaculada Garcés

Universidad de Valencia, Spain. Email address: inmaculada.garces@uv.es

1.1 Introduction

CAGD (Computer Aided Geometric Design) is primarily aimed at providing programmers with solutions for the generation of smooth curves and surfaces. It does this through flexible and efficient tools. This design provides the user with the possibility, through an initial set of control points and the proposal of a subdivision rule, to obtain the indicated results. The system is intuitive, requiring only the user to set the control points. The use of classical schemes such as the four-point interpolation method [3] and its variants [8] justifies its affordability. Other contributions, such as schemes with adjustable parameters [7] and methods dealing with noisy data [5] make the design more adaptable and extend its reliability.

Apart from the CAGD seen, subdivision schemes draw from other sources, such as applications in approximation theory, signal processing and image compression, where, using a finite number of basis functions, the function representation is achieved. The target properties are compression rates, local support and computational speed. All this under the concept of multiresolution analysis [6]. In addition to the classical linear and stationary approaches, nonlinear subdivision schemes have been developed to address more complex data structures and improve flexibility in applications [4].

More recently, it is the subdivision in the context of cell-averages rather than point data that has caused most interest, more specifically for applications involving numerical methods for partial differential equations and volume representation. Other more classical schemes, such as those studied in [2] and [6], have focused on point values. Adapting schemes to cell-average data requires other types of analysis tools and other refinement rules. Thus, the most recent advances in non-oscillatory and high regularity [1] schemes, as well as in [9], raise the current interest in these generalisations.

The present paper first provides a brief review of the fundamentals of subdivision schemes, and then discusses the construction and analysis of new linear subdivision schemes, in this case for univariate cell-average data. These schemes are designed to reproduce cell-average of polynomials of degrees less than or equal to 1, 3 and 5. We analyze the convergence and smoothness of the limit functions using Laurent polynomial representations and the joint spectral radius method, following the analytical framework introduced in [6].

The goal is to study the behavior of new subdivision schemes in the context of cell-averages in 1D, already worked on for point values in [3], [6] and [9]. We analyse the reproduction, approximation and convergence properties of the new schemes and present some numerical experiments to validate the theoretical results and demonstrate the practical behaviour of the proposed methods.

1.2 Main results

Subdivision schemes are based on the successive application of certain rules to a series of control points. First of all, let us look at a general definition.

Definition 1 (A subdivision scheme) *Let $\mathbf{f}^0 = \{f_l^0\}_{l=1}^m$ be a series of control points. Then, there are functions $\Psi_0, \Psi_1 \colon \mathbb{R}^m \to \mathbb{R}$ that define a subdivision scheme (SS) in the following way:*

$$\begin{cases} (S\,\mathbf{f}^0)_{2i} = \Psi_0(f_1^0, \ldots, f_m^0), \\ (S\,\mathbf{f}^0)_{2i+1} = \Psi_1(f_1^0, \ldots, f_m^0), \end{cases}$$

with $i \in \mathbb{Z}$.

The Chaikin algorithm serves as an example of a spline subdivision scheme, offering low complexity converging towards functions in C^1 and without Gibbs phenomenon.

Example 1 (Chaikin scheme) *Chaikin introduced a simple scheme for generating curves from a given control polygon, defined as follows:*

$$\begin{cases} (S_C\mathbf{f}^0)_{2i} = \frac{3}{4}f_i^0 + \frac{1}{4}f_{i+1}^0, \\ (S_C\mathbf{f}^0)_{2i+1} = \frac{1}{4}f_i^0 + \frac{3}{4}f_{i+1}^0. \end{cases}$$

As we are going to work in the context of linear subdivision schemes, we give their particular definition (see [6]).

Definition 2 (A linear binary SS) *A SS, $S_{\mathbf{a}} : \ell_\infty(\mathbb{Z}) \to \ell_\infty(\mathbb{Z})$, with finitely supported mask $\mathbf{a} = \{a_l\}_{l\in\mathbb{Z}}$ is defined to refine the data on the level k, $\mathbf{f}^k = \{f_j^k\}_{j\in\mathbb{Z}} \in \ell_\infty(\mathbb{Z})$, as:*

$$f_{2j+i}^{k+1} := (S_{\mathbf{a}}\mathbf{f}^k)_{2j+i} := \sum_{l\in\mathbb{Z}} a_{2l-i} f_{j+l}^k, \quad j \in \mathbb{Z}, \quad i = 0, 1.$$

We call even mask to $\mathbf{a}^0 = \{a_{2l}\}_{l\in\mathbb{Z}}$ and odd mask $\mathbf{a}^1 = \{a_{2l-1}\}_{l\in\mathbb{Z}}$.

Once we know the definition of subdivision scheme, we ask ourselves how do we understand the points. They can be understood as a point measurement, which is the value of the function at the point and we call it point values, or they can also be understood as an average of a cell, which is an integral and we call it cell-averages.

Let $x_i^k = ih_k$ equispaced points, with $k \in \mathbb{N}$, $i \in \mathbb{Z}$. The discretization operator is defined in each case in this way:

Definition 3 (Discretization operator in point values) *We define the discretization operator as the value of the function at a point of the mesh, i.e.,*

$$f_i^k := f(x_i^k).$$

Definition 4 (Discretization operator in cell-averages) *We define the discretization operator as the average over the cell $c_i^k = (2^{-k}(i-1), 2^{-k}i)$, i.e.,*

$$\bar{f}_i^k := 2^k \int_{c_i^k} f(x)dx.$$

In our case, it is possible to work with cell-averages because in an image we have pixels and the colour of these pixels is determined by the integral, i.e., the average of the intensity in that square.

1.2.1 The cell-average scheme for $p = 1, 3, 5$

In 1D, we consider our data as the cell-average of an integrable function $f \in \mathcal{L}^1(\mathbb{R})$. Thus, let $x_i^k = ih_k$, $k \in \mathbb{N}$, $i \in \mathbb{Z}$ and $h_k = 2^{-k}$ be a grid; the data cells are defined as:

$$\bar{f}_i^k = \frac{1}{h_k} \int_{x_i^k - \frac{h_k}{2}}^{x_i^k + \frac{h_k}{2}} f(x)dx.$$

We will construct new linear refinement rules that reproduce cell-average of polynomials with degrees less than or equal to $p = 1, 3, 5$. Therefore, we will have

$$(S_p \bar{f}^k)_{2j} = \sum_{l=-\frac{p-1}{2}}^{\frac{p-1}{2}} \beta_l^p \bar{f}_{j+l}^k,$$

$$(S_p \bar{f}^k)_{2j+1} = \sum_{l=-\frac{p-1}{2}}^{\frac{p-1}{2}+1} \alpha_l^p \bar{f}_{j+l}^k.$$

In Table 1.1 we show the results of the masks for $p = 1, 3, 5$.

	β_l^1	α_l^1	β_l^3	α_l^3	β_l^5	α_l^5
$l = 0$	1	$\frac{1}{2}$	$\frac{17}{16}$	$\frac{37}{64}$	$\frac{1109}{1024}$	$\frac{1247}{2048}$
$l = 1$		$\frac{1}{2}$	$-\frac{1}{32}$	$\frac{37}{64}$	$-\frac{23}{512}$	$\frac{1247}{2048}$
$l = 2$				$-\frac{5}{64}$	$\frac{7}{2048}$	$-\frac{509}{4096}$
$l = 3$						$\frac{63}{4096}$

Table 1.1: Masks of cell-average subdivision for $p = 1, 3, 5$. Note that $\alpha_l^p = \alpha_{-l}^p$ and $\beta_{l+1}^p = \beta_{-l}^p$, $l = 1, \ldots, \frac{p-1}{2}$.

We will now perform an analysis of the convergence and smoothness of these schemes. The classical definition of a convergent subdivision is the following [6]:

Definition 5 (A uniformly convergent subdivision scheme) *A SS S_a is* uniformly convergent *if for any initial data $\mathbf{f}^0 \in \ell^\infty(\mathbb{Z})$, there exists a continuous function $F : \mathbb{R} \to \mathbb{R}$ such that*

$$\lim_{k \to \infty} \sup_{j \in \mathbb{Z}} |(S_a^k \mathbf{f}^0)_j - F(2^{-k} j)| = 0.$$

We denote by $S_a^\infty \mathbf{f}^0 = F$ to the limit function generated from $\mathbf{f}^0$ and we write $S_a \in \mathcal{C}^d$ if all the limit functions have such smoothness, $S_a^\infty \mathbf{f}^0 \in \mathcal{C}^d$, $\forall\ \mathbf{f}^0 \in \ell^\infty(\mathbb{Z})$.

The 2-cell-average scheme with $p \in \Pi_1$ that we propose is:

$$\begin{aligned} f_{2j}^{k+1} &= f_j^k, \\ f_{2j+1}^{k+1} &= \frac{1}{2} f_j^k + \frac{1}{2} f_{j+1}^k. \end{aligned} \tag{1.1}$$

This is the piecewise linear scheme, which is the simplest example of an interpolatory subdivision scheme. We represent the basic limit function. The original points are represented with

blue spheres while the points resulting from the corresponding iterations are represented with red dots.

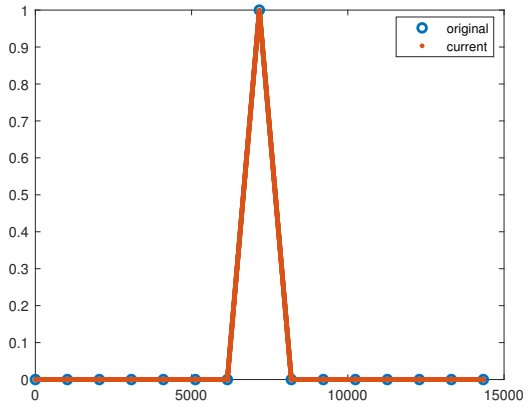


Figure 1.1: Basic limit function of the 2-cell-average scheme defined by (1.1) with the following parameters: 15 initial points, 10 iterations.

We can see in Figure 1.1 that the limit is the piecewise linear interpolant to the data, so we can predict that the scheme will be C^0 but not C^1.

The 4-cell-average scheme with $p \in \Pi_3$ that we propose is:

$$\begin{aligned} f_{2j}^{k+1} &= \left(1 + \frac{1}{16}\right) f_j^k - \frac{1}{32}(f_{j-1}^k + f_{j+1}^k), \\ f_{2j+1}^{k+1} &= \frac{37}{64}(f_j^k + f_{j+1}^k) - \frac{5}{64}(f_{j-1}^k + f_{j+2}^k). \end{aligned} \tag{1.2}$$

We represent the basic limit function, as in the previous case.

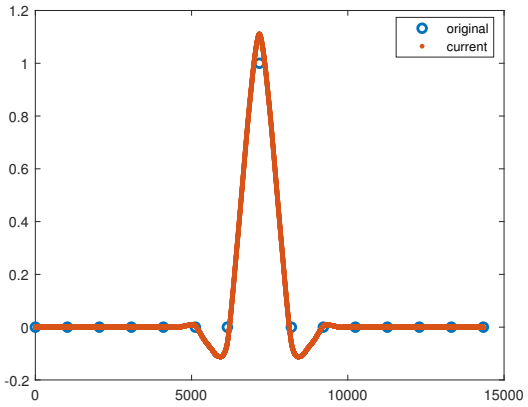
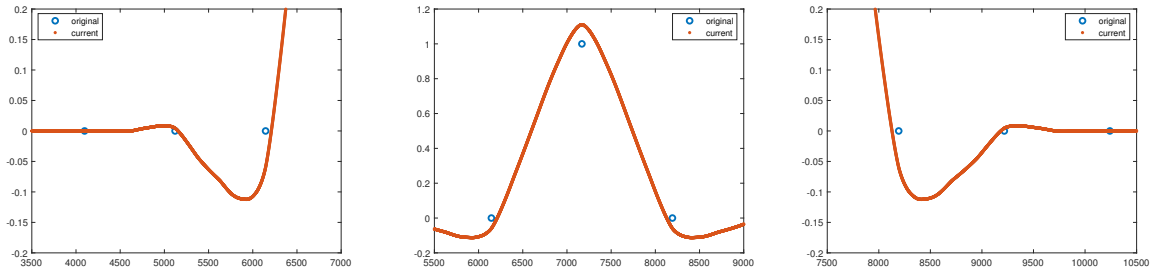


Figure 1.2: Basic limit function of the 4-cell-average scheme defined by (1.2) with the following parameters: 15 initial points, 10 iterations.

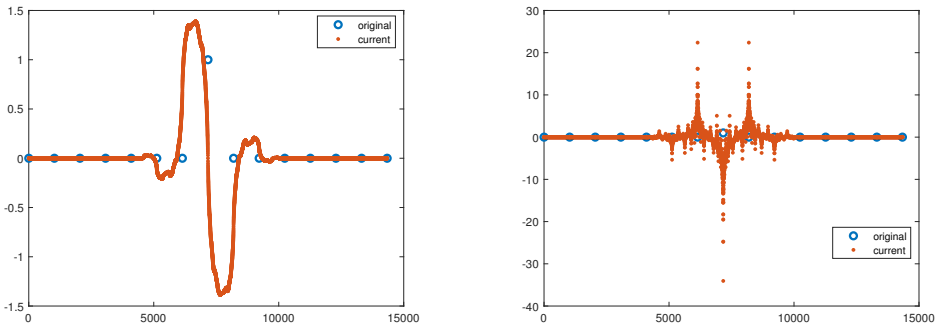
In order to get a closer look at the surroundings near the jump, we performed the following zooms that can be seen in the Figure 1.3.



(a) (b) (c)

Figure 1.3: Zooms of the basic limit function in Figure 1.2 divided into three different regions: **(a)** $x \in [3500, 7000]$; **(b)** $x \in [5500, 9000]$; **(c)** $x \in [7500, 10500]$.

To give a more supported prediction, we plot the approximation of the derivatives corresponding to this case. We can see this in the Figure 1.4.



(a) (b)

Figure 1.4: **(a)** Approximation of the first derivative; **(b)** Approximation of the second derivative.

There are no peaks in the function, so we can predict that the scheme will be C^1 but not C^2.

Finally, the 6-cell-average scheme with $p \in \Pi_5$ that we propose is:

$$\begin{aligned} f_{2j}^{k+1} &= \frac{1109}{1024} f_j^k - \frac{23}{512}(f_{j-1}^k + f_{j+1}^k) + \frac{7}{2048}(f_{j-2}^k + f_{j+2}^k), \\ f_{2j+1}^{k+1} &= \frac{1247}{2048}(f_j^k + f_{j+1}^k) - \frac{509}{4096}(f_{j-1}^k + f_{j+2}^k) + \frac{63}{4096}(f_{j-2}^k + f_{j+3}^k). \end{aligned} \tag{1.3}$$

We perform the same process as before.

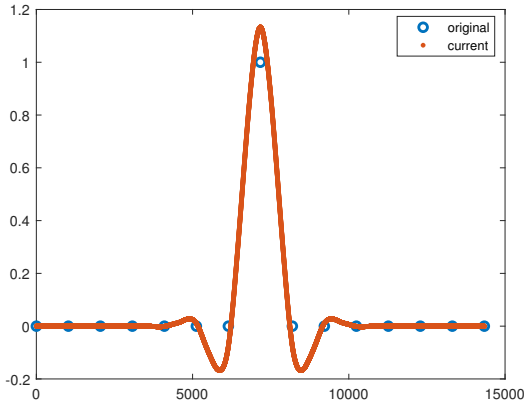
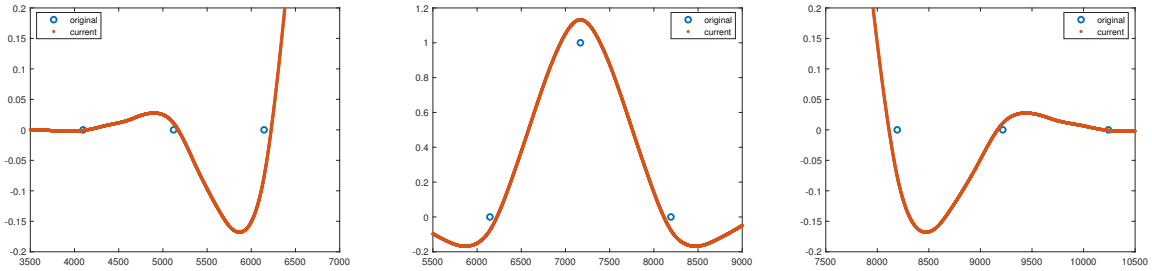
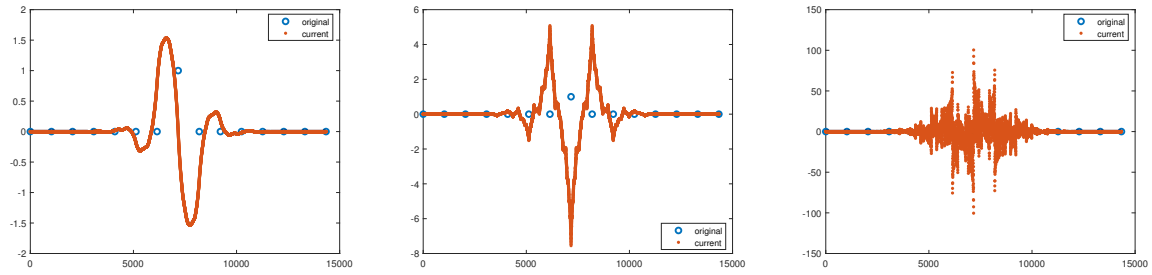


Figure 1.5: Basic limit function of the 6-cell-average scheme defined by (1.3) with the following parameters: 15 initial points, 10 iterations.



(a) (b) (c)

Figure 1.6: Zooms of the basic limit function in Figure 1.5 divided into three different regions: **(a)** $x \in [3500, 7000]$; **(b)** $x \in [5500, 9000]$; **(c)** $x \in [7500, 10500]$.



(a) (b) (c)

Figure 1.7: **(a)** Approximation of the first derivative; **(b)** Approximation of the second derivative; **(c)** Approximation of the third derivative.

By looking at the Figures 1.5, 1.6 and 1.7, we can predict that the scheme will be C^2 but not C^3.

Using classical techniques we prove that it is convergent, which in this case being linear is easy to prove. So let us look at the Laurent polynomial formulation.

Given a subdivision scheme S_a, with coefficients $a = \{a_j\}_{j\in\mathbb{Z}}$, we define an associate Laurent

polynomial

$$a(z) = \sum_{j\in\mathbb{Z}} a_j z^j,$$

which is the symbol of S_a. The values generated at level k of the subdivision, namely $\{f_j^k\}_{j\in\mathbb{Z}}$, define a formal Laurent series

$$F_k(z) = \sum_{j\in\mathbb{Z}} f_j^k z^j,$$

satisfying the relation

$$F_{k+1}(z) = a(z)F_k(z^2).$$

So this polynomial allows us to have the property that if it is contractive (the segments are closing) it finally gives us a continuous function.

Definition 6 (Contractive subdivision scheme) *A subdivision scheme is termed 'contractive' if it sends any initial data to a zero limit.*

Defining

$$b^{[l]}(z) = \prod_{i=0}^{l-1} b\left(z^{2^i}\right) = \sum_i b_i^{[l]} z^i,$$

we have the Laurent series representation of a subdivision scheme transforming values at level k directly to values at level $k+l$. The number of points is multiplied by 2^l, hence, there are 2^l rules. Contractivity follows if the sum of absolute values of the coefficients of each of these rules is less than 1:

$$\sum_i \left|b_{2^l i+r}^{[l]}\right| < 1,\ 0 \le r < 2^l.$$

The following result allows us to check the C^r convergence:

Theorem 1 (Condition for C^r) *S_a is C^r-convergent iff the scheme defined by $\frac{b^{[r]}(z)}{z^{-1}+1}$ is contractive.*

To analyze the convergence and the smoothness of the limit function, we can use the 'joint spectral radius' analysis (JSR).

The joint spectral radius is defined as:

$$\rho(Q_0, Q_1) = \lim_{m\to\infty} \sup\left(\max\left\{\|Q_{i_1}Q_{i_2}\dots Q_{i_m}\|_\infty : i_j \in \{0,1\},\ j = 1,\dots,m\right\}\right)^{1/m},$$

where the matrices Q_0 and Q_1 are obtained by changing the basis.

To show that the scheme is C^k it remains to verify that $\rho(Q_0, Q_1) < 2^{-k}$. The Hölder exponent of the k-th derivative of the limit function is defined by $\nu = -\log_2 \rho(Q_0, Q_1) - k$.

On the one hand, the 4-cell-average scheme defined by (1.2) is C^1 because the JSR is less than 0.5 after 15 iterations, as we can see in the Figure 1.8.

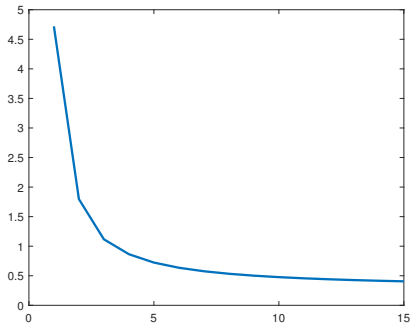


Figure 1.8: The 4-cell-average scheme defined by (1.2) is C^1.

We repeat the previous process and we can see that the JSR is greater than 0.25 after 15 iterations, so the scheme is not C^2, as we can see in the Figure 1.9 with the corresponding graph and a zoom to see it more clearly.

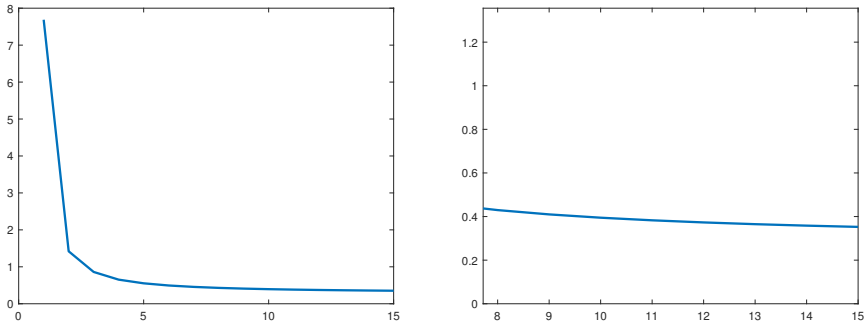


Figure 1.9: The 4-cell-average scheme defined by (1.2) is not C^2.

On the other hand, the 6-cell-average scheme defined by (1.3) is C^2 because the JSR is less than 0.25 after 18 iterations, as we can see in the Figure 1.10. We zoom in slightly to make it clearer.

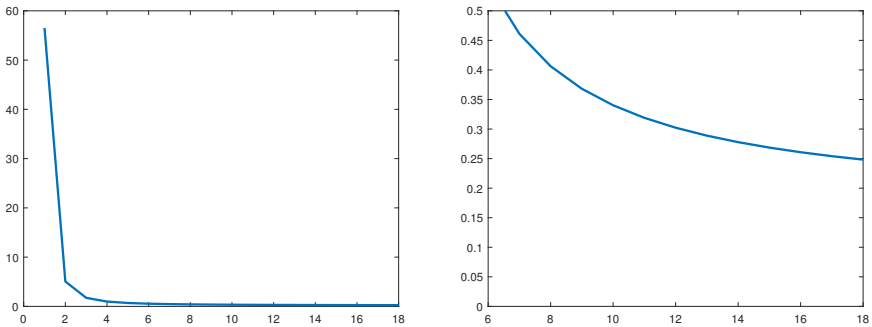


Figure 1.10: The 6-cell-average scheme defined by (1.3) is C^2.

We repeat the previous process and we can see that the JSR is greater than 0.128 after 15 iterations, so the scheme is not C^3, as we can see in the Figure 1.10 and we also zoom in and out.

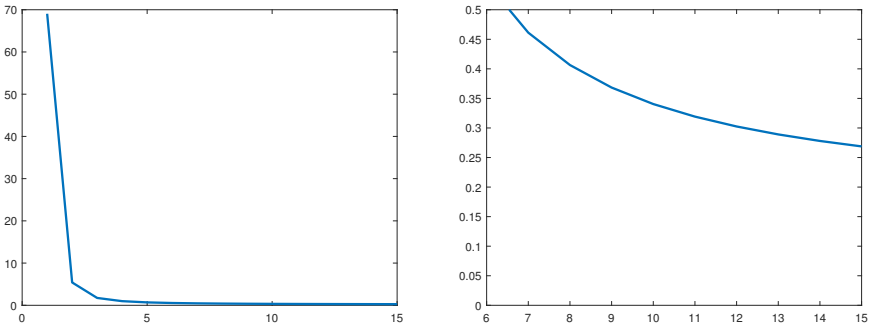
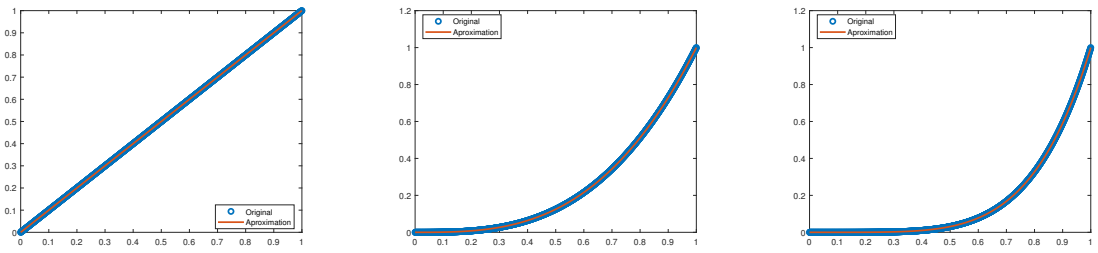


Figure 1.11: The 6-cell-average scheme defined by (1.3) is not C^3.

1.3 Numerical experiments

Finally, we perform numerical experiments in order to analyze the numerical behavior of the schemes. We start by checking the cell-average of polynomial reproduction for each configured mask.

We take a mesh equispaced between 0 and 1 of 50 initial points and apply 5 iterations, where in each iteration the distance between the points is divided by 2. We can see the final result for polynomials of degrees 1, 3 and 5 in Figure 1.12.

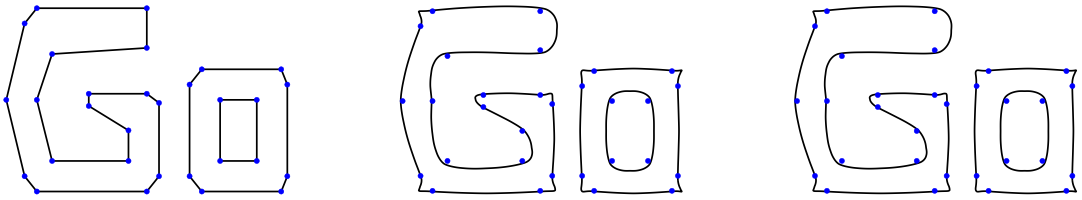


(a) $p = 1$ (b) $p = 3$ (c) $p = 5$

Figure 1.12: The cell-average of polynomial reproduction with degrees 1, 3 and 5 applying the schemes defined by: **(a)** the 2-cell-average scheme (1.1) ; **(b)** the 4-cell-average scheme (1.2); **(c)** the 6-cell-average scheme (1.3).

Evidently we see that the error is zero and therefore we conclude that our schemes reproduce the cell-average of polynomial reproduction.

To end this part, we apply these schemes to the design of 1D curves in order to study the smoothness of the constructed cell-average schemes. We start with a control polygon used in some articles (see [8]) and we can see the graphs obtained in Figure 1.13.

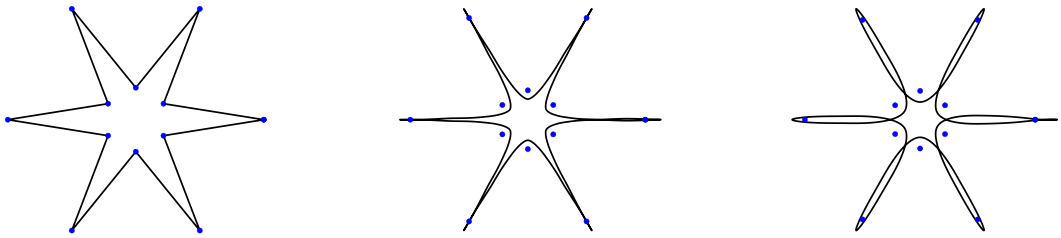


(a) $p = 1$ (b) $p = 3$ (c) $p = 5$

Figure 1.13: Control polygon of the word "Go" (see [8]) applying the schemes defined by: **(a)** the 2-cell-average scheme (1.1) ; **(b)** the 4-cell-average scheme (1.2); **(c)** the 6-cell-average scheme (1.3). The blue dot marker represent the original points while the black line joins the interpolation points by applying the corresponding method.

It is observed that, the higher the degree of the polynomials of the scheme used, the smoother the smoothness is obtained due to the disappearance of peaks in the function.

We continue to apply our cell-average schemes to other control polygons, such as a star. In particular, we take a six-pointed star that can be found in [1].

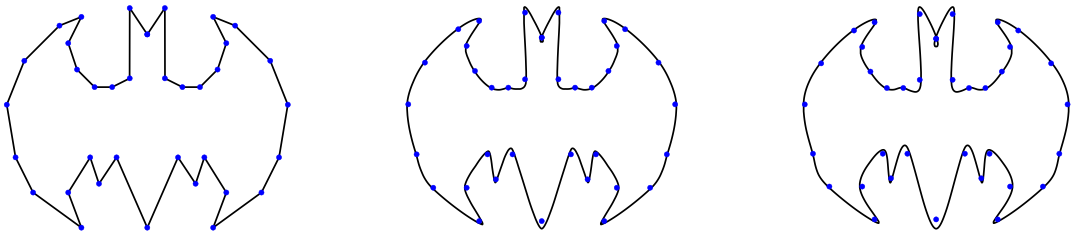


(a) $p = 1$ (b) $p = 3$ (c) $p = 5$

Figure 1.14: Control polygon of the six-pointed star (see [1]) applying the schemes defined by: **(a)** the 2-cell-average scheme (1.1) ; **(b)** the 4-cell-average scheme (1.2); **(c)** the 6-cell-average scheme (1.3). The blue dot marker represent the original points while the black line joins the interpolation points by applying the corresponding method.

In Figure 1.14, we can observe the same fact as we have seen above, so that smoothness is obtained in cell-average schemes (1.2) and (1.3).

We finish the numerical experiments with another figure that can be found in [1] and is a control polygon that has the shape of a bat. The results can be seen in Figure 1.15.



(a) $p = 1$ (b) $p = 3$ (c) $p = 5$

Figure 1.15: Control polygon of a bat (see [1]) applying the schemes defined by: **(a)** the 2-cell-average scheme (1.1) ; **(b)** the 4-cell-average scheme (1.2); **(c)** the 6-cell-average scheme (1.3). The blue dot marker represent the original points while the black line joins the interpolation points by applying the corresponding method.

Therefore, we observe that in the design of 1D curves our schemes are smoother the higher the degree of the polynomial we interpolate, as expected.

1.4 Conclusions

In this paper, new linear refinement rules have been constructed to define new subdivision schemes that reproduce cell-average of polynomials with degrees less than or equal to 1, 3 and 5, as shown by numerical experiments, as well as the study of the smoothness of these schemes. It should be noted that these designed cell-average subdivision schemes can be extended to any degree. In addition, other theoretical properties are being studied. As future work, the procedure will be extended to two-dimensional schemes for further application to digital image processing, with the aim of evaluating its performance in tasks such as reconstruction and resolution enhancement for comparison with existing techniques in this field.

Bibliography

[1] S. Amat, J. Ruiz, J. C. Trillo and D. F. Yáñez, *On a family of non-oscillatory subdivision schemes having regularity C^r with $r > 1$*, Numerical Algorithms, **85**: 543 – 569, 2020.

[2] G. Deslauriers and S. Dubuc, *Symmetric Iterative Interpolation Processes*, Const. Approx.: 49–68, 1989.

[3] N. Dyn, D. Levin, J. A. Gregory, *A 4-point interpolatory subdivision scheme for curve design*, Comput. Aided Geom. Des., **4(4)**: 257–268, 1987.

[4] N. Dyn , *Three Families of Nonlinear Subdivision Schemes*, Studies in Computational Mathematics, **12**: 23–38, 2006.

[5] N. Dyn, A. Heard, K. Hormann and N. Sharon , *Univariate subdivision schemes for noisy data with geometric applications*, Comput. Aided Geom. Des., **37**: 85–104, 2015.

[6] N. Dyn and D. Levin, *Subdivision Schemes in Geometric Modelling*, Acta Numerica: 73–144, 2002.

[7] N. Dyn, D. Levin and C. A. Micchelli, *Using parameters to increase smoothness of curves and surfaces generated by subdivision*, Comput. Aided Geom. Des., **7(1-4)**: 129–140, 1990.

[8] N. Dyn, M. S. Floater and K. Hormann, *Four-point curve subdivision based on iterated chordal and centripetal parameterizations*, Comput. Aided Geom. Des., **26(3)**: 279–286, 2009.

[9] S. Hed, *Analysis of subdivision schemes for surface generation*, Tel-Aviv University, 1992.

SIR model: study of its initial foundations and mathematical development

Jessica Paredes Morales[♭] *

(∗) Universitat Politècnica de València & Universitat de València
(♭) jespamo@alumni.uv.es

1.1 Introduction

Behavioral models are schemes that describe, explain and predict how individuals, groups or systems behave in certain contexts or situations. These models are fundamentally based on how people act in reality, collecting data on their behavior and reducing the complexity of human or social behavior to focus on the most relevant aspects. In addition, it helps to anticipate future reactions to certain stimuli or contexts. These models are used to build design strategies in various areas such as education, business, health and technology.

These models can be reduced to mathematical formulations that bring together all the elements of such models. In this way the elements can be manipulated to obtain new mathematical representations which yield valuable information that would not be obtained by maintaining the language of the science involved.

1.2 Main Transmission Models

In mathematics there are two models: deterministic and stochastic. The former are those models that give exact results, since in the theory the factors involved in the process or phenomenon can be controlled. On the other hand, in a stochastic model these factors cannot be controlled, since the random processes that are present make the results neither simple nor unique.

We must be clear that in a deterministic model a single individual can cause an epidemic, while in a stochastic model it can happen that the epidemic disappears, so that some stochastic solutions converge to a disease-free state even though their corresponding deterministic solution converges to the endemic equilibrium. [**?**]

The SIR (Susceptible-Infected-Recovered) model has its origin in the study of epidemics through mathematical tools, developed by William Ogilvy Kermack and Anderson Gray McKendrick in 1927. Their seminal paper, entitled "A Contribution to the Mathematical Theory of Epidemics", was published in the Proceedings of the Royal Society of London.

In this model, the division of the population into compartments was introduced.:

- $S(t)$: Number of susceptible individuals (those who can contract the disease).
- $I(t)$: Number of infected individuals (those who are infectious and transmit the disease).

- $R(t)$: Number of recovered individuals (immune or deceased, no longer involved in transmission).

1.3 Construction of the SIR model

We need to explain how the SIR arose, since it would be considered a fundamental pillar for the formulation of new dynamical systems for modeling that describe the behavior of infectious diseases.

In this sense, Kermack and McKendrick begin by considering that infections happen at the instant of passing from one interval to another, where the size of the interval is denoted by t. That is, the unit of time, where it can be considered constant. In addition, the number of individuals is denoted by $v_{t,\theta}$, where the number of individuals is influenced by t and by the number of intervals θ, in other words the infections that occur at each instant.

Then, the total number of sick people in the interval t is considered to be

$$y_t = \sum_{\theta=0}^{t} v_{t,\theta}$$

Where the following is denoted: $v_{t,0}$: number of individuals who are starting their infection.

v_t : number of individuals that are in the process of infection during the transition to the next interval.

In wich, $v_{t,0} = v_t$ except at the origin or onset of infection where it is present,,

$$v_{0,0} = v_0 + y_0$$

This is because it is assumed that a certain number y_0 of the population has been recently infected. This described process indicates the process of recovery or death of each individual involved in the infection of the population.

Then, the following parameters are presented:

ψ_θ : elimination rate, which is the sum of the rate of deaths and recoveries. $\psi_\theta v_{t,\theta}$: the number of people removed from each interval t. That is, $\psi_\theta v_{t,\theta} = v_{t,\theta} - v_{t+1,\theta+1}$

Therefore, it can be assumed $v_{t,\theta} = v_{t-\theta,0}\ B_\theta$, donde $B_\theta = (1-\psi(\theta-1))(1-\psi(\theta-2))\dots(1-\psi(0))$. Thus new parameters are presented, in which v_t is defined,

$$v_t = x_t \sum_{\theta=1}^{t} \Phi_\theta v_{t,\theta}.$$

x_t : number of unaffected individuals

Φ_θ : rate of infectious capacity in θ. It is worth noting that $Phi_\theta = 0$ at the time the individual is infected.

This equality is assumed since the chance of infection is proportional to the number of infected and uninfected. Next, $x_t = N - sum_{\theta=0}^{t} v_t - y_0$, with N equal to the initial population density. Next, we define a new variable z_t which indicates the number of individuals killed or recovered, so we can express $N = x_t + y_t + z_t$.

Based on this last equality and what has been explained, the following equations are defined,

$$v_t = x_t \left(\sum_{\theta=1}^{t} A_\theta v_{t-\theta} + A_t y_0 \right)$$

with $A_\theta = \Phi_\theta B_\theta$.

Then, $y_t = \sum_{\theta=0}^{t} B_\theta v_{t-\theta} + B_t y_0$
By definition of v_t we have

$$v_t = x_t - x_{t-1} = x_t \left(\sum_{\theta=1}^{t} A_\theta v_{t-\theta} + A_t y_0 \right).$$

The number of persons removed is denoted by
$z_{t+1} - z_t = \sum_{\theta=1}^{t} C_\theta v_{t-\theta} + C_t y_0$
with $C_\theta = \psi_\theta B_\theta$.
In the same way, the following is defined

$$y_{t+1} - y_t = x_t \left(\sum_{\theta=1}^{t} A_\theta v_{t-\theta} + A_t y_0 \right) - \left(\sum_{\theta=1}^{t} C_\theta v_{t-\theta} + C_t y_0 \right)$$

Thus, if the interval is divided into smaller subintervals, i.e. we tend to the limit the following equations, we have

$$\begin{aligned} -v_t = -x_t + x_{t-1} \Rightarrow v_t &= -\frac{dx_t}{dt} \\ v_t = x_t - x_{t-1} \Rightarrow \frac{dx_t}{dt} &= (-x_t) \int_0^t A_\theta v_{t-\theta} d\theta + A_t y_0 \\ z_{t+1} - z_t \Rightarrow \frac{dz_t}{dt} &= \int_0^t C_\theta v_{t-\theta} d\theta + C_t y_0 \\ y_t \Rightarrow \int_0^t B_\theta v_{t-\theta} d\theta &+ B_t y_0 \end{aligned}$$

Now, $B_\theta = e^{-\int_0^\theta \psi(a) da}$ and these new equalities determine the functions $x,\ y,\ zyv$.
Then omitting t from $\frac{dx-t}{dt}$ when necessary and taking x as a function of $theta$ we have

$$\frac{dx}{dt} = (x) \int_0^t A_{t-\theta} \frac{dx_\theta}{d_\theta} d\theta - A_t y_0$$

When solving this equation, knowing that $A_0 = 0$, given that the individuals are not infectious at the instant of transmission and assuming that $N = x_0 + y_0$ we have

$$\frac{d \log x}{dt} = -A_t N + \int_0^t A'_\theta x_{t-\theta} d\theta.$$

This last integral equation cannot be solved in such a way that it gives us x in terms of t. But, it is observed that this equation is similar to Volterra's equation $f(t) = \Phi(t) + \int_0^t N(T, \theta)\Phi(\theta) d\theta$ where $f(t)$ is taken as $fracdf(t)dt$.

Thus, if we consider the equation of the form

$$\frac{d \log x}{dt} = A_t + \lambda \int_0^t N(t, 0) x(\theta) d\theta,$$

which could be solved through a series of successive approximations, using the same method to solve the Volterra equation, i.e. $x = f_0(t) + f_1(t) + \lambda^2 f_2(t) + \ldots$, substituting in $\frac{dx}{dt} = (x)A_t + \lambda \int_0^t N(t, 0) x(\theta) d\theta$, where it is obtained

$$\begin{aligned} \frac{df_n(t)}{dt} =& f_n(t) A_t + f_{n-1}(t) \int_0^t N(t, \theta) f_0(\theta) d\theta + f_{n-2}(t) \int_0^t N(t, \theta) f_1(\theta) d\theta + \ldots \\ &+ f_0(t) \int_0^t N(t, \theta) f_{n-1}(\theta) d\theta \\ &= L_{n-1}(t). \end{aligned}$$

That when solving the differential equation $\frac{d}{dt}f_n(t) = L_{n-1}(t)$

$$f_n(t)e^{-\int_0^t Atdt} = \int_0^t L_{n-1}(t)e^{-\int_0^t Atdt}dt + C$$

Thus, $f_n(0) = 0$ when $n > 0$, since the initial condition is independent of *lambda* so that the integration constants are 0, except when $f_0(0)$, we have

$$\frac{df_0(t)}{dt} = f_0(t)A_t \Rightarrow f_0(t) = f_0 e^{\int_0^t Atdt} \Rightarrow f_0(0) = x_0.$$

In this sense, the solution for the integral equation is as follows

$$x = (E_t)x_0 + \sum_{\theta=1}^{\infty}\int_0^t \frac{L_{n-1}(t)}{E_t}dt$$

Where, $E_t = \int_0^t A_t dt$ y $\lambda = 1$. Thus, writing in general form $\frac{d\log x}{dt}$ we have

$$\frac{d\log x}{dt} = A_t + \int_0^t Q_{t-\theta}x_\theta d\theta$$

Then, multiplying by e^{-zt} and integrating between 0 and t we have

$$\log x_0 + \int_0^\infty ze^{-zt}\log x dt = F(z) + F_1(z)\int_0^\infty e^{-zt}x_t dt.$$

Where $F(z) = \int_0^\infty e^{-zt}A_t dt$, $F_1(Z) = \int_0^\infty e^{-z\theta}Q_\theta d\theta$ and applying limit when t tends to $e^{-zt}\log x$ is 0, provided that x does not exceed the initial value $N - y_0$. Thus we obtain

$$\int_0^\infty e^{-zt}(z\log x - F_1(z)x)dt = F(z) + \log x_0$$

which can be viewed as a first type Fredholm equation.

$$\int_0^\infty \Phi(x,z)\psi(z,t)dt = \chi(z).$$

Then, we again take the equation $\frac{d\log x}{dt}$ this time integrating between 0 y ∞.

$$-\int_0^\infty \frac{d\log x}{dt}dt = \int_0^\infty\int_0^t A_\theta v_{t-\theta}d\theta dt + y_0\int_0^\infty A_t dt,$$

so that $\log\frac{x_0}{x_\infty} = A(N - x_\infty)$.

with $A = \int_0^\infty A_t dt$ y $\int_0^\infty v_t dt = x_0 - x_\infty$.Now we denote a new parameter $p = \frac{N-x_0}{N}$, which is the proportion of the infected population in the epidemic which is useful for having the following expression,

$$-\log\frac{1-p}{1-\frac{y_0}{N}} = AN_p.$$

Analogously, we have

$$\int_0^\infty y_t dt = Np\int_0^\infty B_\theta d\theta$$

Where $\int_0^\infty B_\theta d\theta$ is the average duration of each case.

Despite the construction of these last equations in terms of x, y and z the information may be incomplete in some cases. That is to say, the problem lies in obtaining information A_θ and

B_θ consequently also $\Phi(\theta)$ y $\psi(\theta)$. Then the equation is taken again $\frac{dx_t}{dt}$, in which Fock's method is used to obtain information A_θ and B_θ. In this way, using v_t y $\frac{d\log x}{dt}$ we have

$$A_\theta = \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} e^{zt}F_2(z)dz$$
$$B_\theta = \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} e^{zt}F_3(z)dz$$

with $F_2(z) =$ y $F_3(z) =$.

It should be noted that the author offers a simpler solution for special cases. In this case, one could say that by stating these special cases, one would be stating the fundamental basis for the SIR model, as it is known today. The case focuses on the early stages of an epidemic in a large population. Thus taking $\frac{dx}{dt} = vt$ and applying Fork's method we have

$$\int_0^\infty e^{-zt}v_t dt = \frac{Ny_0\int_0^\infty e^{-zt}A_t dt}{1 - N\int_0^\infty e^{-zt}A_t dt} := F_4(z)$$
$$\Rightarrow v_t = \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} e^{zt}F_4(z)dz$$

Analogously for y_t

$$\int_0^\infty e^{-zt}y_t dt = \frac{y_0\int_0^\infty e^{-zt}B_t dt}{1 - N\int_0^\infty e^{-zt}A_t dt}$$
$$\Rightarrow y_t = \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} e^{zt}F_5(z)dz.$$

Therefore, the integral equation of y_t is

$$y_t = N\int_0^t A_{t-\theta}y_\theta d\theta + B_t y_0.$$

On the other hand, taking $v_{t,0} = v_t$ assuming that this equality is not true on an interval of $[0, \epsilon]$ and the integral equation $\int_0^\epsilon v_{t,0}dt = y_0$ is not satisfied.

As a result

$$v_{t,0} = v_{t,0} - v_{\epsilon,0} + v_{\epsilon,0} = \int_0^t A_{t-\theta}v_\theta d\theta + A_t y_0.$$

Written differently

$$v_{t,0} = \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} e^{zt}F(z)dz$$

where $F(z) = \frac{y_0}{1-A}$ con $A = \int_0^\infty e^{-z\theta}A_\theta d\theta$.

Now, if v_t has no singularities, the Laplacian solution of $F_4(z)$ is a function without singularities and, therefore, the Laplacian of y_0 corresponds to the singularity.

A review of the parameters that were considered to study the pandemic that caused the Bubonic plague in Bombay, where Kermack and McKendrick (1927) implemented a mathematical model, in which they stated the following equations,

$$\begin{cases} \frac{dx}{dt} = -kxy \\ \frac{dy}{dt} = kxy - ly \\ \frac{dz}{t} = ly \end{cases}$$

There is some discrepancy in how these parameters were treated. Seg'un Bacaer (2012) says that "the Kermack and McKendrick model did not take seasonality into account". That is to say, when this model was first enunciated, the authors did not take into account the periodic variation of the model parameters with respect to time, such as the seasons, social behaviors, biological factors, among others. In addition, he comments that there is no explicit information on N (population), since it allows us to establish a certain possible size at the end of the epidemic, as well as the basic reproductivity number R_0.

Currently, there are several mathematical models that model the spread of epidemics. In continuing the study of the SIR model, it has now been more clearly stated and has undergone some modifications involving new parameters, which have been of great help in understanding the behavior of epidemics.

Therefore, the SIR model is described as follows,

$$\begin{aligned} \frac{dS}{dt} &= -r\beta S\frac{I}{N} \\ \frac{dI}{dt} &= r\beta S\frac{I}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I. \end{aligned}$$

Where the basic reproductivity number is the number of new infections that an infected person can cause in a susceptible population over its entire infectious period, where this number is only valid for homogeneous autonomous models and is denoted as:

$$R_0 = r \times \beta \times \frac{1}{\gamma} = \frac{r\beta}{\gamma},$$

In which the parameters describe

- r : Number of contacts per time unit
- β : probability of contact transmission
- $\frac{1}{\gamma}$: Infectiousness duration
- S : susceptible humans
- I : infectious humans
- R : recovered humans
- $N = S + I + R$

Thus, this model cannot be solved analytically. Where it is possible not to consider the equation of recovered humans R. Moreover, this model can estimate the final size of the epidemic.

Thus, the SIR endemic model can be enunciated, which unlike the SIR model, all equations must be taken into account, where there are two equilibrium points such as endemic and disease-free. Where the reproductivity number R_0 unlike the SIR model the duration of infection is defined as $\dfrac{1}{\gamma + \mu}$.

$$\frac{dS}{dt} = \Lambda - r\beta S\frac{I}{N} - \mu S$$
$$\frac{dI}{dt} = r\beta S\frac{I}{N} - \gamma I - \mu I$$
$$\frac{dR}{dt} = \gamma I - \mu R$$

The endemic break-even point defines the following:

$$S_{ee} = \frac{\Lambda(\gamma+\mu)}{r\beta\mu}$$
$$I_{ee} = \frac{\Lambda(r\beta-(\gamma+\mu))}{r\beta(\gamma+\mu)}$$
$$R_{ee} = \frac{\gamma\Lambda(r\beta-(\gamma+\mu)}{r\beta\mu(\gamma+\mu)}$$

The disease-free break-even point defines the parameters as follows: $S_{efe} = \frac{\Lambda}{\mu}$, $I_{efe} = 0$ y $R_{efe} = 0$, where:

- Λ : Constant recruitment rate
- μ : Per capita disposal rate

Finally, the SEIR model adds a new main parameter based on the models that have been previously stated. Thus, the new parameter that is added refers to exposed (latent) humans E, in which it additionally derives the per capita rate of progression to the infectious state ϵ.

Where the reproducibility number is defined as

$$R_0 = r \times \beta \times \frac{1}{\gamma+\mu} \times \frac{\epsilon}{\epsilon+\mu}$$
$$= \frac{r\beta\epsilon}{(\gamma+\mu)(\epsilon+\mu)}$$

with $\frac{\epsilon}{\epsilon+\mu}$: probability of surviving the exposed stage.

1.3.1 Kermack and McKendrick assumptions

Kermack and McKendrick propose that in limited, i.e. well-demarcated populations, epidemics begin their course and eventually end, furthermore, they rename the $\phi = k$, y $\psi = l$. Thus the equations would be given by

$$\begin{cases} \frac{dx}{dt} = -kxy \\ \frac{dy}{dt} = kxy - ly \\ \frac{dz}{dt} = ly \end{cases}$$

and $x, y, z = N$, in this way we can have that: $\frac{dx}{dz} = -\frac{k}{l}x$, where

$$\frac{dz}{dt} = l(N-x-z) = l(N - x_0e^{-\frac{k}{l}z} - z)$$

Kermack and McKendrick (1927), in their work assume that deaths from natural causes during the course of the epidemic are negligible and that the population growth rate is zero.

At this point, returning to the last equation, it can be said that it is not possible to obtain z as an explicit function of t. An expansion is used for $e^{-\frac{k}{l}z}$, since in initial stages of the epidemic the number of individuals removed z, is small compared to the total population N.

$$e^{-\frac{k}{l}z} = 1 - \left(\frac{k}{l}\right) z + \frac{1}{2!}\left(\frac{k}{l}\right)^2 z^2 - \cdots$$

results in a first-order approximation

$$e^{-\frac{k}{l}z} \approx 1 - \frac{k}{l}z$$

then,

$$\frac{dz}{dt} \approx l\left(N - x_0 + \frac{k}{l}x_0 z - \frac{x_0}{2}\left(\frac{k}{l}\right)^2 z^2 - z\right),$$

If $y_0 = N - x_0$ for a very small value of y_0. This Riccati equation has an explicit solution for z and gives the number of deaths per unit time.

$$z = \frac{l^2}{k^2 x_0}\left(\frac{k}{l}x_0 - (-b)^{1/2} tanh\left(\frac{\sqrt{-b}}{2}lt - \phi\right)\right),$$

where

$$\phi = tanh^{-1}\left(\frac{\frac{k}{l}x_0 - 1}{\sqrt{-b}}\right),$$

$$\sqrt{-b} = \left(\left(\frac{k}{l}x_0 - 1\right)^2 + 2x_0 y_0 \frac{k^2}{l^2}\right)^{1/2}.$$

.

Subsequently, towards the end of the epidemic we have:

$$z = \frac{2l}{kx_0}\left(x_0 - \frac{l}{k}\right),$$

let us take into account that y_0 is small and negligible at the beginning of the epidemic if we compare it with x_0. Now, $x_0 \approx N$ and $x_0 = \frac{l}{k}$, at this point there is no room for an epidemic, but if N slightly exceeds this value a small epidemic would occur, i.e., we would have $N = \frac{l}{k} + n$, in this case

$$z = \frac{2nl}{Nk} = 2n - 2\frac{n^2}{N},$$

if n is very small, $N_0 = \frac{l}{k}$ would be the density threshold. It should be noted that no epidemic can occur if the population density does not exceed this value.

Now let's focus on the expression

$$-log\frac{1-p}{1-\frac{y_0}{N}} = ApN,$$

here $\frac{y_0}{N}$, is a portion of infected and p is a fraction of infected population during the epidemic; and represent how they are connected through the spread, considering the interaction with the transmission and removal rates. Describing the expression in a little more detail, $1 - p$ is the

population that was not affected by the epidemic, and $1 - \frac{y_0}{N}$ as the initial fraction of the population that was susceptible.

If A is large, even a small fraction of initial infected ($fracy_0N$) can lead to significant epedimia (p close to 1), and if A is small the epidemic is rapidly self-limiting.

If $p > 0$, then $N_0 > \frac{1}{A}$ would generate an epidemic.

1.4 Conclusion

Mathematical models have made it possible to understand human behavior in order to predict and manipulate new schemes of development and social behavior through mathematical formulations, where the capacity for analysis facilitates the obtaining of new perspectives applicable to diverse areas of knowledge.

References

[1] Kermack, W. O., & McKendrick, A. G., *A contribution to the mathematical theory of epidemics.*Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, **115(772)**, (1927), pp. 700-721.

[2] Bacaër, N., *Le modèle de Kermack et McKendrick pour la peste à Bombay et la reproductivité nette d'un type avec saisonnalité.* J. Math. Biol, **64(3)**, (2012), pp. 403-422.

Quantum error-correcting codes

Luis Pablo Colmenar ♭, Vicent Miralles Lluch ♮, and Alberto Rodríguez Durá ♯

(♭) Universitat de València, Carrer del Dr. Moliner, 50, 46100 Burjassot, València;
cololuis@alumni.uv.es
(♮) Universitat Politècnica de València, Camí de Vera, Algirós, 46022 València;
vicent@vmiralles.com
(♯) Universtiat de València, Carrer del Dr. Moliner, 50, 46100 Burjassot, València;
albertord.bg@gmail.com

1.1 Abstract

In information transmission and processing, error correction plays a crucial role in both classical and quantum systems. This paper begins with an introduction to classical coding theory, addressing how error-correcting codes allow error detection and correction in noisy channels, highlighting methods such as maximum likelihood decoding in binary symmetric channels. Next, we explore the fundamentals of quantum computing, emphasising how qubits, affected by phenomena such as quantum noise and decoherence, demand new correction strategies. Finally, we will focus on quantum error-correcting codes, including bit-flip, phase-flip, and Shor's code, which combine classical concepts with quantum adaptations. We will also examine CSS codes, which leverage classical algebraic structures to address the challenges of quantum noise, showing their relevance in the construction of stable and functional quantum systems.

1.2 Basics of coding theory

In this section we will offer a brief introduction to coding theory, giving the main definitions and results, and introducing the minimum viable content to understand quantum coding. We will follow the approach made in [8], and taking some ideas from [2].

1.2.1 Introduction

Definition 1. A q-ary code $\mathcal{C}$ written in the alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_q\}$ of q symbols is a finite subset of

$$\mathcal{A}^n = \{(a_{i_1}, a_{i_2}, \ldots, a_{i_n}) \mid a_{i_j} \in \mathcal{A}, \text{ for } j = 1, 2, \ldots, n\}.$$

The elements of $\mathcal{C}$ are called *codewords*.

We will use juxtaposition and write $a_{i_1} a_{i_2} \ldots a_{i_n}$ instead of tuple notation.
To transmit a message written in a specific alphabet, referred to as the *font alphabet*, we first encode the information into a *code alphabet*, which does not necessarily need to match the font alphabet.

Definition 2. An *encoding scheme* of a font alphabet $\mathcal{S}$ is a pair $(\mathcal{C}, \varphi)$ consisting of a q-ary code $\mathcal{C}$ and a bijective map $\varphi\colon \mathcal{S} \to \mathcal{C}$ called *encoding*.

Let's see an example of an encoding scheme.

Example 1. Let $\mathcal{S} = \{\texttt{A},\texttt{B},\texttt{..},\texttt{Z},\texttt{␣}\}$ be the font alphabet consisting on all the letters in the spanish alphabet and the blank space, and the code $\mathcal{C} = \{00, 01, ..., 27\}$ written in the alphabet $\mathcal{A} = \{0, 1, ..., 9\}$. Then, the map $\varphi\colon \mathcal{S} \to \mathcal{C}$ given by

$$\varphi(\texttt{A}) = 00,\ \varphi(\texttt{B}) = 01,\ \varphi(\texttt{C}) = 02,\ \ldots,\ \varphi(\texttt{Z}) = 26,\ \varphi(\texttt{␣}) = 27,$$

is an encoding of $\mathcal{S}$. Therefore, we may use this encoding to encode any message, such as:

$$\texttt{VIVA␣EL␣ALGEBRA}$$

This message is encoded as

$$22082200270411270011060401 1800.$$

We will typically consider $\mathcal{A}$ to be the finite field of q elements $\mathbb{F}_q$, where q is a prime power. Consequently, unless otherwise specified, the alphabet under consideration will be $\mathbb{F}_q$.

The following example is one of the simplest classical code.

Example 2. Consider the code in which each element $\alpha \in \mathbb{F}_q$ is encoded as $\alpha\alpha\cdots\alpha$, an n-tuple of n α symbols. For instance, the binary repetition of length 2 is simply $\mathsf{Rep}_2(2) = \{00, 11\}$. In general, we define

$$\mathsf{Rep}_q(n) = \{\overbrace{00\cdots0}^{(n)}, \overbrace{11\cdots1}^{(n)}, \ldots, \overbrace{(q-1)(q-1)\cdots(q-1)}^{(n)}\}.$$

This code is called a repetition-type code.

1.3 Error-correcting codes

To transmit information encoded as codewords, a communication channel is required. However, errors may occur during transmission, leading to a situation where the word received from the channel does not match the one initially sent. Therefore, a process is needed to identify and, if possible, correct these errors. This is what we call decision scheme.

1.3.1 Introduction and errors in the communication channel

To create an effective decision scheme, we must first introduce some definitions and concepts related to the search for suitable error-correcting codes.

Definition 3. Let $\mathcal{A} = \{a_1, a_2, ..., a_q\}$ be a font alphabet. A set of conditional probabilities

$$\{P(a_j \text{ received} \mid a_i \text{ sent}) \mid 1 \leq i, j \leq q\}$$

is said to be a *probabilities channel* regarding $\mathcal{A}$.

Definition 4. A *communication channel* is a pair $(\mathcal{A}, \mathfrak{F})$ consisting of a font alphabet $\mathcal{A}$ and a probabilities channel $\mathfrak{F}$ regarding $\mathcal{A}$.

Depending on the nature of a communication channel, we can classify different types of these. We will show some of the most important channels:

1. A *no-memory channel* is a channel $(\mathcal{A}, \mathfrak{F})$ in which for two words $x = x_1x_2\cdots x_n$, $y = y_1y_2\cdots y_n \in \mathcal{A}^n$ it is verified

$$P(x \text{ received} \mid y \text{ sent}) = \prod_{i=1}^{n} P(x_i \text{ received} \mid y_i \text{ sent}).$$

2. A *q-ary symmetric channel* is a no-memory channel $(\mathcal{A}, \mathfrak{F})$ in which $|\mathcal{A}| = q$ that satisfies:
   - For every $1 \leq i \leq q$ and some $p \in [0, 1]$ it is verified

$$P(a_i \text{ received} \mid a_i \text{ sent}) = 1 - p.$$

   - The larger the size of $\mathcal{A}$, the larger p could be in the sense that

$$1 - p > \frac{1}{q} \text{ iff } p < \frac{q-1}{q}.$$

1.3.2 Decision schemes: Maximum likelihood method

As we mentioned above, we need methods which helps us to identify and correct errors in the information transmission. We will study the *maximum likelihood method*.

Let $(\mathcal{A}, \mathfrak{F})$ be a no-memory channel and $\mathcal{C}$ a code in $\mathcal{A}$. If we have received a word $x \in \mathcal{A}^n$, the maximum likelihood method consists of decoding the word x as a word $c_x \in \mathcal{C}$ which verifies

$$P(x \text{ received} \mid c_x \text{ sent}) = \max_{c \in C}\{P(x \text{ received} \mid c \text{ sent})\}.$$

Note that the word c_x may not be unique. Then we can distinguish two different maximum likelihood methods based on this fact:

1. *Complete:* We choose one of the words c_x arbitrarily.
2. *Incomplete*: It is admitted that an error has occurred and retransmission of the message is requested.

Example 3. Let $\mathcal{C} = \{00, 10, 01\}$ binary code associated to a binary symmetric channel of probabilities with error probability $p < \frac{1}{2}$. If we receive the word 11, then note that:

1. $P(11 \text{ received} \mid 01 \text{ sent}) = p(1-p)$,
2. $P(11 \text{ received} \mid 10 \text{ sent}) = p(1-p)$,
3. $P(11 \text{ received} \mid 00 \text{ sent}) = p^2$.

Therefore, the complete maximum likelihood method tells us that we can either choose 10 or 01 as a decoded word. Now consider the code $\mathcal{C} = \{00, 10\}$. In this case, applying the same algorithm 10 is the unique codeword that verifies the maximum likelihood condition. Hence, we choose 10 as the decoded word.

In some cases where we can find an easy algorithm which determines the words that verify the maximum likelihood condition.

Theorem 1. *Consider a binary symmetric channel where the probability error in a symbol is $p < \frac{1}{2}$. Then, the maximum likelihood method consists of choosing the codewords that have the least number of different components from the received word.*

This proof can be found in [7].

At this point, we could ask ourselves what kind of benefits coding offers us. As in the hypotheses of the previous theorem, we suppose that we want to transmit information through a binary symmetric channel where the probability error in a symbol is $p < \frac{1}{2}$.

If we want to transmit a single bit to a receiver without using a specific code, then the receiver receives the wrong bit with probability p (we are assuming that each bit gets flipped independently). However, we now suppose that we use the binary repetition code of length 3 and we try to decode the message through the maximum likehood method. By Theorem 1, the probability of decoding correctly a received word is the probability that at most one of the three bits gets flipped. The probability that two bits get flipped during transmission is $3p^2(1-p)$. Furthermore, the probability that three bits get flipped during transmission is p^3. Thus, the probability that we do not decode correctly the information is:

$$3p^2(1-p) + p^3 = 3p^2 - 2p^3$$

In the following graph we compare the probability of not decoding correctly when the repetition code is used and when it is not used:

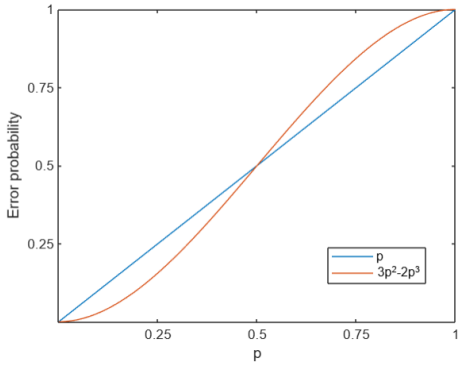


Figure 1.1: Error probability in decoding

Note that for $p < \frac{1}{2}$ the process of encoding and decoding results in a decrease in the probability of receiving a wrong bit. This does not mean the code completely vanishes the probability of error, but rather that it significantly decreases the likelihood of it. In fact, observe that when $p \geq \frac{1}{2}$ the code behalves the opposite, it actually increases the likelihood of receiving the wrong bit.

1.4 Linear and dual codes

In order to work with quantum codes, it is essential to study a specific type of code known as linear codes. Additionally, we must discuss dual codes, which are directly related to linear codes, as they are derived from them. Furthermore, dual codes are fundamental for the construction and study CSS codes, a particular class of quantum error-correcting codes.

Definition 5. A subspace $\mathcal{C}$ of $\mathbb{F}_q^n$, where $n \in \mathbb{N}$, is called a q-ary linear code.

Note that a q-ary linear code is a q-ary code. Since a q-ary linear code is a linear subspace it has a dimension. Thus, if $\mathcal{C} \subseteq \mathbb{F}_q^n$ be a q-ary linear code of dimension k we say that $\mathcal{C}$ is a q-ary $[n,k]$-code. For example, the repetition code $\mathsf{Rep}_2(2) = \{00, 11\}$ is a $[2,1]$-code since $\mathsf{Rep}_2(2) = \langle 11 \rangle$.

The notions of bilinear and sesquilinear forms can also be defined for finite fields.

Definition 6. The function $\langle \cdot, \cdot \rangle : \mathbb{F}_q^n \times \mathbb{F}_q^n \to \mathbb{F}_q$ given by

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n, \text{ for all } x, y \in V,$$

is a non-degenerated symmetric bilinear form.

Now, we will define a dual code using the orthogonal subspace of a subspace of $\mathbb{F}_q^n$ and its properties.

Definition 7. Let $\mathcal{C}$ be a q-ary $[n,k]$-code, then define

$$\mathcal{C}^{\perp} \coloneqq \left\{ x \in \mathbb{F}_q^n \mid \langle x, c \rangle = 0, \text{ for all } c \in \mathcal{C} \right\}.$$

Note that $\mathcal{C}^{\perp}$ is a subspace of $\mathbb{F}_q^n$. Thus, it is a linear code itself.

Theorem 2. *Let f be a non-degenerated symmetric bilinear form in $\mathbb{F}_q^n$ and suppose that $\mathcal{C}$ is a subspace of $\mathbb{F}_q^n$. Then* $\dim(\mathcal{C}) + \dim(\mathcal{C}^{\perp}) = n$.

Definition 8. Let $\mathcal{C}$ be a q-ary $[n,k]$-code. Then the *dual code* of $\mathcal{C}$ is the q-ary $[n, n-k]$-code $\mathcal{C}^{\perp}$.

1.5 Basics of quantum computation

1.5.1 Quantum postulates

Although our goal is not to explore quantum codes from a physical perspective, it is essential to understand and keep in mind the postulates of quantum mechanics, as they provide the mathematical framework on which we will base the entire development of this work.

We will not delve into the origins or physical interpretations of these postulates. Instead, we adopt them as axiomatic, and from now on all mathematical aspects must be ruled by these postulates.

Postulate 1. *Every isolated quantum system has a Hilbert space associated (specifically, a complex vector space with an inner product) known as the state space of the system. The system is fully determined at any given moment by its state vector, which is a unit vector in the state space.*

The most basic quantum system is the qubit. This system has associated a 2-dimensional state space with an orthonormal basis $\{|0\rangle, |1\rangle\}$, which we will reefer as computational state basis. Therefore, we can express each state of the system as:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \quad \alpha, \beta \in \mathbb{C} \text{ such that } |\alpha|^2 + |\beta|^2 = 1,$$

where the last property guarantees that our vector is unitary. The qubit is the fundamental quantum system for the development of quantum codes.

Postulate 2. *The evolution of a closed quantum system is governed by a unitary transformation. Specifically, the state of the system $|\psi_0\rangle$ at time t_0 is related to the state of the same system $|\psi_1\rangle$ at time t_1 through a unitary matrix U, which only depends on the times t_0 and t_1 in the following way:*

$$|\psi_1\rangle = U|\psi_0\rangle.$$

Throughout this discussion, we consider only closed quantum systems. However, if we attempt to observe the system, our interaction with it causes it to stop being a closed system. Therefore, its evolution is no longer unitary over time. The following postulate describes what happens when measurements are performed on a closed quantum system.

Postulate 3. *A quantum measurement is characterised by a set of operators $\{M_m\}$, referred to as measurement operators, which act on the system. The index m corresponds to each of the*

possible outcomes of the measurement (already fixed beforehand). If the state of the system is $|\psi\rangle$ *, the probability of obtaining the result* m *is given by:*

$$p(m) = \langle\psi|M_m^* M_m \,|\, \psi\rangle$$

And the system collapses into the state:

$$\frac{M_m|\psi\rangle}{\sqrt{\langle\psi|M_m^* M_m \,|\, \psi\rangle}}$$

To be a valid measure,the measurement operators must verify the "completeness condition" to ensure that the probabilities of all possible outcomes sum to one. Mathematically, this is expressed as:

$$\sum_m \langle\psi|M_m^* M_m \,|\, \psi\rangle = \sum_m p(m) = 1$$

Theorem 3. *Let* $\{|\psi_i\rangle\}_{i=1,\dots,n}$ *be a set of states in a quantum system. For each* $j = 1, \dots, n$ *it is possible to distinguish the state* $|\psi_j\rangle$ *of the system through a quantum measurement if and only if the states in the set are orthonormal.*

Consequently, we can choose other bases composed of orthonormal states on which measurements can also be realised. An example of this fact is that we could make measurements through the basis $\{|+\rangle, |-\rangle\}$ if it is convenient. Where:

$$|+\rangle = \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle \quad \text{and} \quad |-\rangle = \frac{1}{\sqrt{2}}|0\rangle - \frac{1}{\sqrt{2}}|1\rangle$$

Yet we only described how it works a single qubit. To establish how to concatenate several qubits, as in classical computation, we need another postulate.

Postulate 4. *The state space of a composite quantum system is the tensor product of the state spaces of the individual subsystems that constitute it.*

Therefore, a system composed by n qubits, has associated a Hilbert space of dimension 2^n. Hence, if $|\psi\rangle$ is the state of that system we will write:

$$|\psi\rangle = \sum_{i\in\mathbb{F}_2^n} a_i|i_1\rangle \otimes \cdots \otimes |i_n\rangle = \sum_{i\in\mathbb{F}_2^n} a_i|i_1 \dots i_n\rangle,$$

where $i = (i_1 \dots i_n) \in \mathbb{F}_2^n$, and $\sum_{i\in\mathbb{F}_2^n} |a_i|^2 = 1$, to ensure the unitary condition.

1.5.2 Quantum logic gates and basic circuits

Quantum gates are just linear unitary transformations acting upon the state space of a quantum system. The linearity of these transformations is essential to preserve the physical characteristics of the system. Now, we show an example of how we are going to represent these gates. As we mentioned earlier, they are just a linear unitary transformation that maps our state $|\psi_0\rangle$ at time t_0 to our new state $|\psi_1\rangle$ at time t_1. In the figure below, the movement of our state over time is represented by a wire, then a unitary transformation is applied, which transforms it, and then the qubit continues to move.

For each qubit, there are several matrices that can modify its state, not like in classical computation, where the only nontrivial gate for a bit is the NOT gate (which swaps 0 to 1 and 1 to 0). Another important consequence of being unitary is that all quantum transformations are reversible in the sense that the inverse of any quantum gate can be easily obtained. In other

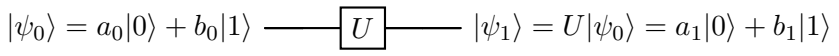


Figure 1.2: Evolution of a quantum state over time.

words, when we concatenate gates and get some results we can follow the same path backwards to give them an interpretation.

Due to the linearity of quantum gates, knowing the image of the basis states $|0\rangle$ and $|1\rangle$, allows us to determine the image of any arbitrary qubit state. So based on this reasoning, we will only present the most common gates for a single qubit. The X gate sends $|0\rangle$ to $|1\rangle$ and $|1\rangle$ to $|0\rangle$. The Y gate sends $|0\rangle$ to $|1\rangle$ and $|1\rangle$ to $-|0\rangle$. The Z gate sends $|0\rangle$ to $|0\rangle$ and $|1\rangle$ to $-|1\rangle$. And finally the H gate sends $|0\rangle$ to $|+\rangle$ and $|1\rangle$ to $|-\rangle$. In terms of its unitary matrices, they may be represented as:

$$X = |1\rangle\langle 0| + |0\rangle\langle 1| = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \sigma_x,$$

$$Y = |1\rangle\langle 0| - |0\rangle\langle 1| = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = -i\sigma_y,$$

$$Z = |0\rangle\langle 0| - |1\rangle\langle 1| = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \sigma_z,$$

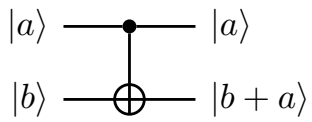
$$H = |+\rangle\langle 0| + |-\rangle\langle 1| = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

where $\{\sigma_x, \sigma_y, \sigma_z\}$ are the Pauli matrices.

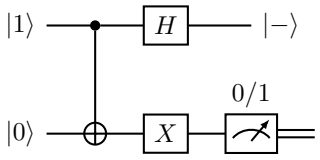
So far we have only talked about logic gates acting on a single qubit. To generalise this to more than a single qubit we need to talk about circuits. A circuit is a model for quantum computing in which a computation is a sequence of logic gates. We read them from left to right. Each line represents a wire in the circuit, i.e., the evolution of a qubit (which is not necessarily a physical wire, it may be the movement of a photon in the space). By default, we assume that the input state of the circuit is a state of the computational basis. Quantum and classical circuits share most of their properties. However, quantum circuits must fulfill some specific properties in order not to violate quantum postulates. We will not delve deeper, since the circuits we are going to use will satisfy these requirements. The Figure 1.2 is an example of a circuit acting upon one qubit.

The most basic and important circuit is known as the CNOT gate, which acts upon two qubits. We understand it as a generalization of the classical XOR gate. It works as follows: the first qubit, *control qubit*, indicates whether the second qubit, *target qubit*, should be modified. That is, if the control qubit is in the state $|0\rangle$ the CNOT leaves the state of the target qubit unchanged. Otherwise, the CNOT gate flips the state of the target qubit (i.e., $|0\rangle \mapsto |1\rangle$ and $|1\rangle \mapsto |0\rangle$). This logic gate is key in the understanding of the most basics quantum correcting codes, as we will soon see. Figure 1.3 shows the CNOT gate and a circuit example.

In the second circuit in the figure we implemented a measurement of a qubit $\alpha|0\rangle + \beta|1\rangle$. The measurement is depicted as a meter that gives as outcome a classical bit 0 or 1 with probability $|\alpha|^2$ or $|\beta|^2$ respectively. These classical bits are manipulated via a classical wire represented as a double wire.



(a) CNOT gate, where the sum is modulo 2 and a, b are 0 or 1.



(b) Example of a circuit where several circuit elements are implemented.

Figure 1.3: Some circuits examples.

1.5.3 Non-cloning Theorem

On the basis of the above, the reader might propose an approach to measure a quantum system without altering it: replicating the quantum system and performing measurements on the copies. Through statistical analysis of the results, the original state could theoretically be determined. This technique is widely used in classical computation, where copying bits a common practice, and even more in classical correcting codes. In classical computation, it is possible to copy bits using an XOR gate by introducing a new bit in state 0 and performing a modulo 2 sum with the bit to be copied. However, from the perspective of quantum information, it is fundamentally impossible to copy any qubit using unitary transformations as the following theorem states.

Theorem 4 (Non-cloning theorem)**.** *An unknown quantum system cannot be copied using unitary transformations.*

A very important remark is that this theorem tells us that in general we are not able to copy as shown in the proof, but for certain values of α, β we can do it.

1.6 Quantum error-correcting codes

As we have said, the aim of coding theory is to provide a way of protecting information from errors made when sending it through a noisy channel. In the case of quantum information, prevent them from quantum errors acting upon qubits corrupting the information.

Therefore, following the line gone with classical codes we could try to construct a code by repetition. In other words, add redundancy to the qubits we want to send so that we detect and recover the original information.

Unfortunately, this strategy would violate the non-cloning principle (Theorem 4). Indeed, if $|\psi\rangle$ is a quantum state, then we cannot directly clone it via a map such that

$$|\psi\rangle \mapsto |\psi\psi\psi\rangle$$

to obtain a 3-quantum state. However, we can achieve the same result using a different approach.

Right before tackle this problem, we must keep in mind the most major differences between classical and quantum codes:

1. No cloning: Theorem 4 states that we cannot clone an arbitrary quantum state. Furthermore, even if we could, we would not be able to measure or compare the states after the sending.
2. A single qubit can be subject to an infinite type of errors.
3. Performing measurements modify their state, so we cannot observe the states to detect errors.

Fortunately, all these problems are solvable and we will be able to construct a theory of quantum codes capable of addressing these issues. We will first look at simple quantum codes.

1.6.1 Repetition code for qubits

In the classical case, the binary repetition code of length code 3 is obtained by encoding the elements of the alphabet $\mathbb{F}_2 = \{0, 1\}$ as follows:

$$0 \mapsto 000 \quad \text{and} \quad 1 \mapsto 111.$$

This encoding allows us to correct up to one error by majority decision (see Theorem 1). In other words, we decode:

$$000,\ 001,\ 010,\ 100 \ \text{ as } \ 0$$

and

$$111,\ 011,\ 110,\ 101, \ \text{ as } \ 1.$$

The same procedure for constructing a quantum repetition code cannot be performed due to the non-cloning theorem. However, we apply a different strategy that leads to the same result.

Suppose $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ is a quantum state. We wish to add redudancy to it knowing that we cannot clone an arbitrary state. Thus, we take two auxiliary qubits in state $|0\rangle$ and add them to the two single states. That is,

$$|\psi\rangle|00\rangle = (\alpha|0\rangle + \beta|1\rangle)|00\rangle = \alpha|000\rangle + \beta|100\rangle.$$

Now we simply encode by applying a CNOT gate twice as shown the circuit in Figure 1.4.

Note that no-cloning theorem has not been violated since we have not cloned any qubit. We have just modified a 3-quantum state. Each qubit of the 3-quantum state is called a *physical qubit*.

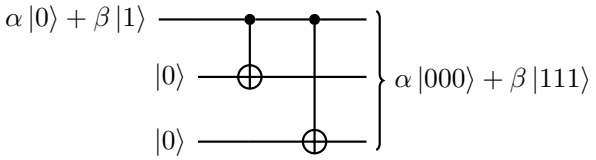


Figure 1.4: Encoding of a quantum state of a qubit.

In fact, the above encoding is an isometric embedding of a 2-dimensional Hilbert space $\mathcal{H}$ with basis $\{|0\rangle, |1\rangle\}$ in the subspace spanned by $\{|000\rangle, |111\rangle\}$ of a Hilbert space $\mathcal{H}'$ of dimension 8. The isometry is precisely $U = |000\rangle\langle 0| + |111\rangle\langle 1| \colon \mathcal{H} \to \mathcal{H}'$ given by

$$U(\alpha|0\rangle + \beta|1\rangle) = \alpha|000\rangle + \beta|111\rangle, \text{ for all } \alpha, \beta \in \mathbb{C}.$$

The picture in Figure 1.5 represents the action of U.

1.6.2 Bit-flip code

Suppose we want to send a qubit through a channel that transforms a state $|\psi\rangle$ into a state $X|\psi\rangle$ with probability p, and leaves it untouched with probability $1 - p$. That is, it exchanges the amplitudes of $|\psi\rangle$. This channel is called the *bit-flip channel*. Now we construct the *bit-flip code* which protects qubits against the noise from this channel.

Now we take any state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ and using the previous procedure to encode it so that the bit-flip code is

$$\mathcal{C} = \{\alpha|000\rangle + \beta|111\rangle \mid \alpha, \beta \in \mathbb{C} \text{ and } |\alpha|^2 + |\beta|^2 = 1\}.$$

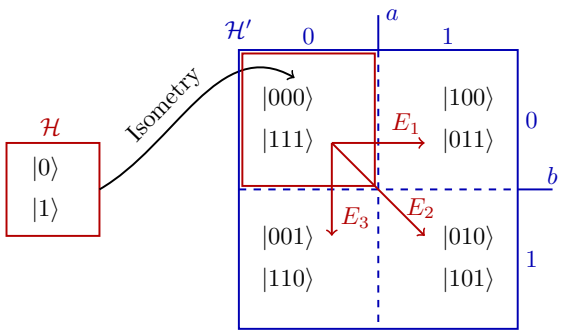


Figure 1.5: Representation of the encoding.

Once the information has been encoded, it is sent through a bit-flip channel that produces bit-flip errors in each physical qubit independently with probability p. The possible states received will depend on the physical qubits on which an error occurs. For instance, we can receive the state with error $\alpha|100\rangle + \beta|011\rangle$ with probability $p(1-p)^2$. The circuit in Figure 1.6 details the encoding, error detection, and error correction.

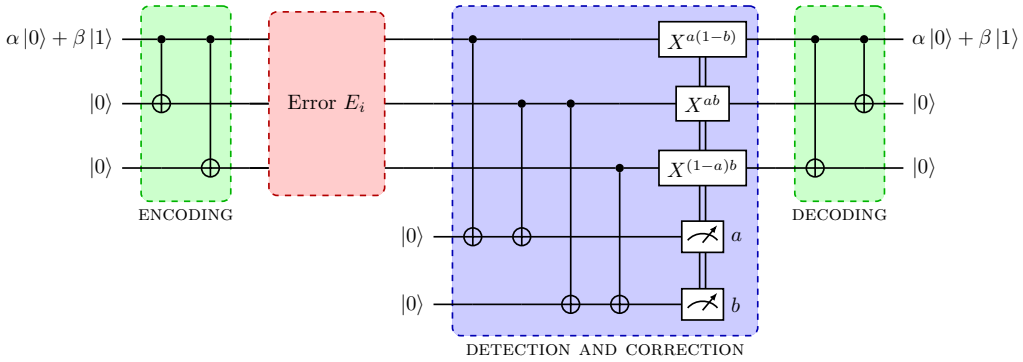


Figure 1.6: Circuit representing the encoding of a qubit bit-flip code, the sending of the encoded word through a bit-flip channel and the detection and correction of the error by the receiver of the message.

In order to recover the original information, we approach it in several ways. One way is to carry out a quantum measurement of the projections:

$$
\begin{aligned}
P_0 &= |000\rangle\langle 000| + |111\rangle\langle 111| \text{ no error} \\
P_1 &= |100\rangle\langle 100| + |011\rangle\langle 011| \text{ bit-flip on qubit one} \\
P_2 &= |010\rangle\langle 010| + |101\rangle\langle 101| \text{ bit-flip on qubit two} \\
P_3 &= |001\rangle\langle 001| + |110\rangle\langle 110| \text{ bit-flip on qubit three}
\end{aligned}
$$

Note that this family of projections $\{P_0, P_1, P_2, P_3\}$ satisfies the conditions of postulate 3.

For $i = 1, 2, 3$ suppose that a one bit-flip error E_i occurs and the state after this error is $|\psi\rangle$. Then the result of these measurements will be i with probability $\langle\psi|P_i|\psi\rangle = 1$. Now, since $P_i|\psi\rangle = |\psi\rangle$ (we are assuming the error E_i has occurred), the state is left unchanged. If no error occurs, then we will obtain 0 with probability 1 and the codeword will remain unchanged. Thus, in total, we will have four possible outcomes (error syndromes) which will specify the

position on which the error occurs. Then we simply correct by applying the bit-flip gate to the corresponding position and recover the codeword (recall that $X^2 = I$).

For example, suppose that the state $|\psi\rangle = \alpha|010\rangle + \beta|101\rangle$ is received. Then $\langle\psi|P_2|\psi\rangle = 1$ with syndrome 2 and $P_2|\psi\rangle = |\psi\rangle$. Thus, we recover the codeword by applying $E_2 = I \otimes X \otimes I$.

This process can be better understood by noting that we are projecting orthogonally the received state onto the four different Hilbert subspaces determined by the four distinct types of error. Once the subspace is detected, we can apply the error to bring the received state back into the code subspace (see again Figure 1.5).

Another and better approach to error detection and correction is the one implemented in the circuit in Figure 1.6. It works as follows:

We take two auxiliary qubits as we do when encoding and apply to each of them two CNOT gates. The first two CNOT gates check the parity of the first two physical qubits, and the next two check the parity of the last two physical qubits. Then we measure thess auxiliary qubits in the computational basis and obtain 0 or 1 in each case. These values form a pair ab which is the error syndrome. Thus, we will apply the corresponding correction according to the error syndrome. The following table lists the possible states received, the corresponding syndromes and their corrections:

State	**Syndrome** ab	**Correction** E_i
$\alpha\|000\rangle + \beta\|111\rangle$	00	$I \otimes I \otimes I$
$\alpha\|100\rangle + \beta\|011\rangle$	10	$X \otimes I \otimes I$
$\alpha\|010\rangle + \beta\|101\rangle$	11	$I \otimes X \otimes I$
$\alpha\|001\rangle + \beta\|110\rangle$	01	$I \otimes I \otimes X$

Figure 1.7: Received states, syndromes and corrections.

If we go back to Figure 1.5 we see that the representation of the associated Hilbert space is divided into four regions. This division is made in terms of the parity of the first two physical qubits followed by the parity of the two second ones. So, in the end, what we are doing is, again, detecting to which subspace the received state belongs.

1.6.3 Phase-flip code

This code will allow us to treat a state that has been altered by a *phase-flip*, i.e., an unwanted Z action on the state.

Suppose we want to send a qubit through a channel that transforms a state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ into a state $Z|\psi\rangle = \alpha|0\rangle - \beta|1\rangle$ with probability p, and leaves it unchanged with probability $1 - p$. That is, it performs a phase change. This channel is called the *phase-flip channel*. Note that there is no equivalent classical channel to this, since classical channels do not possess any equivalent phase property.

At first glance, we would suspect to have to come up with a different strategy to construct the *phase-flip code*. The code that protects qubits against the noise from the phase-flip channel. It turns out that we do not need to work much. Indeed, note that phase-flip gate Z is related to bit-flip gate X through Hadamard gate H since $X = HZH$. Furthermore, observe that the plus and minus states, i.e.

$$|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) \quad \text{and} \quad |-\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle),$$

are eigenvectors of Z as $Z|\pm\rangle = |\mp\rangle$. In short, the Z gate acts upon plus and minus states as the X gate acts upon computational basis states. Thus, it is sufficient to implement the Hadamard gate in the bit-flip code circuit to make it suitable to correct phase-flip errors:

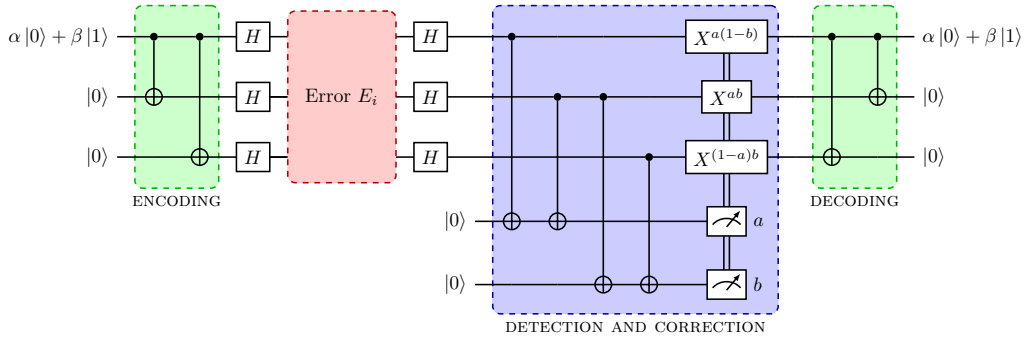


Figure 1.8: Circuit representing the phase-flip code process.

The encoded state that enters the transmission area (where errors can occur) is now $\alpha|+++\rangle + \beta|---\rangle$ due to the action of the Hadamard gate. Note that errors are transformed into orthogonal, and thus detectable states.

When it comes to error detection and correction is exactly the same as for the bit-flip code, since after the error we bring the plus and minus states back to the computational basis states.

1.6.4 Shor code

So far we have been able to find two different codes in order to correct bit-flip and phase-flip errors. The two codes are mutually exclusive, i.e. one can correct bit-flip errors but not phase-flip, and vice versa. Nonetheless, the Shor code is a clever way to implement both codes at the same time and be able to correct both bit-flip and phase-flip. This can be achieved by combining, or more precisely, concatenating bit-flip and phase-flip as follows:

First, we encode our state with the phase-flip code

$$\alpha|0\rangle + \beta|1\rangle \mapsto \alpha|+++\rangle + \beta|---\rangle$$

and then encode each physical qubit with the bit-flip code. In other words, we apply:

$$|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) \mapsto \frac{1}{\sqrt{2}}(|000\rangle + |111\rangle)$$
$$|-\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle) \mapsto \frac{1}{\sqrt{2}}(|000\rangle - |111\rangle)$$

Therefore, this results in:

$$\alpha|0\rangle + \beta|1\rangle \mapsto \alpha\frac{1}{2\sqrt{2}}(|000\rangle + |111\rangle)(|000\rangle + |111\rangle)(|000\rangle + |111\rangle)$$
$$+\beta\frac{1}{2\sqrt{2}}(|000\rangle - |111\rangle)(|000\rangle - |111\rangle)(|000\rangle - |111\rangle)$$

This encoding defines the Shor code which was proposed by Peter Shor in 1995 (see [11]). Now, suppose we send the encoded state through a channel that produces bit-flip, phase-flip and both at the same time on a single physical qubit. Then, in order to detect and correct a bit-flip error we simply study each block of three physical qubits separately. Indeed, if a bit-flip occurs in the first three physical qubits block then we apply the same strategy used for bit-flip code (this strategy can be understood as taking majority decision). Similarly for the rest of the blocks.

On the other hand, if a phase-flip occurs on any single qubit then it switches the sign of the entire block. For instance, if a phase-flip occurs on the first, second or third physical qubit, that

is, on the first block, then $|000\rangle + |111\rangle \mapsto |000\rangle - |111\rangle$ and vice versa. Thus, it flips the sign of the entire block. Therefore, note that the effect of a single phase-flip error on that space is the same regardless of in which of the three qubits it has occurred. This type of code where the location of the error is unknown is called *degenerated code*. Therefore, to correct a phase-flip error, we just take majority decision on the signs of the blocks. For example, if a phase-flip occurs in the sixth qubit then the α term would be

$$\frac{1}{2\sqrt{2}}(|000\rangle + |111\rangle)(|000\rangle - |111\rangle)(|000\rangle + |111\rangle)$$

Thus, by comparing the signs of the first two blocks and those of the last two, we detect the error in the second block and successfully correct it.

Now, when it comes to Y errors note that $Y = iXZ$. Therefore, since any global phase is irrelevant, any Y error is merely an X and Z error occurring on the same qubit. That is the reason why we have omitted the Y error so far. Thus, we detect and correct in two steps: first the bit-flip error and then the phase-flip error. This works because the two decisions we make are independent of each other.

We stated before that Shor's code can correct any type of error, including those not given by a unitary matrix. We will not go into details since this is not the goal of this section although we give the idea behind this result. Indeed, any matrix $U \in \mathcal{M}_2(\mathbb{C})$ can be written as a linear combination of the Pauli matrices and the identity matrix I_2. Thus, we can specify the action of U upon the k-th qubit and making the appropriate measurements we collapse, or more precisely project, the state (with error) onto the Pauli error and then use the strategies we have shown to detect and correct it (see all the details in [6] p. 60).

1.6.5 Some thoughts on general theory

The codes we have worked with so far provide the ideas underpinning the general theory of quantum error-correcting codes. The idea is the same: encode a quantum state via a unitary map into a state inside a quantum error correcting code, which is just a subspace of a larger Hilbert space. This subspace will have an associated projection map. For example, it is easy to check that the bit-flip code has the projection $P = |000\rangle\langle 000| + |111\rangle\langle 111|$ (see again the quantum measurements for bit-flip code) and the phase-flip code $P' = |{+}{+}{+}\rangle\langle{+}{+}{+}| + |{-}{-}{-}\rangle\langle{-}{-}{-}|$. Now, suppose that an encoded state undergoes an error, then, we perform a syndrome measurement so as to identify the type of error. Once identified the error syndrome, perform the correction to bring the state back to the code (see again Figure 1.5).

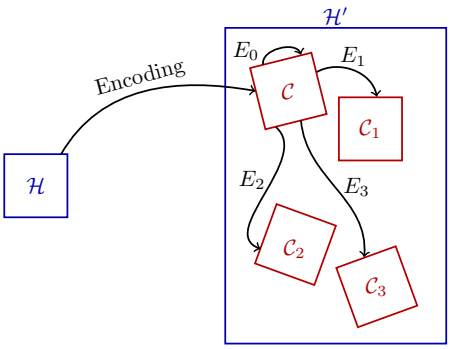
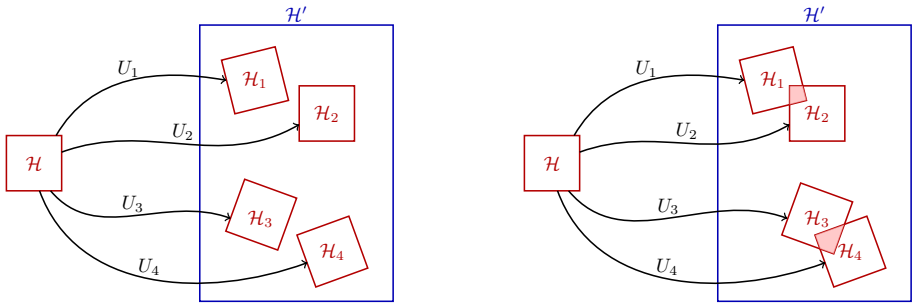


Figure 1.9: The original subspace is embedded (encoded) isometrically in a subspace of a larger Hilbert space and the errors (isometries) can shift the code to another orthogonal subspace.

The codes shall be designed in such a way that error syndromes can be distinguished by syndrome measurements. Thus, errors should send the code subspace into an orthogonal one. Moreover, these errors must preserve the properties of the subspace. In short, the error subspace should be an orthogonal copy of the code space. In this way, orthogonal words will be sent into orthogonal states so that we successfully recover the codeword after the error correction. Figure 1.9 gives an outline of the framework we wish to work with.

In general terms, the actions of the isometry and the errors can be summarised in isometries that map the original Hilbert space in subspaces of the larger Hilbert space. These isometries will preserve all the desired properties mentioned above. The following picture shows two situations that can occur:



(a) Good encoding, the subspaces do not overlap.

(b) Bad encoding, the subspaces do overlap

Figure 1.10: Graphical overview of the encoding and error process.

Since the action of isometry and errors is summarised in these four isometries, we then need the subspaces not to overlap in order to have the error codeword successfully recovered. Thus, we want the first situation to occur.

1.6.6 CSS codes

After this brief introduction to the first and most basic error-correcting quantum codes, it is natural to ask whether there is any relationship between classical codes and quantum codes. The answer is yes, as shown by the type of quantum codes defined below. These codes are called CSS by Calderbank, Steane and the aforementioned Shor (see [3] and [12]). They are the first family of codes that, like Shor's code, are capable of correcting any kind of error. In fact, this family happens to be part of a larger family of codes. The so-called *stabiliser codes*, which we will not delve deeper here.

Since a quantum code is after all a vector space, we can specify the length and dimension just as we do with linear codes. We write $[\![n, k]\!]_q$ to denote the length and dimension q^k of a quantum-error correcting code, that is, a q^k-dimensional subspace of the Hilbert space $\mathcal{H} = (\mathbb{C}^q)^{\otimes n} \cong \mathbb{C}^{q^n}$. We write double brackets to emphasise that this is a quantum code.

Theorem 5. *Suppose that $\mathcal{C}_1, \mathcal{C}_2 \subseteq \mathbb{F}_2^n$ are two $[n, k_1], [n, k_2]$-binary linear codes, respectively, such that $\mathcal{C}_2^\perp \subseteq \mathcal{C}_1$. Let $u \in \mathcal{C}_1$ and define the quantum state*

$$|u + \mathcal{C}_2^\perp\rangle := \frac{1}{\sqrt{|\mathcal{C}_2^\perp|}} \sum_{v \in \mathcal{C}_2^\perp} |u + v\rangle, \tag{1.1}$$

where the sum is the usual coordinate-wise addition in $\mathbb{F}_2^n$. Then, the code $\mathcal{Q}$ spanned by $\{|u + \mathcal{C}_2^\perp\rangle \, | u \in \mathcal{C}_1\}$ is a $[\![n, k_1 + k_2 - n]\!]$-quantum error-correcting code.

This proof can be found in [3], [12] and [9].

Definition 9. We call CSS code, and write $\mathrm{CSS}(\mathcal{C}_1, \mathcal{C}_2)$, to the $[\![n, k_1 + k_2 - n]\!]$-quantum error-correcting code $\mathcal{Q}$ from Theorem 5.

To conclude, let us look at an example of CSS code, the Steane code. It is the best known example of this type of code and is built using a classic *Hamming code* (a specific linear code).

Example 4 (Steane code)**.** Suppose that $\mathcal{C}$ is a $[7, 4]$-binary Hamming binary code and define $\mathcal{C}_1 = \mathcal{C}$ and $\mathcal{C}_2^{\perp} = \mathcal{C}^{\perp}$. It is easy to check that $\mathcal{C}_2^{\perp} \subseteq \mathcal{C}_1$. Thus, since $\mathcal{C}_1$ and $\mathcal{C}_2$ are $[7, 4]$-binary linear codes, from Theorem 5 it follows that $\mathrm{CSS}(\mathcal{C}_1, \mathcal{C}_2)$ is a $[\![7, 1]\!]$-quantum CSS code. Now, since $|\mathcal{C}_2^{\perp}| = 2^3 = 8$ it is also easy to check that:

$$\begin{aligned}\mathcal{C}_2^{\perp} = \{&0000000, 0001111, 0110011, 1010101,\\ &0111100, 1011010, 1100110, 1101001\}\end{aligned}$$

On the other hand, we have that the dimension is $\frac{|\mathcal{C}_1|}{|\mathcal{C}_2^{\perp}|} = 2$. Thus, since the null state (in this case $\mathbf{0} = 0000000$) is always part of the vector space, is sufficient to take $u \in \mathcal{C}_1$ such that $u \notin \mathcal{C}_2^{\perp}$. Taking $u \in \mathcal{C}_1$ to be $\mathbf{1} = 1111111$, the Steane code is:

$$\mathrm{CSS}(\mathcal{C}_1, \mathcal{C}_2) = \{\alpha|\mathbf{0} + \mathcal{C}_2^{\perp}\rangle + \beta|\mathbf{1} + \mathcal{C}_2^{\perp}\rangle \mid \alpha, \beta \in \mathbb{C} \text{ and } |\alpha|^2 + |\beta|^2 = 1\},$$

where,

$$\begin{aligned}|\mathbf{0} + \mathcal{C}_2^{\perp}\rangle = \frac{1}{2\sqrt{2}}(&|0000000\rangle + |0001111\rangle + |0110011\rangle + |1010101\rangle\\ &+|0111100\rangle + |1011010\rangle + |1100110\rangle + |1101001\rangle),\\ |\mathbf{1} + \mathcal{C}_2^{\perp}\rangle = \frac{1}{2\sqrt{2}}(&|1111111\rangle + |1110000\rangle + |1001100\rangle + |0101010\rangle\\ &+|1000011\rangle + |0100101\rangle + |0011001\rangle + |0010110\rangle).\end{aligned}$$

This last example sheds light on the usefulness of classical codes in quantum codes. More specifically, it leads us to find classical codes with the right properties that provide new quantum codes.

1.7 Conclusions

In this work, we have presented an introduction to the fundamental concepts of quantum error-correcting codes. Beginning with classical coding theory, we present the necessary mathematical foundations to understand the transition from classical to quantum error correction. We emphasized how the peculiarities of quantum systems necessitate new approaches compared to classical strategies.

We explored basic quantum codes, including bit-flip and phase-flip codes, and illustrated how the Shor code combines techniques to correct both bit-flip and phase-flip errors simultaneously. Moreover, we highlighted the importance of CSS codes, demonstrating how classical linear codes can be useful to construct quantum codes capable of protecting information against arbitrary quantum noise.

Ultimately, the theory of quantum error correction not only ensures the stability and reliability of quantum information processing but also is fundamental for the development of practical quantum technologies. Further advances in this field will be crucial for realizing more faithful quantum computers.

References

[1] Ball, S. & Centelles, A. & Huber, F., *Quantum error-correcting codes and their geometries*, In: *Ann. Inst. H. Poincare Comb. Phys. Interact.*, **10**, No. 2, (2023), pp. 337-405.

[2] Bierbrauer, J., *Introduction to coding theory*, Boca Raton : Chapman and Hall/CRC, (2023).

[3] Calderbank, A. & Shor, P., *Good quantum error-correcting codes exist*, In: *Phys. Rev. A* **54.2**, (1996), p. 1098.

[4] Dirac, P.A.M., *The Principles of Quantum Mechanics*, 4th ed., Oxford University Press, (1958), pp. 10-15.

[5] Galindo, A. & Martin-Delgado, M. A., *Information and computation: Classical and quantum aspects*, Reviews of Modern Physics **74**, No. 2, (2002), pp. 347.

[6] Lidar, D.; Brun, T., *Quantum Error Correction*, Cambridge University Press, (2013).

[7] Ling, S.; Xing, C., *Coding Theory: A first course*, Cambridge: Cambridge University Press, (2004).

[8] Neubauer, A., *Coding theory : Algorithms, architectures, and applications*, Hoboken, NJ : John Wiley and Sons, (2007).

[9] Nielsen, M.A.; Chuang, I.L., *Quantum Computation and Quantum Information*, Cambridge University Press, (2010).

[10] Preskill, J., *Quantum Computing in the NISQ Era and Beyond*, Quantum **2**, (2018), pp. 79.

[11] Shor, P. W., *Scheme for reducing decoherence in quantum memory*, In: *Phys. Rev. A* **52**, (1995), pp. R2493-2496.

[12] Steane, A., *Simple quantum error-correcting codes*, In: *Phys. Rev. A* **54.6**, (1996), p. 4741.

On the smooth skeleton of affine algebraic sets

Manuel García-García [♭] Oliver Navío-Velázquez 2 [♮]

(♭) Student, Universidad de Valencia, Calle del Doctor Moliner, 50, Burjassot, 46100, Valencia, España. magarg25@alumni.uv.es.

(♮) Student, Universidad de Valencia, Calle del Doctor Moliner, 50, Burjassot, 46100, Valencia, España. onave@alumni.uv.es.

1.1 Introduction

It is well-known that affine algebraic sets need not be a submanifold of the ambient space. The points where they fail to be a submanifold are called singular points. The systematic study of singularities began in the 1950s with the pioneering works of H. Whitney, R. Thom, J. Mather, and others. One of the earliest significant results regarding the singularities of affine algebraic sets is due to H. Whitney (see [1]). It states that every real (or complex) affine algebraic set is the disjoint union of finitely many real (or complex) analytic submanifolds of the ambient space, each having finitely many topological components. So, although we may not have a submanifold, we can divide our space into finitely many disjoint connected submanifolds. This is the simplest way of stratifying a singular space. Stratification is a simple but powerful technique for decomposing singular spaces into smaller but smooth pieces called strata. However, to make further progress in Singularity Theory, it became necessary to understand how these pieces fit together. This is where Whitney's Stratifications come into play. Whitney's conditions were introduced to address this issue, and they provided techniques, such as controlled vector fields, that allowed us to extend results from manifolds to Whitney's stratified spaces. For an overview of Stratification Theory, check [4] and for learning the controlled vector fields technique we refer the reader to [3].

In this work, we revisit Whitney's proof of this elementary but powerful result. This work is divided in three parts. Namely,

- First we define affine algebraic sets and study their basic properties. Next, we introduce the Zariski topology and prove that the affine space, endowed with this topology, is a noetherian topological space. This fact turns out to be crucial in Whitney's proof.

- In the second part, we revisit Whitney's Theorem. We will follow Whitney's proof, using slightly different arguments. For example, for the finiteness of the topological components, we prefer to use Morse Theory, as Milnor did in [2].

- In the final part of this work, we provide some applications to show that this simple technique conceals a powerful tool for dealing with singularities.

1.2 Main Results

1.2.1 Fundamental aspects of algebraic geometry

Let $\mathbb{K}$ be a fixed field. We define the affine n-space over $\mathbb{K}$

$$\mathbb{A}^n_{\mathbb{K}} = \mathbb{A}^n := \mathbb{K}^n = \mathbb{K}\overbrace{\times \cdots \times}^{n \text{ times}}\mathbb{K}.$$

Let $A := \mathbb{K}[x_1, \ldots, x_n]$ be the polynomial ring in n variables over $\mathbb{K}$. The elements of A can be interpreted as functions $f : \mathbb{A}^n \to \mathbb{K}$ by substitution. Thus, if $f \in A$, we can talk about its set of zeros

$$V(f) := \{P \in \mathbb{A}^n \ : \ f(P) = 0\}.$$

More generally, if $F \subset A$, we can talk about its common zero set

$$V(F) := \{P \in \mathbb{A}^n \ : \ f(P) = 0, \ \ \forall f \in F\}.$$

Notice that if $F \subset G \subset A$, then $V(G) \subset V(F)$, i.e., V is inclusion-reversing.

Definition 1. *A subset $Y \subset \mathbb{A}^n$ is called an affine algebraic set if there exists $F \subset A$ such that $Y = V(F)$.*

Apparently, affine algebraic sets can be the common zero set of too many polynomials. The first thing we are going to prove is that there is always a suitable finite choice of the family F. In fact, this is a direct consequence of the Hilbert's basis theorem.

Corollary 1. *Every affine algebraic set is the common zero set of finitely many polynomials.*

Let $S \subset \mathbb{A}^n$. We define the ideal

$$I(S) := \{f \in A \ : \ S \subset V(f)\}.$$

Notice that I is inclusion-reversing.

In order to define the Zariski topology on the affine space, the following proposition is needed.

Proposition 1. *Affine algebraic sets are closed to finite union and arbitrary intersections. Additionally, the emptyset and the whole space are affine algebraic sets.*

Definition 2. *We define the Zariski topology on $\mathbb{A}^n$ by taking as open sets the complementary of affine algebraic sets. By the preceding proposition, this is a well-defined topology on $\mathbb{A}^n$, whose closed sets are exactly the affine algebraic sets.*

It is easy to see that, if $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, then Zariski open sets are open in the usual topology of the affine space.

Now we aim to prove that the affine space $\mathbb{A}^n$ with the Zariski topology is a Noetherian space, i.e., that for every descendent chain of algebraic sets $C_1 \supseteq C_2 \supseteq \cdots$, there is some $n_0 \in \mathbb{N}$ such that $C_n = C_{n_0}, \ \ \forall n \geq n_0$. This will follow from the following proposition.

Proposition 2. *Let $S \subset \mathbb{A}^n$. Then*

$$V(I(S)) = \bar{S}.$$

Proof. By definition of the Zariski topology, $V(I(S))$ is closed. Adding this to the trivial inclusion $S \subset V(I(S))$, we get $\bar{S} \subset V(I(S))$. Conversely, write $\bar{S} = V(J)$ for some ideal $J \subset A$. Therefore $S \subset V(J)$. Applying the inclusion-reversing property of I, we obtain $I(V(J)) \subset I(S)$. So, from the trivial inclusion $J \subset I(V(J))$, it follows that $J \subset I(S)$. Finally, using the inclusion-reversing property of V, we get the desired inclusion $V(I(S)) \subset V(J) = \bar{S}$. □

Theorem 1. *The affine space $\mathbb{A}^n$ with the Zariski topology is a Noetherian space.*

Proof. Indeed. Every descendant succession of algebraic sets

$$C_1 \supseteq C_2 \supseteq \cdots \supseteq C_n \supseteq \cdots$$

leads to an ascendant succession of ideals in A

$$I(C_1) \subseteq I(C_2) \subseteq \cdots \subseteq I(C_n) \subseteq \cdots$$

Since A is Noetherian, there is some $n_0 \in \mathbb{N}$ such that $I(C_n) = I(C_{n_0}), \quad \forall n \geq n_0$. Applying Proposition 2, it follows that

$$C_n = \overline{C_n} = V(I(C_n)) = V(I(C_{n_0})) = \overline{C_{n_0}} = C_{n_0}, \quad \forall n \geq n_0.$$

□

1.2.2 Whitney's theorem

In what follows $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. Our main objective in this work is to review Whitney's proof of the following,

Theorem 2 (Whitney's Theorem)**.** *Let $E \subset \mathbb{K}^n$ be an affine algebraic set. Then*

$$E = \bigsqcup_{j=1}^{l} M_j,$$

where M_j is a connected $\mathbb{K}$-analytic submanifold of $\mathbb{K}^n$ for every $j = 1, \ldots, l$.

Since the proof is very extense, we will divide it in several lemmas. But first, let us define the singular set of an affine algebraic set.

Let $E \subset \mathbb{K}^n$ be an affine algebraic set. For each $x \in E$, we define

$$\rho(x) := \dim_{\mathbb{K}} \left(\langle \{ df_x \ : \ f \in I(E) \} \rangle_{\mathbb{K}} \right).$$

First we observe that this dimension is always finite. Indeed. From the relation $d(fg)_x = f(x)dg_x + g(x)df_x$ we deduce that

$$d(fg)_x = g(x)df_x, \quad \forall f \in I(E), \ g \in A.$$

Suppose $I(E) = \langle f_1, \ldots, f_r \rangle_A$. Then

$$\rho(x) = \dim_{\mathbb{K}} \left(\langle d(f_1)_x, \ldots, d(f_r)_x \rangle_{\mathbb{K}} \right) \leq r < \infty, \quad \forall x \in E.$$

This implies that exists

$$\rho_0 := \max_{x \in E} \rho(x) \leq r < \infty.$$

The singular points of E are the points $x \in E$ such that $\rho(x) < \rho_0$. The set of singular points of E will be denoted by $\mathrm{Sing}E$. Notice that

$$\begin{aligned} x \in \mathrm{Sing}E &\longleftrightarrow \dim_{\mathbb{K}}(\langle d(f_1)_x, \ldots, d(f_r)_x \rangle_{\mathbb{K}}) < \rho_0 \\ &\longleftrightarrow \operatorname{rank} \left(\frac{\partial f_i}{\partial x_j}(x) \right)_{\substack{1 \leq i \leq r \\ 1 \leq j \leq n}} < \rho_0. \end{aligned} \tag{1.1}$$

In particular, equation (1.1) shows that $\mathrm{Sing}E$ is an affine algebraic set. The following lemma is probably the most important one. This is why we decided to include here its proof.

Lemma 1. *$E - \mathrm{Sing}E$ is a $\mathbb{K}$-analytic submanifold of $\mathbb{K}^n$ of dimension $n - \rho_0$.*

Proof. It suffices to show that each point $x_0 \in E - \mathrm{Sing}E$ has an open neighborhood (nbhd.) V in $\mathbb{K}^n$ such that $V \cap (E - \mathrm{Sing}E)$ is an analytic submanifold of $\mathbb{K}^n$. Let $x_0 \in E - \mathrm{Sing}E$. We can suppose without loss of generality that $x_0 = 0$ and that $d(f_1)_0, \ldots, d(f_{\rho_0})_0$ are $\mathbb{K}$-linearly independent. Thus, by (1.1), we have

$$\operatorname{rank}\left(\frac{\partial f_i}{\partial x_j}(0)\right)_{\substack{1\leq i\leq \rho_0\\ 1\leq j\leq n}} = \rho_0.$$

So some minor of order ρ_0 does not vanish. With a coordinates permutation, we can suppose that this minor is determined by the first ρ_0 columns, i.e.,

$$\frac{\partial(f_1, \ldots, f_r)}{\partial(x_1, \ldots, x_{\rho_0})}(0) \neq 0.$$

Consider the map germ

$$\begin{array}{rccc} \phi: & (\mathbb{K}^n, 0) & \longrightarrow & (\mathbb{K}^n, 0) \\ & x & \longmapsto & (f_1(x), \ldots, f_{\rho_0}(x), x_{\rho_0+1}, \ldots, x_n) \end{array}.$$

By construction, the jacobian of ϕ does not vanish at the origin. Apply the Inverse Function Mapping Theorem for analytic germs (notice that the coordinate functions are polynomials) and then choose a good representative of the germ to get an analytic diffeomorphism

$$\phi : V \longrightarrow I_\varepsilon^n,$$

where V is an open neighborhood of the origin of $\mathbb{K}^n$ and $I_\varepsilon^n :=]-\varepsilon/2, \varepsilon/2[^n$ for some $\varepsilon > 0$. Define now

$$V_0 := \{x \in V \ : \ \phi_i(x) = 0, \ \ \forall i = 1, \ldots, \rho_0\} = \phi^{-1}(\{0\} \times I_\varepsilon^{n-\rho_0}).$$

Since ϕ is an analytic diffeomorphism and $\{0\} \times I_\varepsilon^{n-\rho_0}$ is a $\mathbb{K}$-analytic submanifold of $\mathbb{K}^n$ of dimension $n - \rho_0$, we deduce that V_0 is a $\mathbb{K}$-analytic submanifold of $\mathbb{K}^n$ of dimension $n - \rho_0$. Now, from the fact that ϕ is a diffeomorphism, it follows that $d(f_1)_x, \ldots, d(f_{\rho_0})_x$ are $\mathbb{K}$-linearly independent for every $x \in V$. Consequently, $V \cap (E - \mathrm{Sing}E) = V \cap E$. So, it now suffices to prove that $V_0 = V \cap E$. The inclusion $V \cap E \subseteq V_0$ is obvious and, since $V_0 \subseteq V$ by definition, it only remains to show that $V_0 \subseteq E$.

For this purpose, let $\mathscr{C}^w_{(\mathbb{K}^n,0)}$ be the $\mathbb{K}$-algebra of analytic function germs of the form $h : (\mathbb{K}^n, 0) \longrightarrow \mathbb{K}$. We claim that if $F \in W := \mathscr{C}^w_{(\mathbb{K}^n,0)} I(E)$ (i.e., the ideal in $\mathscr{C}^w_{(\mathbb{K}^n,0)}$ generated by $I(E)$), then $\partial F/\partial x_j \in W$ for each $\rho_0 + 1 \leq j \leq n$. It is sufficient to show it for the case $F \in I(E)$. Fix $\rho_0 + 1 \leq j \leq n$ and set $f_l := x_l$ if $\rho_0 + 1 \leq j \leq n$ and $l \neq j$, $f_j := F$ and $f = (f_1, \ldots, f_n)$. Since $F(0) = 0$, there is an open nbhd. $U \subset V$ of the origin such that $F(U) \subset]-\varepsilon/2, \varepsilon/2[$. Thus

$$\phi^{-1} \circ f(x) = (x_1, \ldots, x_{\rho_0}, x_{\rho_0} + 1, \ldots, F(x), \ldots, x_n), \ \ \forall x \in U.$$

On the one hand we have

$$J(\phi^{-1} \circ f)(x) = \frac{\partial F}{\partial x_j}(x), \ \ \forall x \in U. \tag{1.2}$$

On the other one, by applying the Chain Rule we obtain

$$J(\phi^{-1} \circ f)(x) = (J(\phi^{-1}) \circ f)(x) Jf(x), \ \ \forall x \in U. \tag{1.3}$$

Since $J(\phi^{-1})\circ f$ is analytic in U, we have $J(\phi^{-1})\circ f \in \mathscr{C}^w_{(\mathbb{K}^n,0)}$. Moreover, since $f_1,\dots,f_{\rho_0},F \in I(E)$, it follows that

$$Jf(x) = \frac{\partial(f_1,\dots,f_{\rho_0},F)}{\partial(x_1,\dots,x_{\rho_0},x_j)}(x) = 0,\ \ \forall x \in E,$$

because otherwise we would have points in E with rank greater than ρ_0. In particular

$$\frac{\partial F}{\partial x_j} \overset{(1.2)}{=} J(\phi^{-1}\circ f) \overset{(1.3)}{=} (J(\phi^{-1})\circ f)Jf \in W.$$

Then, by induction one gets that

$$(\partial^{\alpha_{\rho_0+1}}/\partial x_{\rho_0+1}^{\alpha_{\rho_0+1}})\cdots(\partial^{\alpha_n}/\partial x_n^{\alpha_n})F \in W,\ \ \forall \alpha=(\alpha_{\rho_0+1},\dots,\alpha_n),\ \ \forall F\in W. \tag{1.4}$$

In particular

$$(\partial^{\alpha_{\rho_0+1}}/\partial x_{\rho_0+1}^{\alpha_{\rho_0+1}})\cdots(\partial^{\alpha_n}/\partial x_n^{\alpha_n})F(0) = 0,\ \ \forall \alpha=(\alpha_{\rho_0+1},\dots,\alpha_n),\ \ \forall F\in W.$$

Since $f_i \in I(E) \subset W$, for every $i=1,\dots,\rho_0$, it follows by analyticity and by (1.4) that $f_i(x) \equiv f_i(x_1,\dots,x_{\rho_0})$. Thus,

$$\phi(u;v) = (f_1(u),\dots,f_{\rho_0}(u);v),\ \ \forall (u;v)\in V,\ \ u\in\mathbb{K}^{\rho_0},\ v\in\mathbb{K}^{n-\rho_0}.$$

Consequently,

$$\phi^{-1}(x;y) = ((\phi^{-1})_1(x),\dots,(\phi^{-1})_{\rho_0}(x);y),\ \ \forall (x;y)\in I^n_\varepsilon,\ x\in\mathbb{K}^{\rho_0},\ y\in\mathbb{K}^{n-\rho_0}.$$

From here it follows that, for every $i=1,\dots,\rho_0$,

$$(\partial^{\alpha_{\rho_0+1}}/\partial y_{\rho_0+1}^{\alpha_{\rho_0+1}})\cdots(\partial^{\alpha_n}/\partial y_n^{\alpha_n})(\phi^{-1})_i(0) = 0,\ \ \forall \alpha=(\alpha_{\rho_0+1},\dots,\alpha_n). \tag{1.5}$$

We claim now that (1.4) and (1.5) imply that

$$(\partial^{\alpha_{\rho_0+1}}/\partial y_{\rho_0+1}^{\alpha_{\rho_0+1}})\cdots(\partial^{\alpha_n}/\partial y_n^{\alpha_n})(F\circ\phi^{-1})(0) = 0,\ \ \forall \alpha=(\alpha_{\rho_0+1},\dots,\alpha_n),\ \ \forall F\in W. \tag{1.6}$$

The argument is by induction over $|\alpha|$. If $|\alpha|=0$ it is trivial. We will do it for $|\alpha|=1,2$ to show the reader the idea. Let $\rho_0+1\le i,j\le n$. Then

$$\begin{aligned}
\frac{\partial(F\circ\phi^{-1})}{\partial y_i}(0) &= \sum_{k=1}^n \frac{\partial F}{\partial x_k}(\phi^{-1}(0))\frac{\partial(\phi^{-1})_k}{\partial y_i}(0)\\
&= \sum_{k=1}^n \frac{\partial F}{\partial x_k}(0)\frac{\partial(\phi^{-1})_k}{\partial y_i}(0)\\
&\overset{(1.4)}{=} \sum_{k=1}^{\rho_0} \frac{\partial F}{\partial x_k}(0)\frac{\partial(\phi^{-1})_k}{\partial y_i}(0)\\
&\overset{(1.5)}{=} 0.
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2(F\circ\phi^{-1})}{\partial y_j\partial y_i}(0) &= \sum_{k=1}^n\left(\sum_{l=1}^n \frac{\partial^2 F}{\partial x_l\partial x_k}(0)\frac{\partial(\phi^{-1})_l}{\partial y_i}(0)\frac{\partial(\phi^{-1})_k(0)}{\partial y_j}(0)\right)\\
&\quad + \frac{\partial F}{\partial x_k}(0)\frac{\partial^2(\phi^{-1})_k}{\partial y_i\partial y_j}(0)\\
&\overset{(1.5)}{=} \sum_{k=\rho_0+1}^n\left(\sum_{l=\rho_0+1}^n \frac{\partial^2 F}{\partial x_l\partial x_k}(0)\frac{\partial(\phi^{-1})_l}{\partial y_i}(0)\frac{\partial(\phi^{-1})_k(0)}{\partial y_j}(0)\right)\\
&\quad + \frac{\partial F}{\partial x_k}(0)\frac{\partial^2(\phi^{-1})_k}{\partial y_i\partial y_j}(0)\\
&\overset{(1.4)}{=} 0.
\end{aligned}$$

An induction argument proves (1.6). And thus, by analyticity, we have

$$F \circ \phi^{-1} \mid_{(\{0\} \times I_\varepsilon^{n-\rho_0}, 0)} = 0, \ \ \forall F \in W.$$

Since $\phi^{-1} : (\{0\} \times I_\varepsilon^{n-\rho_0}, 0) \longrightarrow (V_0, 0)$ is a diffeomorphism germ, the previous relation is equivalent to say that $F \mid_{(V_0,0)} = 0, \quad \forall F \in W$. In particular, if $F \in I(E) \subset W$, since F is analytic in the connected submanifold V_0, it follows that $F \mid_{V_0} = 0$, i.e., $F \in I(V_0)$. So $I(E) \subseteq I(V_0)$. To conclude, using that I is inclusion-reversing, we obtain the desired inclusion $V_0 \subseteq E$. □

For the finiteness of the topological components we first reduce the problem to the real case and then use Morse Theory arguments as Milnor did in [2]. This is just an easy lemma.

Lemma 2. *Let $F \subset \mathbb{C}^n$ be a complex affine algebraic set. Then there is a homeomorphism $\Phi : \mathbb{C}^n \longrightarrow \mathbb{R}^{2n}$ such that $\Phi(F) \subseteq \mathbb{R}^{2n}$ is a real affine algebraic set.*

Three more lemmas are needed.

Lemma 3. *Let $E \subset \mathbb{K}^n$ an affine algebraic set, $f_1, \ldots, f_r \in I(E)$, $x_0 \in E$ such that*

$$\det \left(\frac{\partial f_i}{\partial x_j}(x_0) \right)_{1 \leq i,j \leq n} \neq 0.$$

Then $E - \{x_0\}$ is an affine algebraic set.

Proof. Suppose without loss of generality that $x_0 = 0$. Since the polynomials $f_1, \ldots, f_n$ vanish at the origin, we can choose polynomials g_{ij} such that

$$f_i = x_1 g_{i1} + \cdots + x_n g_{in}, \ \ \forall 1 \leq i \leq n.$$

On the one hand

$$\frac{\partial f_i}{\partial x_j}(0) = g_{ij}(0) \longrightarrow \det(g_{ij}(0))_{1 \leq i,j \leq n} = \det \left(\frac{\partial f_i}{\partial x_j}(0) \right)_{1 \leq i,j \leq n} \neq 0.$$

And on the other one, the relation

$$\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} g_{11}(0) \\ \vdots \\ g_{1n}(0) \end{pmatrix} x_1 + \cdots + \begin{pmatrix} g_{1n}(0) \\ \vdots \\ g_{nn}(0) \end{pmatrix} x_n,$$

implies by Cramer's Rule that $\det(g_{ij}(x))_{1 \leq i,j \leq n} = 0, \ \ \forall x \in \mathbb{K}^n - \{0\}$. Therefore

$$E - \{x_0\} = E \cap (\mathbb{K}^n - \{x_0\}) = E \cap V(\det(g_{ij})_{1 \leq i,j \leq n}).$$

□

Lemma 4. *Let $E \subset \mathbb{K}^n$ be an affine algebraic set of isolated points. Then E is finite.*

Proof. Since the points are isolated, E is a submanifold of $\mathbb{K}^n$ of dimension 0. Suppose that $E = V(f_1, \ldots, f_r)$. Denote by $\rho_0 := \max_{x \in E} \rho(x)$. Then $\rho_0 = n$. Otherwise, applying the Lemma 1, we would obtain that $E - \mathrm{Sing}E$ is a submanifold of $\mathbb{K}^n$ of dimension $n - \rho_0 > 0$. This contradicts the fact that $E - \mathrm{Sing}E \subset E$ and E is a submanifold of $\mathbb{K}^n$ of dimension 0. So $\rho_0 = n$. This means that there is at least one point $x_0 \in E$ such that $\rho(x_0) = n$. This implies

that some minor of the matrix $(\partial f_i/\partial x_j(0))$ does not vanish. We can suppose without loss of generality that this minor is determined by the first n rows. This means that $f_1, \ldots, f_n \in I(E)$ verify that

$$\det\left(\frac{\partial f_i}{\partial x_j}(x_0)\right) \neq 0.$$

Applying the Lemma 4, we obtain that $E - \{x_0\}$ is an affine algebraic set. If $E - \{x_0\}$ is not vacuous, we could reiterate the arguments. Since the affine space is Noetherian with the Zariski topology, this process ends in a finite number of iterations, yelding to the conclusion. □

The following lemma will give us the finiteness of the topological components.

Lemma 5. *Let $E, F \subset \mathbb{K}^n$ be affine algebraic sets such that $E - F$ is non-singular. Then $E - F$ has finitely many topological components.*

Proof. Since we are studying a topological property, we can suppose $\mathbb{K} = \mathbb{R}$ by Lemma 2. Suppose that $F = V(g_1, \ldots, g_r)$. Then $F = V(h)$, with $h = g_1^2 + \cdots + g_r^2$. Therefore, $E - F$ is a $\mathbb{R}$-analytic submanifold of $\mathbb{R}^n$ wich is diffeomorphic to the graph of

$$\begin{array}{rccc} \varphi: & E-F & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & \varphi(x) = \dfrac{1}{h(x)} \end{array}.$$

The graph $\Gamma(\varphi)$ of φ is a smooth submanifold of $\mathbb{R}^n \times \mathbb{R}$. Moreover, $\Gamma(\varphi)$ is an affine algebraic set given by $(E \times \mathbb{R}) \cap V(h(x)t - 1)$. In [5] it is proven that, for almost every point $p \in \mathbb{R}^n \times \mathbb{R}$, the function

$$\begin{array}{rccc} r_p: & \Gamma(\varphi) & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & r_p(x) = \|x - p\|^2 \end{array},$$

is a Morse function. Let $p \in \mathbb{R}^n \times \mathbb{R}$ such that r_p is a Morse function. The critical points of this Morse function, $\mathrm{Sing} r_p$, is an affine algebraic set of isolated points. Applying the Lemma 4, we get that $\mathrm{Sing} r_p$ is finite. Now the topological components of $E - F$ are in bijective correspondence with the ones of $\Gamma(\varphi)$. The last ones are closed in $\mathbb{R}^n \times \mathbb{R}$, so they must determine a minimum of the function r_p, that has to be a point of $\mathrm{Sing} r_p$. This completes the proof. □

Finally, we are in conditions to prove Whitney's Theorem.

Proof of 2. Let $E \subset \mathbb{K}^n$ be an affine algebraic set. Applying the Lemma 1 we get that

$$E = (E - \mathrm{Sing}E) \bigsqcup \mathrm{Sing}E,$$

where $E - \mathrm{Sing}E$ is an analytic submanifold of $\mathbb{K}^n$. Define $E_0 := E$ and $E_{i+1} := \mathrm{Sing}E_i$ for all $i > 0$. By induction

$$E = \left(\bigsqcup_{i=0}^{N} E_i - E_{i+1}\right) \bigsqcup E_{N+1},$$

with $E_i - E_{i+1}$ an analytic submanifold of $\mathbb{K}^n$ and E_{N+1} an affine algebraic set contained in E_N for every $N \in \mathbb{N}$. We are constructing a descendent chain of Zariski closed sets

$$E = E_0 \supseteq E_1 \supseteq E_2 \supseteq \cdots$$

Using the Noetherian property of the affine space with the Zariski topology, we know that there is some natural $N \in \mathbb{N}$ such that $E_{N+1} = E_i, \ \forall i \geq N + 1$. From the definition of E_{N+2}, this implies that $E_j = \emptyset, \ \forall j \geq N + 1$. So we have that

$$E = \bigsqcup_{i=1}^{N} E_i - E_{i+1},$$

where $E_i - E_{i+1}$ is a $\mathbb{K}$-analytic submanifold of $\mathbb{K}^n$. To conclude, apply the Lemma 5 to each $E_i - E_{i+1}$, to obtain that each one has finitely many topological components. □

1.2.3 Some applications

In this final section, we give some applications of Theorem 2.

We begin with an application that can be found in Milnor's book [2]. This one is probably the most well-known one. Here we give the same proof.

Corollary 2. *Let $f : \mathbb{K}^n \longrightarrow \mathbb{K}$ be a polynomial function. Then the set of critical values of f is a finite set.*

Proof. The set of critical points of f is an affine algebraic set given by

$$\mathrm{Sing}f := \{x \in \mathbb{K}^n \ : \ df_x = 0\} = \{x \in \mathbb{K}^n \ : \ \frac{\partial f}{\partial x_i}(x) = 0, \ \ \forall 1 \leq i \leq n\}.$$

Applying the Theorem 2, we can find connected $\mathbb{K}$-analytic submanifolds of $\mathbb{K}^n$, say $\{M_i\}_{i=1}^l$, such that

$$\mathrm{Sing}f = \bigsqcup_{i=1}^{l} M_i.$$

The restriction $f\mid_{M_i}$ of f to each submanifold M_i is a $\mathbb{K}$-analytic function that verifies

$$d(f\mid_{M_i})_x = (df_x)\mid_{T_xM_i} = 0 \mid_{T_xM_i} = 0, \ \ \forall x \in M_i.$$

Since M_i is connected, this implies that $f\mid_{M_i} = y_i$ is a constant function. Therefore, the set of critical values of f is finite:

$$f(\mathrm{Sing}f) = f\left(\bigsqcup_{i=1}^{l} M_i\right) = \bigcup_{i=1}^{l} f(M_i) = \bigcup_{i=1}^{l} \{y_i\} = \{y_1, \ldots, y_l\}.$$

□

The following application shows a deep difference between real and complex affine algebraic sets.

Corollary 3. *Let $\emptyset \neq E \subset \mathbb{C}^n$ be a complex affine algebraic set. If E is compact, then E is a finite set.*

Proof. Apply Whitney's Theorem to get a finite partition $\{M_i\}_{i=1}^l$ of E onto smooth[1] submanifolds of $\mathbb{C}^n$. Since M_i is a smooth submanifold of $\mathbb{C}^n$, we have that the inclusion $i : M_i \hookrightarrow \mathbb{C}^n$ is a smooth mapping.

Let us denote for each $1 \leq j \leq n$, $\pi_j : \mathbb{C}^n \longrightarrow \mathbb{C}$ the projection onto the j-coordinate, i.e., $\pi_j(x) = x_j$. Consider the smooth application

$$\pi_j \circ i : M_i \longrightarrow \mathbb{C}.$$

By the Open Mapping Theorem, $(\pi_j \circ i)(M_i)$ is a point or an open subset of $\mathbb{C}$. If it was an open subset of $\mathbb{C}$, we would have that $(\pi_j \circ i)(M_i)$ is open and closed[2] in $\mathbb{C}$. By the connectedness of

[1] Notice that here in the complex case, smooth means holomorphic.

[2] Indeed. M_i is compact because it is a closed subset of the compact set E. Thus, its image under $\pi_j \circ i$ is compact and in particular a closed subset of $\mathbb{C}$

the complex plane this would imply that $(\pi_j \circ i)(E) = \mathbb{C}$. However, being E compact, we know that its image under $\pi_j \circ i$ is compact. This is a contradiction. This contradiction shows that

$$(\pi_j \circ i)(M_i) = \{y_j^{(i)}\}, \ \ \forall 1 \leq j \leq n, \ \ \forall 1 \leq i \leq l.$$

From this we deduce that

$$M_i = i(M_i) = \{y^{(i)}\}, \ \text{ where } y^{(i)} = (y_1^{(i)}, \ldots, y_n^{(i)}), \ \ \forall 1 \leq i \leq l.$$

And hence

$$E = \bigsqcup_{i=1}^{l} M_i = \bigsqcup_{i=1}^{l} \{y^{(i)}\} = \{y^{(1)}, \ldots, y^{(l)}\}.$$

□

There are many more applications of Theorem 2. Another one can be found in the excellent Milnor's book [2], where he uses this kind of arguments to define the link of an isolated point of an affine algebraic set.

As we said in the introduction, Theorem 2 is the simplest way of stratifying a singular space. Stratification is a simple but powerful technique for decomposing singular spaces into smaller but smooth pieces called strata. However, to make further progress in Singularity Theory, it became necessary to understand how these pieces fit together. This is where Whitney's Stratifications come into play.

Whitney's conditions were introduced to address this issue, and they provided techniques, such as controlled vector fields, that allowed us to extend results from manifolds to Whitney's stratified spaces. For an overview of Stratification Theory, check [4] and for learning the controlled vector fields technique we refer the reader to [3].

References

[1] H. Whitney, *Elementary structure of real algebraic varieties*, Annals of Mathematics, **66(3)** : 545–556, 1957.

[2] J. Milnor, *Singular points of complex hypersurfaces*, Annals of Mathematics Studies, **61**, 1968.

[3] J. Mather, *Notes on topological stability*, Bull. Amer. Math. Soc. (N.S.), **49(4)**, 475-506, 2012.

[4] D. Trotman.: In: J.L. Cisneros Molina, D.T. Lê, J. Seade (eds.), *Stratification Theory*, Springer, Cham, 243-273, 2020.

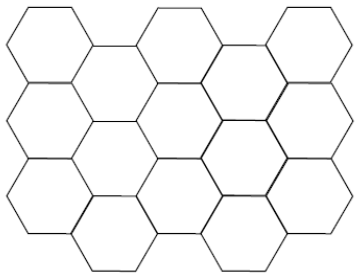
[5] J. Milnor, *Morse Theory*, Annals of Mathematics Studies, **51**, 1969.

Gersho's conjecture, Voronoi tessellations and applications

Clément Collin ♭, Marco Schipani ♮ and Martina Pascuzzo ♭

(♭) Universitat de València, clementcollin2@gmail.com
(♮) Università della Calabria, mschipani75@gmail.com
(♭) Università della Calabria, martinapascuzzo99@gmail.com

1.1 Introduction



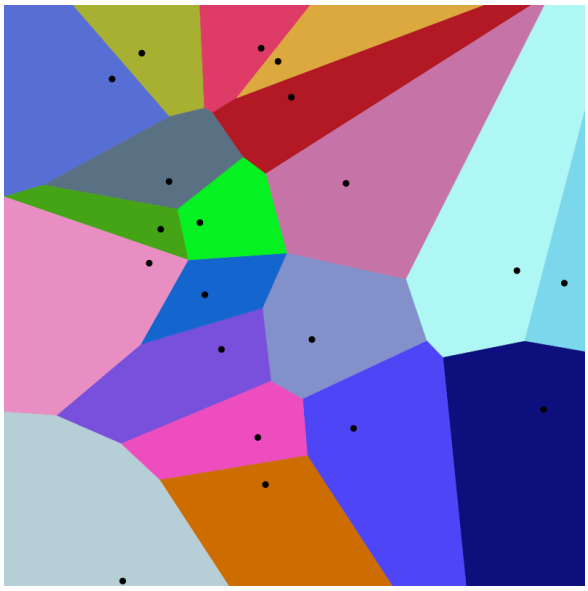
The origines of the honeycomb problem are somewhat obscure since it has very deep roots in history. The first fragments date back to $36B.C$, when Marcus Terentius Varro, in his book on agriculture, wrote about the hexagonal form of the bee's honeycomb. About two centuries later Pappus of Alexandria mentions this problem in his fifth book. He followed the much earlier approach of Zenodorus (ca 180 $B.C$). In this era there were basically two competing theories of the hexagonal structures. One theory held that the hexagon better accommodated the bee's six feet. The other theory, supported by the mathematitian of the day, was that the structure was explained by an isoperimetric property of the hexagonal honeycomb. Varro wrote, *does the cambers in the comb has six angles [...] the geometricians prove that this hexagon inscribed in a circoular figure encloses the greates amount of space.* Therefore, scholars of antiquity, for many centuries, retrace the well-known, and much earlier, Pythagorean isoperimetric problems. In fact it was known to Pythagorean that only three figures can tile the plane: the triangle, the square and the hexagon. Pappus basically uses this result in order to attribute a certain *geometric sense* to the bees for the choice of hexagons. However his reason for restricting these three figures are not strictly mathematical, in fact he also excludes gaps between the cells of the honeycomb since "foreign matter could enter in the interstices between then and so defile the purity of their produce". This easy-looking result will remain a mere "evidence in nature", therefore a conjecture, for many centuries. None turned it into a theorem until 1999, when T.Hales provided a complete proof of the 2D case. To date the conjecture is still open for higher dimensions. It is possible to reformulate the convex case of the problem in terms of Centroidal Voronoi tessellations (CVT's), historically known as Gersho's conjecture (1979). In 1999 Gruber

gave the first complete analytic proof of the Gersho's Conjecture in $2D$, basically extending the work of L. Fejes Tóth (1943). The first aim of this article is to highlights how the variational approach to the problem, via CVT's, was fundamental in reaching a solution. We detect it just following the Gruber's approach. Then we briefly discuss the 3D case, still open. We set some open question and propose some related recent approach. In the second part of the article, we will introduce the well-known Fortune's algorithm which is an incremental approach based on a sweepline technique to compute the Voronoi diagram of a bounded space given some point sites. After a quick introduction to algorithm complexity, we will show how to improve Fortune's algorithm's time complexity. Finally, we will discuss some optimization

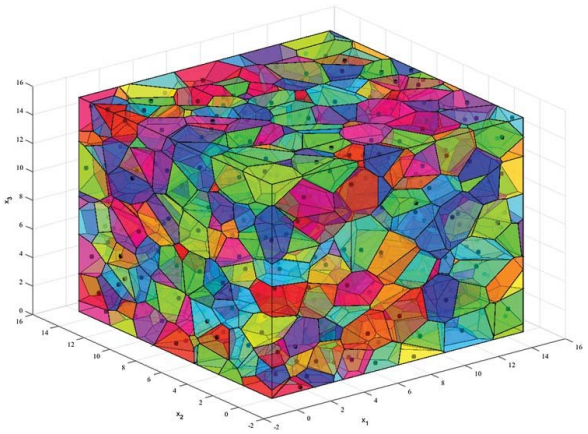
1.2 Voronoi Diagrams

Voronoi diagrams are among the most important structures in computational geometry. A Voronoi diagram records information about what is close to what. Let $P = \{p_1, p_2, \ldots, p_n\}$ be a set of n distinct points in the plane (or in any dimensional space), which we call *sites*. We define $V(p_i)$, the Voronoi cell for p_i, to be the set of points q in the plane that are closer (with respect to a defined distance d) to p_i than to any other site. That is, the Voronoi cell for p_i is defined to be:

$$V(p_i) = \{q | d(p_i, q) < d(p_j, q), \forall j \neq i\}.$$



(a) *Voronoi cells in 2D*



(b) *Voronoi cells in 3D*

Another way to define $V(p_i)$ is in terms of the intersection of halfplanes.

Definition 1.2.1. *Given two sites p_i and p_j in the plane, we define the bisector of p_i and p_j as the perpendicular bisector of the line segment $p_i\bar{p}_j$.*

$$B(p_i, p_j) = \{x | d(p_i, x) = d(p_j, x)\}$$

This bisector splits the plane into to half-planes.
Let's denote the open halfplane that contains p_i as $h(p_i, p_j)$ and the open half-plane that contains p_j as $h(p_j, p_i)$. Notice that $r \in h(p_i, p_j)$ if and only if $d(r, p_i) < d(r, p_j)$. From this, it is easy to see that a point q lies in $V(p_i)$ if and only if q lies within the intersection of $h(p_i, p_j)$ $\forall j \neq i$.

In other words, $V(p_i) = \bigcap_{j \neq i} h(p_i, p_j)$.

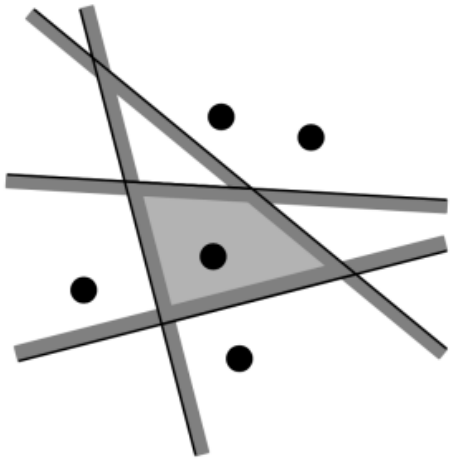


Figure 1.1: The Voronoi cell of the central point is the intersection of the halfplanes that contains it

By definition, each Voronoi cell $V(p)$ is the intersection of $n-1$ open halfplanes containing the site p. Since the intersection of halfplanes is a (possibly unbounded) convex polygon, it is easy to see that $V(p)$ is a (possibly unbounded) convex polygon. Notice that different Voronoi cells are disjoint. Finally, we define the Voronoi diagram of P, denoted $Vor(P)$.

Definition 1.2.2. *The common boundary of two Voronoi regions is called a Voronoi edge, if it contains more than one point.*

Definition 1.2.3. *Endpoints of Voronoi edges are called Voronoi vertices; they belong to the common boundary of three or more Voronoi cells.*

Definition 1.2.4. *Abusing the terminology slightly, we will use $Vor(P)E$ to indicate only the edges and vertices of the subdivision*

1.3 Gersho's conjecture

The Gersho's conjecture is basically a riformulation of the Honeycomb conjecture, under the hypotesis of convexity, in terms of centroidal Voronoi tassellations (CVT). A CVT is realized when the sited p_i are exactly the centroid of their associated Voronoi region V_k.

CVT's enjoy a variational characterization, based upon minimization of the following nonlocal energy

$$E(Y) := \int_Q dist^2(x, Y)dx \tag{1.1}$$

This is the second moment energy associated to the centroids $\{y_k\}, k = \{1, \cdots, n\}$

A well-known conjecture attributed to Gersho adresses the periodic nature of the configuration with least error, or equivalently the *CVT with lowest energy.* The following is the original version of the conjecture:

Conjecture 1.3.1. (Gersho's conjecture)

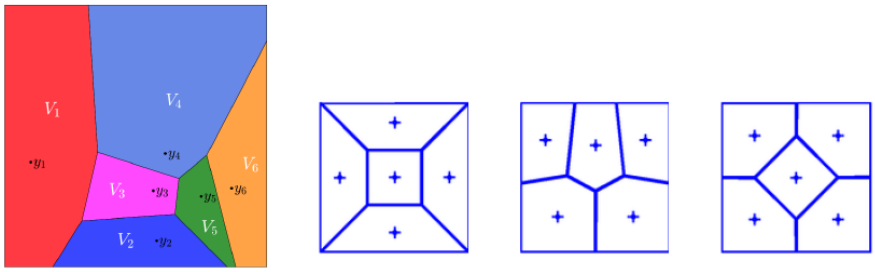


Figure 1.2: Left: A Voronoi diagram (the Voronoi regions associated with six generators). Right: Three centroidal Voronoi tassellations with five generators.

(a) *There exist a polytope V with $|V| = 1$ which tiles the space with congruent copies such that the following holds: let $(Y_n)_n$ be a sequence of minimizers, with Y_n minimizer with n points,then the Voronoi cells of points Y_n are asymptotically congruent to $n^{-1/N}V$ as $n \to +\infty$.*
(b) *For dimension $N = 2$, the optimal polytope V is a regular hexagon, corresponding to an optimal placemment of points on a triangluar lattice (cf. Figure 2 left). For dimension $N = 3$, the optimal polytope V is the truncated octaedron, corresponding corresponding an optimal placement of points on a BBC (body centered cubic) lattice (cf. Figure 2 right).*

In other words this conjecture says that asymptotically speaking, all cells of the optimal CVT, while forming a tessellation, are congruent to a basic cell which depends on the dimension.

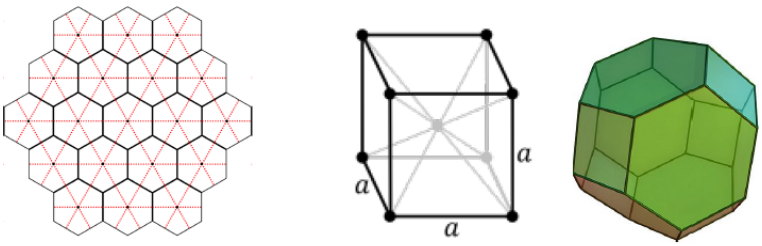


Figure 1.3: Left: $2D$ optimal placement of points on a triangular lattice with associated optimal Voronoi polytope a regular hexagon. Right: $3D$ conjectured optimal placement of points on a BBC lattice and the associated optimal Voronoi polytope the truncated octaedron.

Gruber presented an elementary proof in $2D$ of Gersho's conjecture. For convenience he took the domain Ω to be a suitably-chosen n-gon; however, one can work on the arbitrary domain at the expense of smaller-order boundary errors. This is the statemanet he proved.

Teorema 1.3.2. *Let $f : [0, +\infty) \to \mathbb{R}$ be a non-decreasing function and let H be a convex $3, 4, 5$ or 6-gon in $\mathbb{E}^2$, for any set P of n points in $\mathbb{E}^2$,*

$$S \coloneqq \int_H min\{f(\|x - p\|)\} : p \in P\}dx \geq n \int_{H_n} f(\|x\|)dx,$$

where H_n is the regular hexagon in $\mathbb{E}^2$ of area $|H|/n$ and center at the origin o.

The problem is widely addressed also in the three-dimensional case, where, as one might expect, it becomes much more difficult. The foundamental difficulty of applying Gruber's argoumments in $3D$ case is establishing the convexity in m of

$$G(a, m) \coloneqq \min_V \int_V |x - y|^2 dx, \qquad y = \text{centroid of } V$$

where V is a convex polytope having at most m faces and such that $|V| = a$.

We do not have regular $m-$hedron in $3D$, and computations are unfeasible. A priori, the maximum number of possible faces of the Voronoi polygons associated with the critical point can grow with n. Rustum an Choksi proved some upper bounds on the geometric complexity (including the number of faces) of such polygons which are independent of n. even if at this point one expects that these results guarantee a way to extend the Gersho's conjecture in $3D$, the two mathematicians reflect on the structural difference between the two cases, due to which the presence, in 3D, is not expected of a *universally optimal* configuration. This is in stark contrast with the $2D$ case, where the triangular lattice is almost surely to be universally optimal, although no rigorous proof is available. Gersho's conjecture would not be the first one in which such issue appears: it it is well known that the solution to the optimal foam problem in $2D$ is given the honeycomb structure, whose barycenters lie on the triangular lattice, while in $3D$ this is still open, and the long conjectured solution, i.e. the bitruncated cubic honeycomb, is surely not optimal, as it has higher energy than the Weaire-Phelan structure (see Figure 3).

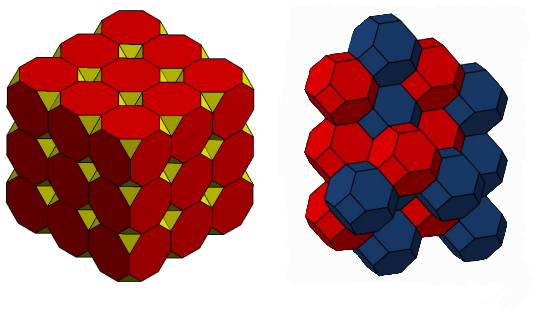


Figure 1.4: Left: The bitruncated honeycomb. Right: The Weaire-Phelan structure

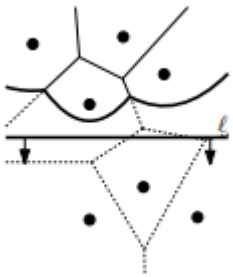
In conclusion, to date, the conjecture remains open in $3D$. Barnes and Sloan have proven the optimality of the BCC configuration amongst all lattice configurations, while Du and Wang have presented numerical evidence supporting the conjecture. The non-local and non-convex character of (1) insures a highly nontrivial energy landscape associated with a multitude of critical points with complex, albeit polygonal, Voronoi regions.

1.4 Fortune's algorithm

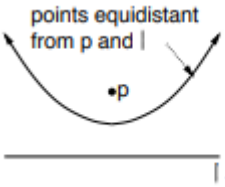
Fortune's algorithm represents a milestone in the field of computational geometry. The strategy in a plane sweep algorithm is to sweep a horizontal line—the *sweep line* —from top to bottom over the plane. While the sweep is performed, information is maintained regarding the structure that one wants to compute. More precisely, information is maintained about the intersection of the structure with the sweep line. While the sweep line moves downwards the information does not change, except at certain special points—the *event points*, which are any event that changes the topological structure of the Voronoi diagram and the beach line.

Let's try to apply this general strategy to the computation of the Voronoi diagram of a set $P = p_1, p_2, \ldots, p_n$ of point sites in the plane. According to the plan sweep paradigm we move a horizontal sweep line l from top to bottom over the plane. The paradigm involves maintaining the intersection of the Voronoi diagram with the sweep line. Unfortunately this is not so easy, because the part of $Vor(P)$ above l depends not only on the sites that lie above l but also on sites below l.

Stated differently, when the sweep line reaches the topmost vertex of the Voronoi cell $V(p_i)$ it has not yet encountered the corresponding site p_i. Hence, we do not have all the information needed to compute the vertex. We are forced to apply the plane sweep paradigm in a slightly different fashion: instead of maintaining the intersection of the Voronoi diagram with the sweep line, we maintain information about the part of the Voronoi diagram of the sites above l that cannot be changed by sites below l.



Denote the closed half-plane above l by l^+. What is the part of the Voronoi diagram above l that cannot be changed anymore? In other words, for which points $q \in l^+$ do we know for sure what their nearest site is? The distance of point $q \in l^+$ to any site below l is greater than the distance of q to l itself. Hence, the nearest site of q cannot lie below l if q is at least as near to some site $p_i \in l^+$ as q is to l^+. The locus of points that are closer to some site $p_i \in l^+$ than to l is bounded by a parabola. Hence, the locus of points that are closer to any site above l than to l itself is bounded by parabolic arcs.



We call this sequence of parabolic arcs the **beach line**. It is a dynamic data structure that plays a crucial role in handling events and updating the Voronoi diagram as the sweep line progresses. Notice that the portion of the Voronoi diagram that lies above the beach line is "safe" in the sense that we have all the information that we need in order to compute it (without knowing about which sites are still to appear below the sweep line).

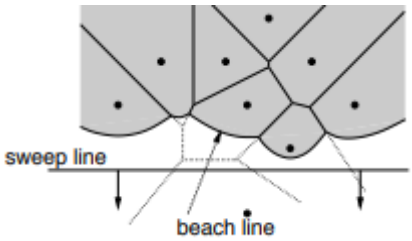


Figure 1.5: Only the portion of the Voronoi diagram that lies above the beach line is computed

Another way to visualize the beach line is the following. Every site p_i above the sweep line defines a complete parabola β_i. The beach line is the function that, for each x-coordinate, passes

through the lowest point of all parabolas.

Figure 1.6: The beachline

Observation 1.4.1. The beach line is x-monotone, that is, every vertical line intersects it in exactly one point.

It is easy to see that one parabola can contribute more than once to the beach line. We'll worry later about how many pieces there can be. This behavior is a key feature that allows the algorithm to maintain an accurate representation of the evolving Voronoi diagram. This dynamic reintroduction of parabolic arcs ensures that the algorithm can handle complex scenarios, including the creation and removal of Voronoi vertices, as the sweep line progresses.

Notice that the breakpoints, the points along the beachline where adjacent parabolic arcs intersect, equidistant from two sites and the sweep line, lie on edges of the Voronoi diagram and may represent potential Voronoi vertices. This is not a coincidence: the breakpoints exactly trace out the Voronoi diagram while the sweep line moves from top to bottom. These properties of the beach line can be proved using elementary geometric arguments.From this we have the following important characterization.

Lemma 1.4.2. *The breakpoints of the beach line lie on Voronoi edges of the final diagram.*

So, instead of maintaining the intersection of Vor(P) with l we maintain the beach line as we move our sweep line l. We do not maintain the beach line explicitly, since it changes continuously as l moves. This happens when a new parabolic arc appears on it, and when a parabolic arc shrinks to a point and disappears.

First we consider the events where a new arc appears on the beach line. One occasion where this happens is when the sweep line l reaches a new site. The parabola defined by this site is at first a degenerate parabola with zero width: a vertical line segment connecting the new site to the beach line. As the sweep line continues to move downward the new parabola gets wider and wider. The part of the new parabola below the old beach line is now a part of the new beachline. Figure 1.7 illustrates this process. We call the event where a new site is encountered a **site event**.

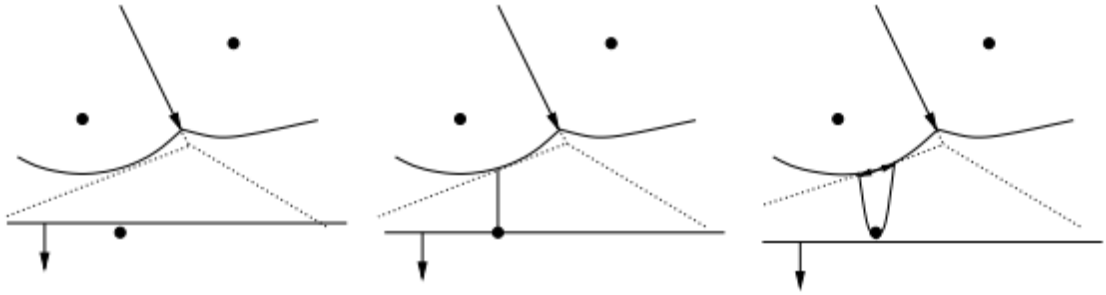


Figure 1.7: A new arc appears on the beach line because a site is encountered

What happens to the Voronoi diagram at a site event? Recall that the breakpoints on the beach line trace out the edges of the Voronoi diagram. At a site event two new breakpoints appear, which start tracing out edges. In fact, the new breakpoints coincide at first, and then move in opposite directions to trace out the same edge. Initially, this edge is not connected to the rest of the Voronoi diagram above the sweep line. Later on, the growing edge will run into another edge, and it becomes connected to the rest of the diagram. So now we understand what happens at a site event: a new arc appears on the beach line, and a new edge of the Voronoi diagram starts to be traced out. Is it possible that a new arc appears on the beach line in any other way? The answer is no :

Lemma 1.4.3. *The only way in which a new arc can appear on the beach line is through a site event.*

An immediate consequence of the lemma is that the beach line consists of at most $2n - 1$ parabolic arcs: each site encountered gives rise to one new arc and the splitting of at most one existing arc into two, and there is no other way an arc can appear on the beach line. The nice thing about site events is that they are all known in advance. Thus, after sorting the points by y-coordinate, all these events are known.

The second type of event in the plane sweep algorithm is where an existing arc of the beach line shrinks to a point and disappears. Let $\alpha^{'}$ be the disappearing arc, and let α and $\alpha^{''}$ be the two neighboring arcs of $\alpha^{'}$ before it disappears, as in the figure 1.8

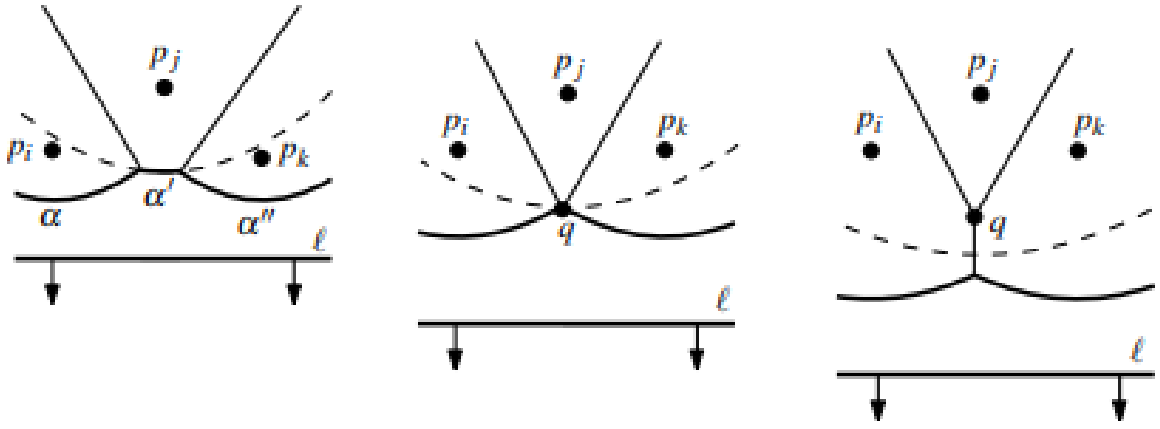


Figure 1.8: An arc disappears from the beach line

Let $\alpha^{'}$ be the disappearing arc, and let α and $\alpha^{''}$ be the two neighboring arcs of $\alpha^{'}$ before

it disappears. The arcs $\alpha^{'}$ and $\alpha^{''}$ cannot be part of the same parabola; this possibility can be excluded in the same way as the first possibility in the proof of Lemma (1.4.3) was excluded. Hence, the three arcs α, $\alpha^{'}$ and $\alpha^{''}$ are defined by three distinct sites p_i, p_j and p_k. At the moment $\alpha^{'}$ disappears, the parabolas defined by these three sites pass through a common point q. Point q is equidistant from l and each of the three sites. Hence, there is a circle passing through p_i,p_j and p_k with q as its center and whose lowest point lies on l.

There cannot be a site in the interior of this circle: such a site would be closer to q than q is to l, contradicting the fact that q is on the beach line. It follows that the point q is a vertex of the Voronoi diagram.

This is not very surprising, since we observed earlier that the breakpoints on the beach line trace out the Voronoi diagram. So when an arc disappears from the beach line and two breakpoints meet, two edges of the Voronoi diagram meet as well. We call the event where the sweep line reaches the lowest point of a circle through three sites defining consecutive arcs on the beach line a **circle event**. From the above we can conclude the following lemma.

Lemma 1.4.4. *The only way in which an existing arc can disappear from the beach line is through a circle event.*

Now we know where and how the combinatorial structure of the beach line changes: at a site event a new arc appears, and at a circle event an existing arc drops out. We also know how this relates to the Voronoi diagram under construction: at a site event a new edge starts to grow, and at a circle event two growing edges meet to form a vertex.

Our goal is to compute the Voronoi diagram, so we need a data structure that stores the part of the Voronoi diagram computed thus far. We also need the two 'standard' data structures for any sweep line algorithm: an **event queue**, a data structure (often implemented as a priority queue) that organizes events based on their x-coordinates, and a structure that represents the status of the sweep line (and in our case the beachline). These data structures are implemented in the following way.

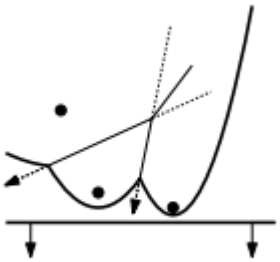
- We store the Voronoi diagram under construction in our usual data structure for subdivisions, the doubly-connected edge list. A Voronoi diagram, however, is not a true subdivision as defined before: it has edges that are half-lines or full lines, and these cannot be represented in a doubly-connected edge list. During the construction this is not a problem, because the representation of the beach line will make it possible to access the relevant parts of the doubly-connected edge list efficiently during its construction. But after the computation is finished we want to have a valid doubly-connected edge list. To this end we add a big bounding box to our scene, which is large enough so that it contains all vertices of the Voronoi diagram. The final subdivision we compute will then be the bounding box plus the part of the Voronoi diagram inside it.

- The event queue Q is implemented as a priority queue, where the priority of an event is its y-coordinate. It stores the upcoming events that are already known. For a site event we simply store the site itself. For a circle event the event point that we store is the lowest point of the circle, with a pointer to the leaf in T that represents the arc that will disappear in the event.

■ The beachline is represented by a balanced binary search tree T. The choice of such data structure will be discussed later.

All the site events are known in advance, but the circle events are not. This brings us to one final issue that we must discuss, namely the detection of circle events.

During the sweep the beach line changes its topological structure at every event. This may cause new triples of consecutive arcs to appear on the beachline and it may cause existing triples to disappear. Our algorithm will make sure that for every three consecutive arcs on the beach line that define a potential circle event, the potential event is stored in the event queue Q.



There are two subtleties involved in this.

- First of all, there can be consecutive triples whose two breakpoints do not converge, that is, the directions in which they move are such that they will not meet in the future; this happens when the breakpoints move along two bisectors away from the intersection point. In this case the triple does not define a potential circle event.

- Secondly, even if a triple has converging breakpoints, the corresponding circle event need not take place: it can happen that the triple disappears (for instance due to the appearance of a new site on the beach line) before the event has taken place. In this case we call the event a **false alarm**.

So what the algorithm does is this. At every event, it checks all the new triples of consecutive arcs that appear. For instance, at a site event we can get three new triples: one where the new arc is the left arc of the triple, one where it is the middle arc, and one where it is the right arc. When such a new triple has converging breakpoints, the event is inserted into the event queue Q.

Observe that in the case of a site event, the triple with the new arc being the middle one can never cause a circle event, because the left and right arc of the triple come from the same parabola and therefore the breakpoints must diverge. Furthermore, for all disappearing triples it is checked whether they have a corresponding event in Q. If so, the event is apparently a false alarm, and it is deleted from Q. This can easily be done using the pointers we have from the leaves in T to the corresponding circle events in Q.

Lemma 1.4.5. *Every Voronoi vertex is detected by means of a circle event*

1.5 Solving optimization problems through Voronoi diagrams

General setting. We consider n points in a bounded space $\bar{S}$.

$$P = \{p_i \mid i \leq n, p_i \in \bar{S}\}$$

The points are meant to represent facilities and the bounding space some region containing the facilities. Under this setting, we can consider the Voronoi diagram of $\bar{S}$ generated by the points of P and some distance (the Euclidean distance will be denoted d_E). Let's take a look at two locational optimization problems in $\bar{S}$ which can easily be solved once we have computed its Voronoi diagram.

1.5.1 Largest empty circle problem

This problem consists in finding the location from which the distance to the nearest facility is the longest in a bounded space $\bar{S}$, i.e. find

$$\max_{p \in \bar{S} / \bigcup_{i \in I} s_i} \min_{i \in I} \{d_E(p, s_i)\} \tag{1.2}$$

The circle centered at the worst location with radius r given by expression 1.2 is the largest circle in which there are no facilities. Thus, the problem of finding the worst location comes down to finding the largest empty circle.

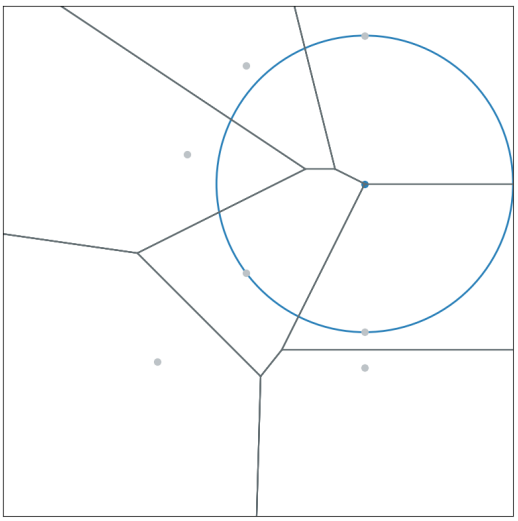


Figure 1.9: The Voronoi diagram of some set of points (in grey) and the largest empty circle. We can see that its center lies on a vertex of the Voronoi diagram

Lemma 1.5.1. *If V_i and V_j are two cells of a Voronoi diagram whose sites are respectively s_i and s_j and $p, p' \in V_i$, then $d(p, s_i) \leq d(p', s_j)$.*

Proof. From the definition of the Voronoi diagram, we know that $d(p, s_i) \leq d(p, s_j) \overset{\text{triangular inequality}}{\leq} d(p, p') + d(p', s_j) \leq d(p', s_j)$. □

Lemma 1.5.2. *In any cell V_i of a Voronoi diagram, the farthest point q_i^* from the site s_i exists on one of its vertices.*

Proof. Let V be a Voronoi diagram generated by $s_1, ..., s_n$. Let V_i be the cell whose site is s_i. Let $p_i \in V_i$, $p_{ij} \in V_i, V_j$ and $p_{ijk} \in V_i, V_j, V_k$. In other words, p_{ij} is on the edge between V_i and V_j and p_{ijk} is on the vertex at the junction of V_i, V_j and V_k.

From lemma 1.5.1 we know that, $d_E(p_i, s_i) \leq d_E(p_{ij}, s_j)$. Therefore, $d_E(p_{ij}, s_i) = d_E(p_{ij}, s_j) \implies d_E(p_i, s_i) \leq d_E(p_{ij}, s_i)$. What's more, $d_E(p_{ij}, s_i) \leq d_E(p_{ijk}, s_k)$ and $d_E(p_{ijk}, s_i) = d_E(p_{ijk}, s_k) \implies d_E(p_{ij}, s_i) \leq d_E(p_{ijk}, s_i)$. Everything put together, we have that the farthest point from s_i in V_i is on one of its vertices.

□

From lemma 1.5.2, we deduce that once we have the Voronoi diagram of $\bar{S}$, the solution of the largest circle problem is readily obtained by finding the maximum value among $\{d_E(q_{ij}, p_i) \mid i = 1, \ldots, n \ ; \ j = 1, \ldots, m_i\}$ where q_{ij} is the j-th vertex of the i-th cell.

If the Voronoi diagram has been computed with Fortune's algorithm, all these values (which correspond to the radius of the circles at each Circle Event) have already been calculated. Therefore, a few minor adjustments that don't increase the algorithm's complexity enable it to also solve the largest empty circle problem.

let Q be the ordered list of upcoming events
let B the beachline
let V the Voronoi diagram under construction
$r^* \leftarrow 0$
while Q *is not empty* **do**
  $E \leftarrow$ topmost event in Q
  if E *is a site event* **then**
    update B consequently
    look for new circle events
  end
  if E *is a circle event* **then**
    insert the corresponding vertex into V
    remove the shrunk arc A from B
    remove upcoming events involving A from Q
    look for new circle events
    if $r_C > r^*$ **then**
      $r^* \leftarrow r_C$
    end
  end
end

Algorithm 1: Description of Fortune's algorithm modified to also solve the Largest Circle Problem

1.5.2 Smallest enclosing circle problem

It consists in finding the location from which the distance to the farthest facility is the shortest, i.e. find

$$\min_{p \in \bar{S} / \bigcup_{i \in I} s_i} \max_{i \in I} \{d_E(p, s_i)\}$$

This problem can be solved through the *farthest-point Voronoi diagram* which is one obtained by using not the Euclidean distance but its opposite ($d = -d_E$). The *farthest-point Voronoi diagram* partitions the plane in convex regions in each of which the farthest site is the same.

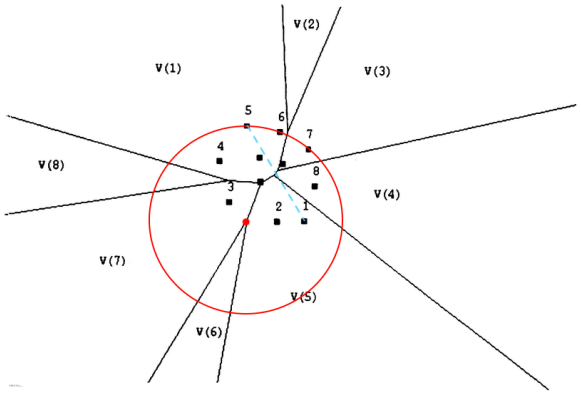


Figure 1.10: A farthest-point Voronoi diagram, its diameter (the blue dashed line) and an enclosing circle (in red). Region $V(i)$ contains all the points whose most distant site is number i

This diagram has the property that for every vertex q_i there exists a unique circle C_i centered at q_i which passes through three or more sites and encloses all other sites. This circle is called an enclosing circle.

The center of the smallest enclosing circle may exist on the middle of the diameter (the segment whose ends are the two sites furthest apart) or may be on a vertex of the *farthest-point Voronoi diagram*. Thus, the procedure for solving this problem consists of two steps. First, we examine if the circle whose diameter is the diameter of the diagram is an enclosing circle. If so, its center is solution to the problem. If not, the solution can be found among the enclosing circles C_i centered at q_i. Once again, if the diagram has been computed with Fortune's algorithm, these circles have already been processed.

References

[1] Franz Aurenhammer. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3):345–405, 1991.

[2] Rustum Choksi and Xin Yang Lu. Bounds on the geometric complexity of optimal centroidal voronoi tesselations in 3d, 2019.

[3] Mark de Berg, Marc van Kreveld, Mark Overmars, and Otfried Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, second edition, 2000.

[4] Qiang Du and Desheng Wang. The optimal centroidal voronoi tessellations and the gersho's conjecture in the three-dimensional space. *Computers Mathematics with Applications*, 49(9):1355–1373, 2005.

[5] A. Gersho. Asymptotically optimal block quantization. *IEEE Transactions on Information Theory*, 25(4):373–380, 1979.

[6] Peter M. Gruber. A short analytic proof of fejes tóth's theorem on sums of moments. *aequationes mathematicae*, 58:291–295, 1999.

[7] T. C. Hales. The honeycomb conjecture. *Discrete Comput. Geom.*, 25(1):1–22, jan 2001.

[8] Atsuyuki Okabe and Atsuo Suzuki. Locational optimization problems solved through voronoi diagrams. *European Journal of Operational Research*, 98(3):445–456, 1997.

[9] Fortune Steven. Sweepline algorithm for voronoi diagrams. *Springer-Algorithmica*, 2:153–174, 1987.

The Riemann hypothesis through Dirichlet polynomials

Mario Guillén[♭], Miguel Rodríguez[♮] and Marc Ventura[†]

(♭) mguisan3@posgrado.upv.es, Universitat Politècnica de València
(♮) miroa4@alumni.uv.es, Universitat de València
(†) marcven2@alumni.uv.es, Universitat de València

1.1 Introduction

In the 19th century, while studying the distribution of prime numbers, Bernhard Riemann revolutionized number theory by introducing the Riemann zeta function, establishing a deep connection between prime numbers and complex analysis. In his 1859 memoir [13], he proposed what is now known as the Riemann hypothesis: the conjecture that all non-trivial zeros of the zeta function lie on the critical line $\mathrm{Re}\,(s) = 1/2$. This hypothesis remains one of the most important unsolved problems in mathematics and is listed among the seven Millennium Prize Problems, with the Clay Mathematics Institute offering a one-million-dollar reward for a correct proof.

An indirect approach to understanding the hypothesis involves bounding the number of zeros that may lie off the critical line. These so-called zero-density estimates have significant implications for the distribution of prime numbers and are closely linked to the behavior of Dirichlet polynomials, which play a key role in detecting zeros of the zeta function through their large-value behavior.

This article begins with an exposition of the Riemann zeta function and its functional properties, laying the groundwork for the Riemann hypothesis. We then explore the relationship between zero-density estimates and Dirichlet polynomials, highlighting classical contributions such as those of Ingham. The final part is devoted to the recent breakthrough by Larry Guth and James Maynard (2024), who, through tools from harmonic analysis, established improved bounds that represent a significant step forward in this line of research.

1.1.1 Preliminaries

Definition 1.1. The Riemann zeta function is defined for all $s \in \mathbb{C}$ with $\mathrm{Re}(s) = \sigma > 1$ as

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

The function ζ defines a holomorphic function in $\mathrm{Re}\,(s) > 1$.

Theorem 1.2. *Let $s \in \mathbb{C}$ with $\mathrm{Re}(s) > 1$. Then*

$$\zeta(s) = \prod_{p \text{ prime}} \frac{1}{1 - 1/p^s}.$$

In particular, $\zeta(s)$ *has no zeros in* $\mathrm{Re}\,(s) > 1$.

To extend this function to a larger domain, it is useful to introduce the gamma function.

Definition 1.3. Let $s \in \mathbb{C}$ with $\mathrm{Re}(s) > 0$. We define

$$\Gamma(s) = \int_0^\infty t^{s-1} e^{-t} \mathrm{d}t.$$

The gamma function can be extended to a holomorphic function in $\mathbb{C} \setminus \{0, -1, -2, -3, \ldots\}$ that has no zeros and where $0, -1, -2, \ldots$ are simple poles.

A proof of the result above can be found in [9, Chapter 6, Theorem 1.4].

Definition 1.4. Let $s \in \mathbb{C}$ with $\mathrm{Re}(s) > 1$. We define

$$\xi(s) = \Gamma(s/2)\zeta(s)\pi^{-s/2}.$$

ξ can be extended to a holomorphic function in $\mathbb{C} \setminus \{0, 1\}$ and where 0 and 1 are simple poles with residues -1 and 1, respectively. Moreover, $\xi(s) = \xi(1 - s)$ for all $s \neq 0, 1$.

Corollary 1.5. *The Riemann zeta function can be extended to a holomorphic function* $\mathbb{C} \setminus \{1\}$ *where* 1 *is a simple pole.*

$$\zeta(s) = \pi^{s/2}\xi(s)\frac{1}{\Gamma(s/2)}.$$

Definition 1.6. ζ has simple zeros in $s = -2, -4, -6, \ldots$. These complex numbers are called the **trivial zeros** of ζ.

The case $s = 0$ is different because ξ has a pole in 0. Taking into account that $\Gamma(1) = 1$ and $\mathrm{Res}(\xi, 1) = -1$, we can calculate $\zeta(0) = -1/2$.

As a consequence of $\xi(s) = \xi(1 - s)$ we obtain the following equality.

Corollary 1.7 (Functional equation)**.** *For all* $s \neq 0, 1$

$$\Gamma(s/2)\zeta(s)\pi^{-s/2} = \Gamma\left(\frac{1-s}{2}\right)\zeta(1-s)\pi^{(s-1)/2},$$

regarding the evaluation as a limit when $s = -2, -4, \ldots,$ *or* $s = 1, 3, 5 \ldots$

Corollary 1.8. *The set of non-trivial zeros of* ζ *is symmetric respect to the line* $\mathrm{Re}(s) = 1/2$.

The case of $s = -2, -4, \ldots$ is different because Γ has a simple pole in $s/2$ and the RHS of the functional equation does not vanish.

From Theorem 1.2 and 1.8, we get that if $\mathrm{Re}(s) < 0$ then $\zeta(s) = 0$ only if s is a trivial zero. We conclude that the only non-trivial zeros of ζ must be in the strip $0 \leq \mathrm{Re}(s) \leq 1$. Moreover we have this result

Theorem 1.9. ζ *does not have zeros on the lines* $\mathrm{Re}(s) = 0$ *and* $\mathrm{Re}(s) = 1$.

We are now in a position to state the Riemann hypothesis

Conjecture 1.10 (Riemann hypothesis)**.** All non-trivial zeros of ζ are in the line $\mathrm{Re}(s) = 1/2$.

Remark 1.11. The set of non trivial zeros of ζ is also symmetric respect to the real axis. This is a consequence of $\zeta(s) = \overline{\zeta(\overline{s})}$, which can be proved directly when $\mathrm{Re}(s) > 1$ and then extended by the Analytic Continuation Principle.

Definition 1.12. Let $N(\sigma, T)$ be the number of zeros of the Riemann zeta function $\zeta(s)$ in the rectangle $\mathrm{Re}(s) \geq \sigma, 0 \leq \mathrm{Im}(s) \leq T$.

Remark 1.13. Note that by Theorem 1.2, if $\sigma < 1$, $N(\sigma, T)$ is the number of zeros of ζ in the rectangle $\sigma \leq \mathrm{Re}(s) \leq 1, 0 \leq \mathrm{Im}(s) \leq T$, which is a compact set. Since a holomorphic function non identically zero cannot have infinite zeros in a compact set, $N(\sigma, T)$ is a natural number.

Now, by Corollary 1.8 and Remark 1.11, we can write the Riemann hypothesis as follows:

Conjecture 1.14 (Riemann hypothesis)**.** $N(\sigma, T) = 0$ for all $\sigma > 1/2$ and $T > 0$.

In relation to this notation we have the following result stated by Riemann [13] and proved by von Mangoldt. A proof can be found in [4, Theorem 9.4].

Theorem 1.15 (Riemann-von Mangoldt Formula)**.** *Let $T > 0$ and let $N(T)$ denote the number of zeros of the function $\zeta(s)$ in the rectangle $0 \leq \mathrm{Re}(s) \leq 1$, $0 \leq \mathrm{Im}(s) \leq T$. Then*

$$N(T) = \frac{T}{2\pi} \log \frac{T}{2\pi} - \frac{T}{2\pi} + O(\log T).$$

Theorem 1.2 gave us a relationship between ζ and prime numbers. We now present a famous theorem whose proof involves the zeta function.

We will first introduce several notations. $A \lesssim B$ means that there exists a constant C verifying $A \leq CB$. $A \lesssim_z B$ means that for every z there exists a constant $C(z)$ depending on z so that $A \leq C(z)B$. $A \asymp B$ means that $A \lesssim B$ and $B \lesssim A$ both hold. $A \lessapprox B$ means that for every $\varepsilon > 0$ there exists a constant $C(\varepsilon)$ depending on ε verifying $A \leq C(\varepsilon) T^{\epsilon} B$ for all large T. $f(x) \sim g(x)$ means that $\lim_{x\to\infty} f(x)/g(x) = 1$.

Theorem 1.16 (Prime Number Theorem)**.** *Let $\pi(x)$ denote the number of primes not exceeding x. Then*

$$\pi(x) \sim \frac{x}{\log(x)} \quad \text{as } x \to \infty.$$

This theorem was originally proved by Hadamard [8] and de la Vallée Poussin [10] in 1896 using properties of the zeta function. Later, in 1949 Selberg [12] and Erdos [11] an elementary proof of this theorem. In fact, the title of Riemann's original article [13] translated into English is "On the Number of Prime Numbers less than a Given Quantity".

A consequence of prime number theorem is

$$\pi(2x) - \pi(x) \sim \frac{x}{\log(x)} \quad \text{as } x \to \infty.$$

So, we can ask if

$$\pi(x + x^{\theta}) - \pi(x) \sim \frac{x^{\theta}}{\log(x)} \quad \text{as } x \to \infty, \tag{1.1}$$

for some $\theta < 1$. Ingham proved in 1937 that the values of θ satisfying (1.1) are related to some bounds of $N(\sigma, T)$.

Theorem 1.17. *If $N(\sigma, T) \lesssim T^{A(1-\sigma)} \log^B T$ for all $1/2 \leq \sigma \leq 1$ then (1.1) is satisfied by*

$$1 - \frac{1}{A} < \theta < 1.$$

1.1.2 State of the art

Experience has shown that the following quantity is the most important to control when bounding the number of zeros of ζ (for example, see Theorem 1.17).

Definition 1.18. For $1/2 \leq \sigma \leq 1$, we define $A(\sigma)$ as the least non-negative constant for which one has an asymptotic

$$N(\sigma, T) \lesssim T^{A(\sigma)(1-\sigma)+o(1)}$$

as $T \to \infty$.

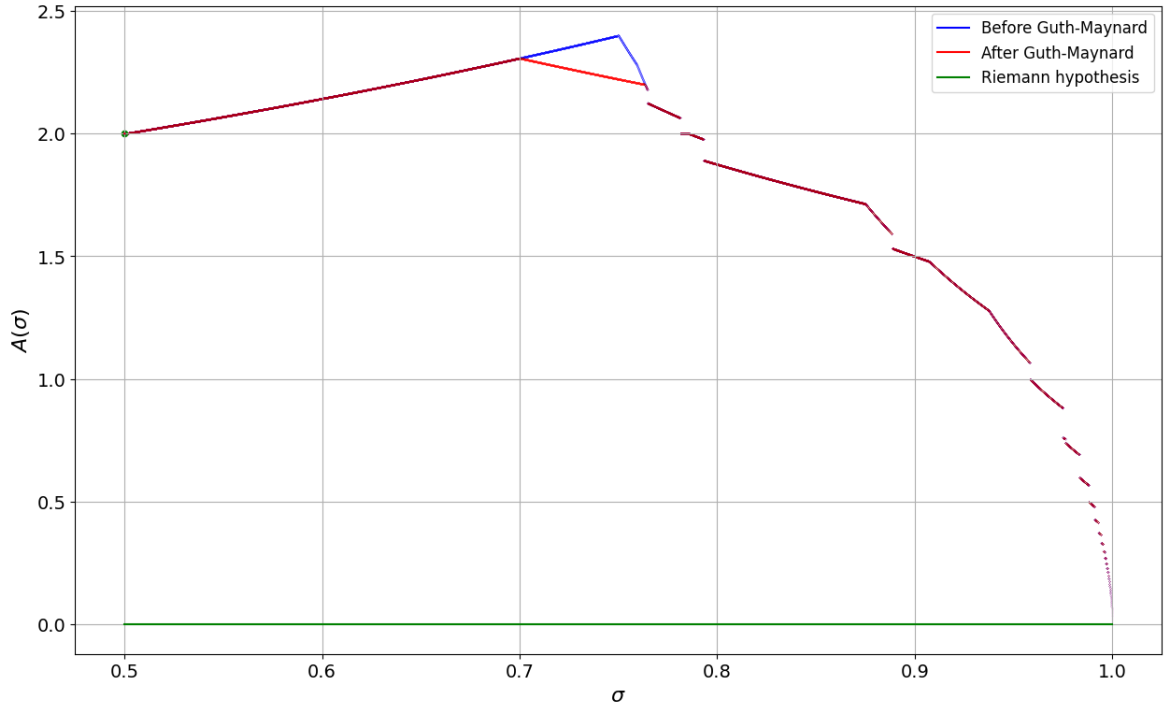


Figure 1.1: Best-known bounds for $A(\sigma)$

Figure 1.1 shows the best-known bounds for $A(\sigma)$, showing the improvement of Guth and Maynard's result over previous bounds. We obtained these bounds from a collection in [14].

1.2 Main results

We present the main results of this work. We first give an overview of important bounds on the number of zeros of the Riemann zeta function. Then, we present the main result of the preprint of Guth and Maynard [1], which, using bounds for the frequency of zeros of Dirichlet polynomials taking large values, gives a new bound on the number of zeros of the Riemann zeta function.

1.2.1 Bounds on zeros of the Riemann zeta function

First, we present key bounds on the number of zeros of the Riemann zeta function known to date, which, as we have seen in Theorem 1.17, give information about the distribution of primes in short intervals. We start by presenting a trivial bound. Then we give an overview of Ingham's bound from 1940. Lastly, we study the recent Guth and Maynard's bound, where we develop

the zero-detecting setup, which is a classical set-up for detecting zeros of the Riemann zeta function.

On the one hand, the Riemann-von Mangoldt formula 1.15 gives us the asymptotic

$$N(1/2,T) \asymp T\log T = T^{1+o(1)},$$

which, based on the definition of $A(\sigma)$, implies that $A(1/2) = 2$. On the other hand, Theorem 1.2 and Theorem 1.9 tell us that

$$N(1,T) = 0,$$

which implies that $A(1) = 0$. Since $N(\sigma,T)$ is decreasing in σ, and hence so is $A(\sigma)(1-\sigma)$, we get for $1/2 \le \sigma \le 1$

$$0 = A(1)(1-1) \le A(\sigma)(1-\sigma) \le A(1/2)(1-1/2) = 1,$$

which gives us the trivial "von Mangoldt" zero density theorem

$$A(\sigma) \le \frac{1}{1-\sigma}.$$

Ingham's result from 1940 gives us a better bound for $A(\sigma)$, and is still the best known bound for $1/2 \le \sigma \le 0.7$. This result was proven in [3] using already developed results in [2], so we refer the interested reader to these works for a proof.

Theorem 1.19 (Ingham, 1940)**.** *We have*

$$N(\sigma,T) \lesssim T^{\frac{3}{2-\sigma}(1-\sigma)+o(1)}$$

for every $1/2 \le \sigma \le 1$.

This gives $A(\sigma) \le \frac{3}{2-\sigma}$, which is better than the trivial bound.

We review the classical set-up for detecting zeros of the Riemann zeta function, which was unified by Montgomery in [5, Chapter 12] in 1971 using previous developments. A modern development of the set-up can be found in [6, Appendix C].

Recall that the Möbius function is defined by

$$\mu(n) = \begin{cases} 1 & \text{if } n = 1, \\ (-1)^k & \text{if } n \text{ is a product of } k \text{ distinct primes}, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 1.20 (Type I/II zeros)**.** Let $\rho = \beta + i\gamma$ be a non-trivial zero of ζ with $\gamma \in [T, 2T]$.

1. We say ρ is a 'Type I zero' if

$$|D_N(\rho)| \ge \frac{1}{3\log T}$$

for some $N = 2^j \in [T^{1/100}, T^{1/2}(\log T)^2]$, where D_N is given by

$$D_N(s) := \sum_{n\in[N,2N]} \frac{a(n)}{n^s} \exp\Big(-\frac{n}{T^{1/2}}\Big), \tag{1.2}$$

$$a(n) := \sum_{\substack{d\mid n \\ d\le 2T^{1/100}}} \mu(d).$$

2. We say ρ is a 'Type II zero' if

$$\Big|\frac{1}{2\pi i}\int_{(-\beta+1/2)} T^{s/2}\Gamma(s)M(\rho+s)\zeta(\rho+s)\mathrm{d}s\Big| \geq \frac{1}{3},$$

where

$$M(s) := \sum_{m\leq 2T^{1/100}} \frac{\mu(m)}{m^s}.$$

Lemma 1.21. *Every non-trivial zero $\rho = \beta+i\gamma$ of the Riemann zeta function ζ with $\gamma \in [T, 2T]$ for T sufficently large is either a Type I zero or a Type II zero (or both).*

The strategy is to show that these conditions can only hold infrequently. Let $R_I(\sigma, T)$ denote the number of Type I zeros with $\beta \geq \sigma$ and $\gamma \in [T, 2T]$, and let $R_{II}(\sigma, T)$ denote the number of Type II zeros with $\beta \geq \sigma$ and $\gamma \in [T, 2T]$. The Type II zeros cause few problems.

Lemma 1.22. *We have*

$$R_{II}(\sigma, T) \ll T^{2(1-\sigma)}(\log T)^{O(1)}.$$

A proof of this Lemma is given in [6, Lemma 6.3]. The hardest part of this zero detecting set-up is bounding the number of Type I zeros. For this we need to control the Dirichlet polynomial $D_N(s)$ defined in (1.2), which is the main object of study in the next section.

The main result of Guth and Maynard's article [1] is the following. We give an overview of the proof in the next subsection.

Theorem 1.23 (Large values estimate)**.** *Suppose (b_n) is a sequence of complex numbers with $|b_n| \leq 1$, and $(t_r)_{r\leq R}$ is a sequence of 1-separated points in $[0, T]$ such that*

$$\Big|\sum_{n=N}^{2N} b_n n^{it_r}\Big| \geq V$$

for all $r \leq R$. Then we have

$$R \leq T^{o(1)}\Big(N^2V^{-2} + N^{18/5}V^{-4} + TN^{12/5}V^{-4}\Big).$$

Theorem 1.23 allows to bound the number of Type I zeros, which gives the following novel result of Guth and Maynard, giving a bound on the number of zeros of the Riemann zeta function. This is known as a zero density estimate.

Theorem 1.24 (Zero density estimate)**.** *We have*

$$N(\sigma, T) \lesssim T^{\frac{15}{3+5\sigma}(1-\sigma)+o(1)}.$$

for every $1/2 \leq \sigma \leq 1$.

This gives $A(\sigma) \leq 15/(3+5\sigma)$, which is better than previous bounds for $0.7 \leq \sigma \leq 0.8$. Most importantly, this bound reduces $\|A\|_\infty$ from $12/5 = 2.4$ (thanks to a 1972 result of Huxley), to $30/13 = 2.307\ldots$. This gives more information on the behavior of primes, for example, applying $A = \|A\|_\infty$ to Theorem 1.17.

Proof of Theorem 1.24. Theorem 1.24 follows from Ingham's result 1.19 if $\sigma \leq 7/10$ and a result given by Huxley in [7] if $\sigma \geq 8/10$, which is $A(\sigma) \leq 3/(3\sigma - 1)$, so we may assume that $\sigma \in [7/10, 8/10]$. It suffices to show the bound of Theorem 1.24 for zeros with imaginary part in $[T, 2T]$, since the result for $[0, T]$ then follows by considering $T/2, T/4, \dots$ in place of T.

Now, given a parameter N, we consider the Dirichlet polynomials defined in (1.2). Since we already know that the number of Type II zeros is sufficiently small because of Lemma 1.21, it suffices to bound the number of Type I zeros. There are $O(\log T)$ choices of N so we focus on the value of N which gives the largest number of Type I zeros.

We now make a slight modification to D_N to remove the dependencies on the real parts. Let $\rho = \beta + i\gamma$ be a Type I zero with $\beta \geq \sigma$, and let $\psi(u)$ be a smooth function equal to $e^{u(\beta-\sigma)}$ on $[\log N, \log 2N]$ and supported on $[(\log N)/2, 2\log N]$ with $\|\psi^{(j)}(t)\|_\infty \lesssim_j t^{-j}$ for all $j \in \mathbb{N}$. We then note that by Fourier expansion

$$D_N(\rho) = \sum_{n \in [N,2N]} b_n n^{-\sigma - i\gamma} \psi(\log n) = \frac{1}{2\pi i} \int_\xi \hat{\psi}(\xi) \Big(D_N(\sigma + i(\gamma + 2\pi\xi)) \Big) \mathrm{d}\xi.$$

Since $\hat{\psi}$ is rapidly decreasing, we may truncate the integral to $\xi \lessapprox 1$ at the cost of an $O(T^{-100})$ error term. Therefore we see that if ρ is a Type I zero, we have $|D_N(\sigma + i\gamma + i\xi)| \gtrapprox 1$ for some $\xi \lessapprox 1$. There are $O(\log T)$ non-trivial zeros $\rho = \beta + i\gamma$ with $\gamma \in [t, t+1]$ for any $t \in [T, 2T]$. Therefore we can find a 1-separated set of points $(s_r)_{r \leq R}$ in $[T, 2T]$ with $|D_N(\sigma + is_r)| \gtrapprox 1$ and the number R of points satisfies $R \gtrapprox N(\sigma, 2T) - N(\sigma, T)$. Let

$$\tilde{b}_n := \Big(\frac{N}{n}\Big)^\sigma b_n, \qquad \tilde{D}_N(t) := \sum_{n \in [N,2N]} \tilde{b}_n n^{it} = N^\sigma D_N(\sigma + it).$$

Thus it suffices to show that if $N < T^{1/2 + o(1)}$ and W is a 1-separated set in $[T, 2T]$ such that $|\tilde{D}_N(t)| \gtrapprox N^\sigma$, we have $|W| \lessapprox T^{15(1-\sigma)/(3+5\sigma) + o(1)}$.

If $T^{5/(3+5\sigma)} \leq N^2 \leq T^{75(1-\sigma)/(54+30\sigma-100\sigma^2)}$, then we use Theorem 1.23 applied to the Dirichlet polynomial $\tilde{D}_N^2$ of length N^2, which shows that (noting that $\sigma \in [7/10, 8/10]$ implies that the $N^2 V^{-2}$ term is dominated by the $N^{18/5} V^{-4}$ term)

$$\begin{aligned} |W| &\lessapprox (T^{75(1-\sigma)/(54+30\sigma-100\sigma^2)})^{18/5-4\sigma} + T(T^{5/(3+5\sigma)})^{12/5-4\sigma} \\ &\lessapprox T^{15(1-\sigma)/(3+5\sigma)}. \end{aligned}$$

If instead N lies outside of these ranges, using classical estimates applying the Mean Value Theorem gives good enough bounds. To see proof of these explicitly we refer the reader to [1, Theorem 1.2].

□

1.2.2 Main result of Guth and Maynard

Our main task is to prove Theorem 1.23. Theorem 1.23 would be a consequence of the following proposition.

Proposition 1.25. *Let $\sigma \in [0.7, 0.8]$, $w \in C^\infty(\mathbb{R})$, a real function supported on $[1, 2]$ that values 1 on $\big(t[\frac{6}{5}, \frac{9}{5}]\, ght]$, and $||w^{(j)}||_\infty < \infty$ for every $j \geq 0$. Suppose (b_n) is a sequence of complex numbers with $|b_n| \leq 1$, and $W \subset [0, T]$, is a set of 1-separated points such that*

$$\left(t| \sum_n w\left(t(\frac{n}{N})\, ght \right) b_n n^{it} \right) ght| \geq N^\sigma, \quad \forall t \in W,$$

then

$$|W| \leq CT^{o(1)}(N^{\frac{18}{5} - 4\sigma} + TN^{\frac{12}{5} - 4\sigma}).$$

Recalling a Dirichlet Polynomial as the expression

$$D_N(t) = \sum_n w\left(\frac{n}{N}\right) b_n n^{it}.$$

We will prove Proposition 1.25. Given $W = \{t_1, \dots, t_r\} \subset \mathbb{R}$, let M_W be the $r \times (N-1)$-sized matrix defined as

$$M_W := \begin{bmatrix} w\left(t(\frac{N+1}{N})\, ght\right)(N+1)^{it_1} & w\left(t(\frac{N+2}{N})\, ght\right)(N+2)^{it_1} & \cdots & w\left(t(\frac{2N-1}{N})\, ght\right)(2N-1)^{it_1} \\ w\left(t(\frac{N+1}{N})\, ght\right)(N+1)^{it_2} & w\left(t(\frac{N+2}{N})\, ght\right)(N+2)^{it_2} & \cdots & w\left(t(\frac{2N-1}{N})\, ght\right)(2N-1)^{it_2} \\ \vdots & \vdots & \ddots & \vdots \\ w\left(t(\frac{N+1}{N})\, ght\right)(N+1)^{it_r} & w\left(t(\frac{N+2}{N})\, ght\right)(N+2)^{it_r} & \cdots & w\left(t(\frac{2N-1}{N})\, ght\right)(2N-1)^{it_r} \end{bmatrix},$$

or equivalently,

$$M_W(k,j) = w\left(t(\frac{N+j}{N}\right) ght)(N+j)^{it_k}.$$

Let us see that the number of large values of D_N is controlled by this matrix's singular values.

Lemma 1.26. *Let $s_j(M_W)$ be the j^{th} largest singular value of M_W. If $|D_N(t)| \geq N^\sigma$ for all $t \in W$, and $|b_n| \leq 1$, then*

$$|W| \leq N^{1-2\sigma} |s_1(M_W)|^2.$$

Proof. Let $b = (b_n)_n$. First, observe that

$$D_N(t) = \sum_n w\left(\frac{n}{N}\right) b_n n^{it} = (M_W b)_t.$$

On the other hand, since $N^{2\sigma} \leq |D_N(t)|^2$, we have

$$|W| N^{2\sigma} \leq \sum_{t \in W} |D_N(t)|^2 = \|M_W b\|_2^2 \leq \|M_W\|_2^2 \|b\|_2^2 \leq |s_1(M_W)|^2 N.$$

□

We also know that the infinity norm of a vector is smaller than any r-norm, so we can find out the following inequality

$$|s_1(M_W)|^2 = s_1(\overline{M_W}^T M_W) \leq \left(\sum_j s_j(\overline{M_W}^T M_W)^r\right)^{\frac{1}{r}} = \left(tr((\overline{M_W}^T M_W)^r)\right)^{\frac{1}{r}}.$$

We will find bounds for the trace of $(\overline{M_W}^T M_W)^3$. If we notice that

$$(\overline{M_W}^T M_W)_{t_1,t_2} = \sum_n w\left(\frac{n}{N}\right)^2 n^{i(t_1-t_2)},$$

it seems natural to study the function $h_t(u) := w(u)^2 u^{it}$. We will expand the trace of the cubed matrix, and then perform Poisson summation, so we need to bound the Fourier transform of h_t.

Lemma 1.27. *For any integer $j \geq 0$, we have*

$$\widehat{h_t}(\xi) \lesssim_j \frac{(1+|t|)^j}{|\xi|^j}$$

and

$$\widehat{h_t}(\xi) \lesssim_j \frac{(1+|\xi|)^j}{|t|^j}.$$

We will apply these two lemmas to bound the trace of the cubed matrix.

Proposition 1.28. *If W is T^ε-separated, then*

$$tr((\overline{M_W}^T M_W)^3) = N^3|W|||w||_{L^2}^6 + \sum_{m\in\mathbb{Z}^3-\{0\}} I_m + O_\varepsilon(T^{-100}),$$

where $m = (m_1, m_2, m_3)$ and

$$I_m = N^3 \sum_{t_1,t_2,t_3\in W} \widehat{h}_{t_1-t_2}(m_1N)\widehat{h}_{t_2-t_3}(m_2N)\widehat{h}_{t_3-t_1}(m_3N).$$

Proof. Let us expand the trace of the cubed matrix

$$\begin{aligned} tr((\overline{M_W}^T M_W)^3) &= \sum_{m_1,m_2,m_3\in\mathbb{Z}} \sum_{t_1,t_2,t_3\in W} w\left(\frac{m_1}{N}\right)^2 w\left(\frac{m_2}{N}\right)^2 w\left(\frac{m_3}{N}\right)^2 m_1^{i(t_1-t_2)} m_2^{i(t_2-t_3)} m_3^{i(t_3-t_1)} \\ &= \sum_{m_1,m_2,m_3\in\mathbb{Z}} \sum_{t_1,t_2,t_3\in W} h_{t_1-t_2}\left(\frac{m_1}{N}\right) h_{t_2-t_3}\left(\frac{m_2}{N}\right) h_{t_3-t_1}\left(\frac{m_3}{N}\right), \end{aligned}$$

then by Poisson summation, we get

$$\begin{aligned} tr((\overline{M_W}^T M_W)^3) &= \sum_{m_1,m_2,m_3\in\mathbb{Z}} \sum_{t_1,t_2,t_3\in W} \widehat{h_{t_1-t_2}\left(\frac{m_1}{N}\right)} \widehat{h_{t_2-t_3}\left(\frac{m_2}{N}\right)} \widehat{h_{t_3-t_1}\left(\frac{m_3}{N}\right)} \\ &= N^3 \sum_{m_1,m_2,m_3\in\mathbb{Z}} \sum_{t_1,t_2,t_3\in W} \widehat{h}_{t_1-t_2}(m_1N)\, \widehat{h}_{t_2-t_3}(m_2N)\, \widehat{h}_{t_3-t_1}(m_3N). \end{aligned}$$

Let us see the term $I_0 = N^3 \sum_{t_1,t_2,t_3\in W} \widehat{h}_{t_1-t_2}(0)\, \widehat{h}_{t_2-t_3}(0)\, \widehat{h}_{t_3-t_1}(0)$. By Lemma 1.27, if $t_1 \neq t_2$, then $\widehat{h}_{t_1-t_2}(0) \lesssim \frac{1}{|t_1-t_2|^{100}} \lesssim_\varepsilon T^{-100}$. So all the terms are negligible unless $t_1 = t_2 = t_3$, which gives

$$I_0 = N^3|W|\widehat{h}_0(0)^3 + N^3 O_\varepsilon(T^{-100}) = N^3|W|||w||_{L^2}^6 + N^3 O_\varepsilon(T^{-100}).$$

□

By Lemma 1.26 and Proposition 1.28, we get the following result.

Proposition 1.29. *Let W be T^ε-separated, and $|b_n| \leq 1$ such that $|D_N(t)| \geq N^\sigma$ for all $t \in W$, then*

$$|W| \lesssim_\varepsilon N^{2-2\sigma} + N^{1-2\sigma}\left(\sum_{m\in\mathbb{Z}^3-\{0\}} I_m\right)^{\frac{1}{3}}.$$

The remaining work is to bound $S = \sum_{m\in\mathbb{Z}^3-\{0\}} I_m$. It has been explained that if $t_1 \neq t_2$, then

$$\widehat{h}_{t_1-t_2}(0) \lesssim_\varepsilon T^{-100},$$

and $\widehat{h}_0(0) = \int w^2 \approx 1$. But we have not used the first claim of Lemma 1.27. Indeed, if $t_1 = t_2$ and $m \neq 0$, by first claim of Lemma 1.27, then

$$\widehat{h}_0(mN) \lesssim m^{-100}N^{-100}.$$

The last bound is useful when $mN > T^{1+\varepsilon}$, since $|t_1 - t_2| \leq T$ for all $t_1, t_2 \in W$ we get with $j > \frac{200}{\varepsilon} + 100$,

$$|\widehat{h}_{t_1-t_2}(mN)| \lesssim_\varepsilon \frac{(1+|t_1-t_2|)^j}{m^j N^j} \lesssim \frac{T^{100}}{m^{100}N^{100}} \frac{T^{\frac{200}{\varepsilon}}}{T^{(1+\varepsilon)200/\varepsilon}} = \frac{T^{100}}{m^{100}N^{100}} \frac{1}{T^{200}} \leq T^{-100}m^{-100}.$$

It seems natural to separate the sum S into different pieces, $S = S_1 + S_2 + S_3$, where S_1 are the summands where m has two zero components, S_2 the summands where m has only one zero component, and S_3 the summands where m has non-zero components. The results, which we enounce below, are [1, Proposition 5.1], [1, Proposition 6.1] taking $k = 4$, and [1, Proposition 11.2], respectively. In these we use the notation of Proposition 1.25.

Proposition 1.30. *We have*

$$S_1 = O_\varepsilon(T^{-10}),$$

$$|S_2| \lessapprox N^2|W|^2 + TN|W|^{7/4} + N^2|W|^{2-3/8}T^{1/8}.$$

Moreover, if $T \geq N^{3/4}$, then

$$|S_3| \lessapprox T^2|W|^{3/2} + T|W|N^{3-2\sigma} + T|W|^2N^{3/2-\sigma} + T^{9/8}|W|^{29/16}N^{3/2-\sigma}.$$

We can use these bounds to find a bound for $|W|$ following the argument left by Proposition 1.29.

Proof of Proposition 1.25. We got from Proposition 1.29 that

$$|W|N^{2\sigma-1} \lesssim \varepsilon N + \left(\sum m \in \mathbb{Z}^3 - 0 I_m\right)^{\frac{1}{3}}.$$

This and Proposition 1.30 imply that

$$\begin{aligned} |W|^3N^{6\sigma-3} &\lesssim_\varepsilon N^3 + S_1 + S_2 + S_3 \\ &\lessapprox N^3 + N^2|W|^2 + TN|W|^{7/4} + N^2|W|^{2-3/8}T^{1/8} \\ &\quad + T^2|W|^{3/2} + T|W|N^{3-2\sigma} + T|W|^2N^{3/2-\sigma} + T^{9/8}|W|^{29/16}N^{3/2-\sigma}, \end{aligned}$$

choosing $T = N^{\frac{6}{5}}$,

$$\begin{aligned} |W| &\lessapprox T(N^{(4-10\sigma)/5} + N^{(19-30\sigma)/5} + N^{(74-120\sigma)/25} + N^{(298-480\sigma)/95} \\ &\quad + N^{(12-20\sigma)/5} + N^{(9-14\sigma)/5} + N^{(354-560\sigma)/95}). \end{aligned}$$

Finally, if $\sigma \in [0.7, 0.8]$ we get $|W| \lessapprox TN^{(12-20\sigma)/5}$. □

1.3 Conclusions

We have seen that finding better bounds for $A(\sigma)$ function needs the development of new techniques, and these bounds are more and more difficult to improve. Dirichlet polynomials are the main tool that Guth and Maynard treated, and a better bound was proven in a short interval. For instance, researchers are seeking for better estimations of the $A(\sigma)$ function, which will not be sufficient to prove Riemann hypothesis. It seems that Riemann hypothesis will remain as a Millennium problem for a long time.

References

[1] Larry Guth and James Maynard, *New large value estimates for Dirichlet polynomials*, arXiv preprint arXiv:2405.20552 [math.NT], 2024.

[2] A. E. Ingham, *On the difference between consecutive primes*, The Quarterly Journal of Mathematics, **os-8** : 255–266, 1937.

[3] A. E. Ingham, *On the estimation of $N(\sigma, T)$*, The Quarterly Journal of Mathematics, **os-11** : 201–202, 1940.

[4] E. C. Titchmarsh and D. R. Heath-Brown, *The Theory of the Riemann Zeta-function*, Oxford Science Publications, Clarendon Press, 1986.

[5] H. L. Montgomery, *Topics in Multiplicative Number Theory*, Lecture Notes in Mathematics, Springer Berlin Heidelberg, 2006.

[6] James Maynard and Kyle Pratt, *Half-Isolated Zeros and Zero-Density Estimates*, International Mathematics Research Notices, **2024** : 12978–13014, 2024.

[7] M. N. Huxley, *On the Difference between Consecutive Primes*, Inventiones mathematicae, **15** : 164–170, 1971/72.

[8] J. Hadamard, *Sur la distribution des zéros de la fonction $\zeta(s)$ et ses conséquences arithmétiques*, Bulletin de la Société Mathématique de France, **24** : 199–220, 1896.

[9] E. M. Stein and R. Shakarchi, *Complex Analysis*, Princeton Lectures in Analysis, Princeton University Press, 2010.

[10] C. J. de La Vallée Poussin, *Recherches analytiques sur la théorie des nombres premiers*, Hayez, 1897.

[11] P. Erdős, *On a New Method in Elementary Number Theory Which Leads to An Elementary Proof of the Prime Number Theorem*, Proceedings of the National Academy of Sciences, **35** : 374–384, 1949.

[12] Atle Selberg, *An Elementary Proof of the Prime-Number Theorem for Arithmetic Progressions*, Canadian Journal of Mathematics, **2** : 66–78, 1950.

[13] Bernhard Riemann, *Über die Anzahl der Primzahlen unter einer gegebenen Grösse*, Monatsberichte der Berliner Akademie, 671–680, 1859.

[14] Timothy S. Trudgian and Andrew Yang, *Toward optimal exponent pairs*, arXiv preprint arXiv:2306.05599 [math.NT], 2024.

Comparative Analysis of Iterative Methods for Real-Time Selective Harmonic Elimination in Multilevel Inverters

Maldonado María Emilia [♭] *, Tarazona Lévano Mauro [♮] *, Vivert Miguel [♯] *

(∗) Universitat Politècnica de València & Universitat de València

(♭) memalmac@posgrado.upv.es

(♮) mgtarlev@teleco.upv.es

(♯) mevivdel@upv.edu.es

1.1 Introduction

Multilevel inverters have gained substantial attention in high-power and high-voltage applications due to their ability to reduce harmonic distortion and enhance overall efficiency. One widely studied technique for improving the output waveform quality is Selective Harmonic Elimination (SHE), which aims to eliminate specific harmonic orders while maintaining the desired fundamental component. In the existing literature, several methods rely on symbolic computations (e.g., Groebner basis transformations in tools like Maple) to find the switching angles required for SHE. Although these approaches yield accurate solutions, they are often computationally demanding and less adaptable in real-time scenarios. To address this gap, the present work builds upon the strategy proposed in "Adaptive Real-Time Selective Harmonic Elimination for a Cascaded Full-Bridge Multilevel Inverter," which introduced a fourth-degree nonlinear equation describing the switching angles. Previous results confirmed the effectiveness of the polynomial formulation in reducing significant odd-order harmonics (e.g., 3rd, 5th, and 7th) while satisfying the fundamental voltage requirement. However, to achieve real-time adaptability and lower computational overhead, alternative numerical methods—encompassing Newton-based schemes, evolutionary algorithms, and swarm intelligence—are investigated here. By systematically comparing the convergence rate, accuracy, and execution cost of these methods, this work offers new insights and guidelines for selecting robust and efficient numerical approaches for adaptive control in multilevel inverters.

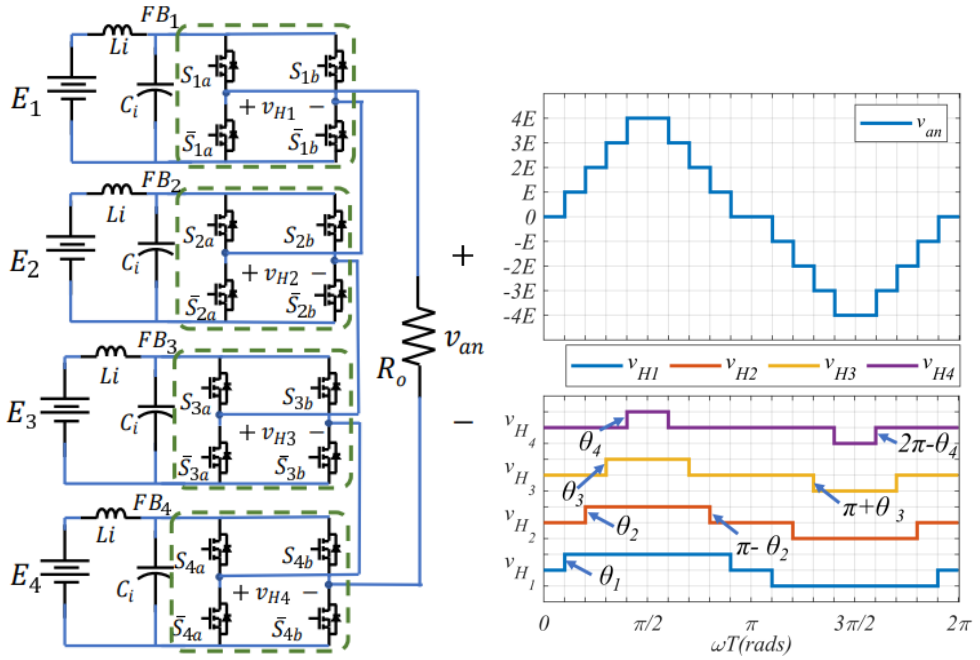
1.2 System Description and Problem Formulation

Consider a *Cascaded Full-Bridge Multilevel Inverter* (CFBMI) of four full-bridge cells (FBCs) like figure-1.1, each supplied by a DC voltage E_k. If each FBC switches at angles θ_k symmetrically around $\pi/2$, the output voltage per cell is

$$v_{H_k}(t) \;=\; \sum_{\substack{l=1 \\ l\,\text{odd}}}^{\infty} \frac{4\,E_k}{l\,\pi}\,\cos(l\,\theta_k)\,\sin(l\,\omega t), \tag{1.1}$$

hence the total inverter output is

$$v_{an}(t) \;=\; \sum_{k=1}^{4} v_{H_k}(t) \;=\; \frac{4}{\pi} \sum_{\substack{l=1 \\ l\,\text{odd}}}^{\infty} \Big(\frac{1}{l} \sum_{k=1}^{4} E_k \,\cos(l\,\theta_k) \Big) \,\sin(l\,\omega t). \tag{1.2}$$



(a) (b)

Figure 1.1: (a) Cascaded Full-Bridge Inverter topology; (b) Output voltage waveform.

Defining the l-th harmonic amplitude as

$$h_l \;=\; \frac{4}{l\,\pi} \sum_{k=1}^{4} E_k \,\cos(l\,\theta_k), \tag{1.3}$$

and letting each $E_k = E\,(1+\delta_k)$, one obtains normalized harmonics

$$h'_l \;=\; \frac{h_l}{E} \;=\; \frac{4}{l\,\pi} \sum_{k=1}^{4} \big(1+\delta_k\big) \,\cos(l\,\theta_k). \tag{1.4}$$

Utilizing Chebyshev polynomials $T_l(x_k)$ with $x_k = \cos(\theta_k)$, these amplitudes become

$$h'_l \;=\; \frac{4}{l\,\pi} \sum_{k=1}^{4} \big(1+\delta_k\big)\, T_l\big(x_k\big). \tag{1.5}$$

Selective Harmonic Elimination (SHE) imposes

$$h'_1 \;=\; H'_1, \quad h'_3 \;=\; 0, \quad h'_5 \;=\; 0, \quad h'_7 \;=\; 0,$$

yielding a fourth-degree polynomial system in the variables $\{x_k\}$. A Gröbner-basis transformation (via Buchberger's algorithm) reshapes this system into a triangular form:

$$Q\big(x_1, x_2, x_3, x_4,\, H'_1\big) \;=\; 0, \tag{1.6}$$

where each polynomial enforces fundamental amplitude while annihilating targeted harmonics. Solving this system produces the switching angles $\theta_k = \arccos(x_k)$, which ensures the desired harmonic elimination for a given fundamental voltage H_1' and any voltage perturbations δ_k. In what follows, we explore numerical methods capable of solving these polynomial equations efficiently for real-time or near-real-time control.

1.3 Proposed Iterative Methods

Efficiently solving the polynomial system derived from the SHE conditions requires robust algorithms. Below, we outline six iterative methods, highlighting their core update schemes and convergence properties.

1.3.1 Monovariable Approaches

Newton-Raphson

For a scalar equation $f(x) = 0$, the classical iteration is

$$x_{n+1} \;=\; x_n \;-\; \frac{f(x_n)}{f'(x_n)}, \tag{1.7}$$

with quadratic convergence ($\mathcal{O}(|x-\alpha|^2)$) if $f'(x_n) \neq 0$. Applied to the univariate polynomial arising from, e.g., a Gröbner-basis factor, it rapidly converges to a root provided the initial guess x_0 is close to the actual solution.

Modified Newton-Raphson

Including second-derivative information,

$$x_{n+1} \;=\; x_n \;-\; \frac{f(x_n)\, f'(x_n)}{\big[f'(x_n)\big]^2 \;-\; f(x_n)\, f''(x_n)}, \tag{1.8}$$

yields cubic convergence ($\mathcal{O}(|x-\alpha|^3)$). This enhanced rate is beneficial when function evaluations, including $f''(x)$, are tractable. In SHE polynomials of degree 4–7, the second derivative is analytically feasible.

Ehrlich-Aberth Method

Specifically aimed at finding *all* roots of a polynomial $F(x)$ of degree n, it updates each approximate root $x_k^{(m)}$ via

$$x_k^{(m+1)} \;=\; x_k^{(m)} - \; \frac{F\big(x_k^{(m)}\big)}{F'\big(x_k^{(m)}\big) \;-\; F\big(x_k^{(m)}\big) \sum_{j\neq k} \dfrac{1}{x_k^{(m)} - x_j^{(m)}}}. \tag{1.9}$$

The correction term prevents different x_k from converging to the same root. This method is particularly advantageous in polynomial-based SHE, ensuring all physical roots ($\cos(\theta_k)$) are obtained in a single procedure.

1.3.2 Multivariable and Hybrid Approaches

Newton-Raphson for Systems

For $\mathbf{F}(\mathbf{x}) = \mathbf{0}$, each iteration uses the Jacobian $\mathbf{J}(\mathbf{x}_n)$:

$$\mathbf{x}_{n+1} \;=\; \mathbf{x}_n \;-\; \mathbf{J}\big(\mathbf{x}_n\big)^{-1}\,\mathbf{F}\big(\mathbf{x}_n\big), \tag{1.10}$$

with quadratic convergence near the solution if $\mathbf{J}(\mathbf{x}_*)$ is nonsingular. For the SHE system, $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$, this approach may converge rapidly to $\cos(\theta_k)$ when given a well-chosen initial estimate.

Newton + Differential Evolution

To mitigate the sensitivity of Newton's method to initial guesses, a *hybrid* algorithm first applies Differential Evolution (DE), which evolves a population $\{\mathbf{x}_i\}$ through mutation,

$$\mathbf{v}_i = \mathbf{x}_{r_1} + F\big(\mathbf{x}_{r_2} - \mathbf{x}_{r_3}\big),$$

and crossover, then selects the best candidate as $\mathbf{x}_0$. Subsequently, a few Newton steps refine this candidate to achieve faster local convergence. This global-local strategy ensures robust and efficient solutions for polynomial systems in SHE.

Particle Swarm Optimization

PSO is a derivative-free global approach where each particle $\mathbf{x}_i$ updates velocity $\mathbf{v}_i$ and position based on local ($\mathbf{pbest}_i$) and global ($\mathbf{gbest}$) optima:

$$\mathbf{v}_i^{(k+1)} = w\,\mathbf{v}_i^{(k)} \;+\; c_1\,\mathbf{R}_1\big(\mathbf{pbest}_i - \mathbf{x}_i^{(k)}\big) \;+\; c_2\,\mathbf{R}_2\big(\mathbf{gbest} - \mathbf{x}_i^{(k)}\big), \tag{1.11}$$

$$\mathbf{x}_i^{(k+1)} = \mathbf{x}_i^{(k)} + \mathbf{v}_i^{(k+1)}, \tag{1.12}$$

leading to wide exploration in multimodal solutions. For SHE, PSO converges reliably without requiring derivatives, although generally with higher computational effort than Newton-based methods.

1.4 Numerical Experiments

Numerical tests were conducted to evaluate convergence speed, residue accuracy, and computational cost under a uniform tolerance of 10^{-10} and a maximum of 100 iterations for each method.

1.4.1 Implementation Details

All algorithms were implemented in MATLAB (R2022b) on a 3.0 GHz Intel Core i7 PC with 16 GB RAM. Each method was given:

- **Tolerance:** $\epsilon = 10^{-10}$
- **Max. iterations:** 100
- **Performance metrics:** CPU time (in seconds), iteration count, and the final residue $\|\mathbf{F}(\mathbf{x}^*)\|$

For monovariable approaches (Newton, Modified Newton, Ehrlich-Aberth), we used the univariate polynomials derived from the Gröbner factorization; for multivariable methods (Newton System, Newton+DE, PSO), the full system $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ was directly solved.

1.4.2 Comparison of Methods

Table 1.1 summarizes the key results. Reported values represent median performance over multiple trials with different initial guesses:

Table 1.1: Comparative Performance of Iterative Methods

Method	**Iterations**	**Order**	**CPU Time (s)**	**Final Residue**
Newton Mono	5–9	2	0.6–0.7	$\sim 10^{-13}$
Modified Newton	4–8	3	0.08–0.09	$\sim 10^{-12}$
Ehrlich-Aberth	5–27	3	0.7–0.8	$\sim 10^{-15}$
Newton System	6–8	2	0.003–0.004	$\sim 10^{-16}$
Newton + DE	250+2	2	0.18–0.20	$\sim 10^{-11}$
PSO	100+	–	0.04–0.05	$\sim 10^{-11}$

Key observations:

- *Newton-based methods* (monovariable or multivariable) yield rapid convergence and minimal residue when the initial guess is sufficiently close.
- The *Hybrid Newton+DE* approach consistently finds good initial guesses, proving robust under parameter mismatches.
- *PSO*, although derivative-free and more global, requires a larger iteration count and thus higher overall computational effort.

1.5 Conclusions

Numerical experiments reveal that multivariable Newton-Raphson converges most rapidly for real-time Selective Harmonic Elimination, provided a suitable initial guess is available. In scenarios where the starting point is unknown or subject to uncertainty, the hybrid *Newton+DE* approach offers robust global exploration followed by fast local refinement. Traditional monovariable Newton-type methods are practical for reduced systems or when Gröbner-based factorization yields tractable polynomial factors. Particle Swarm Optimization remains a useful alternative for highly nonlinear or parameter-uncertain models, albeit with higher computational expense.

References

[1] A. M. Ostrowski, *Solution of Equations and Systems of Equations*, 2nd ed., Academic Press, New York, 1966.

[2] L. W. Ehrlich, "A Modified Newton Method for Polynomials", *Communications of the ACM*, vol. 10, no. 2, pp. 107–108, 1967.

[3] R. Storn and K. Price, "Differential Evolution — A Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Spaces", *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[4] J. Kennedy and R. Eberhart, "Particle Swarm Optimization", in *Proceedings of IEEE International Conference on Neural Networks*, Perth, Australia, 1995, pp. 1942–1948.