

Una incompleta y pobremente ilustrada
Introducción al Cálculo Numérico
para Ciencias Naturales
(usando Python)

Los sospechosos de siempre

29 de febrero de 2020

Índice general

1. Aproximación de Funciones	13
1.1. Polinomios de Taylor	13
1.1.1. Estimación del error	14
1.1.2. Aproximación de Padé*	18
1.2. Interpolación polinomial	19
1.2.1. Forma de Lagrange	20
1.2.2. Forma de Newton	23
1.2.3. Diferencias divididas	25
1.3. Fórmula de error de interpolación	26
1.4. Interpolación de Hermite	28
1.5. Interpolación lineal y cúbica segmentada	29
1.5.1. Interpolación lineal	30
1.5.2. Interpolación cúbica	30
1.6. Aplicación	31
1.6.1. Temperatura ambiente	31
1.6.2. Integración	32
1.7. Ejercicios	33
2. Resolución de Ecuaciones no Lineales	35
2.1. Método de bisección	35
2.1.1. Algoritmo	35
2.1.2. Fórmula del error	37
2.2. Método de Newton	37
2.2.1. Interpretación geométrica	37
2.2.2. Algoritmo de Newton	38
2.2.3. Fórmula del error para el método de Newton	40
2.2.4. Resultados de convergencia global	41
2.3. Método de punto fijo	42
2.3.1. Análisis de convergencia para funciones derivables	44
2.4. Métodos de punto fijo para sistemas de ecuaciones	44
2.5. Método de la secante	46
2.5.1. Fórmula del error para el método de las secante	47
2.6. Aplicaciones	49
2.6.1. Red de resistores	49
3. Resolución Numérica de Ecuaciones Diferenciales	51
3.1. Problemas de valores iniciales	51
3.1.1. Crecimiento poblacional	51
3.1.2. Circuitos R-C y R-L	52
3.1.3. Caída libre	53

3.2.	Método de Euler	53
3.2.1.	Error de truncamiento del método de Euler	55
3.2.2.	Error global	56
3.2.3.	Estimación del error	57
3.2.4.	Extrapolación	58
3.2.5.	Método de Euler implícito	59
3.3.	Métodos de Taylor	59
3.4.	Métodos Runge-Kutta	60
3.5.	Problemas en dimensión mayor	61
3.5.1.	Ecuaciones de orden superior	64
3.6.	Aplicaciones	65
3.6.1.	Modelo depredador-presa (Lotka–Volterra)	65
3.6.2.	Modelo FitzHugh–Nagumo	67
3.6.3.	Atractor de Lorenz	68
3.6.4.	Métodos homotópicos para ecuaciones no lineales	69
3.6.5.	Teoría del cable	71
3.7.	Ejercicios	73
4.	Cadenas de Markov	79
4.1.	Introducción	79
4.2.	Estados	81
4.3.	Matrices de Markov	82
4.3.1.	Espectro de matrices de Markov	83
4.4.	Dinámica	84
4.4.1.	Estados de equilibrio	84
4.4.2.	Estados límite	85
4.4.3.	Existencia de estados límite	85
4.4.4.	Dependencia del estado inicial	86
4.4.5.	Cadenas de Markov regulares	87
4.4.6.	Cadenas de Markov absorbentes	88
4.4.7.	Evolución probabilística*	88
4.5.	Aplicaciones	89
4.5.1.	Evolución de poblaciones	89
4.5.2.	Genética	89
4.5.3.	Modelos epidemiológicos	90
4.6.	Ejercicios	90
5.	Análisis de Datos	95
5.1.	Ajuste por cuadrados mínimos	95
5.1.1.	Modelo lineal	95
5.1.2.	Interpretación geométrica	97
5.1.3.	Linealización	99
5.1.4.	Mínimos cuadrados generalizados y no lineales*	101
5.1.5.	Mínimos cuadrados para suma de funciones	102
5.1.6.	Modelo polinomial	103
5.2.	Modelo lineal multivariado	105
5.3.	Regresión de componentes principales	106
5.3.1.	Motivación gráfica	107
5.3.2.	Componentes principales mediante autovalores y autovectores	108
5.3.3.	Reducción de la dimensión	109

5.4.	Métodos de agrupamiento	109
5.4.1.	Ejemplos gráficos	109
5.4.2.	Método de k-medias	110
5.4.3.	Algoritmo de desplazamiento de medias	114
5.4.4.	DBSCAN	114
5.4.5.	Agrupamiento jerárquico	115
5.4.6.	Distancia entre clases	115
6.	Análisis de Fourier y Filtros	119
6.1.	Señales periódicas	119
6.1.1.	Señales armónicas	119
6.1.2.	Períodos y período mínimo	121
6.1.3.	Operaciones con señales periódicas	122
6.1.4.	Polinomios trigonométricos	124
6.1.5.	Cálculo de los coeficientes	124
6.1.6.	Aproximación de Fourier	125
6.1.7.	Potencia de una señal y convergencia en media cuadrática	128
6.1.8.	Soluciones periódicas de ecuaciones diferenciales	130
6.1.9.	Convolución	132
6.2.	Transformada discreta de Fourier	134
6.2.1.	Método de los trapecios	134
6.2.2.	Muestreo de señales	135
6.2.3.	Definición de DFT	136
6.2.4.	Propiedades de DFT	137
6.2.5.	Reconstrucción exacta y solapamiento	138
6.2.6.	Transformada rápida de Fourier	140
6.2.7.	Espectro de potencia	142
6.2.8.	Escala logarítmica. Decibeles	144
6.2.9.	Decibeles	144
6.3.	Señales aperiódicas	144
6.3.1.	Transformada de Fourier	145
6.3.2.	Propiedades de la transformada de Fourier	147
6.3.3.	Convolución de señales aperiódicas	148
6.3.4.	Teorema de muestreo	149
6.3.5.	Criterio de Nyquist	149
6.3.6.	Aproximación por funciones sinc	149
6.4.	Filtros	150
6.4.1.	Filtros pasa bajos, pasa banda y pasa altos	150
6.4.2.	Filtros en el dominio de la frecuencia	150
6.4.3.	Filtros en el dominio del tiempo	150
6.4.4.	Respuesta unitaria	150
6.4.5.	Función de transferencia	150
6.5.	Aplicación	150
6.5.1.	Detección de ondas Alfa y Beta en señales EEG	150
A.	Conceptos Básicos	151
A.1.	Trigonometría y números complejos	151
A.1.1.	Identidades trigonométricas	151
A.2.	Números complejos	152
A.2.1.	Exponenciales complejas	152

A.3. Serie geométrica	153
A.4. Delta de Kronecker	153
A.5. Notación de Landau	153
A.6. Operaciones aritméticas	154
A.6.1. Error relativo y dígitos significativos	154
A.6.2. Representación en punto flotante	154
A.6.3. Operaciones en punto flotante	156

Índice de figuras

1.1.	Gráfico de $f(x) = \text{sen}(x)$ y las aproximaciones de Taylor de orden $n = 3, 5, 7, 9$.	16
1.2.	Gráfico de $f(x) = \ln(1+x)$ y $P_n(x)$ con $n = 2, 3, 4, 5$.	17
1.3.	Gráfico de $f(x) = \text{sech}(x)$ y $P_n(x)$ con $n = 2, 4, 6, 20$.	18
1.4.	Gráfico de $f(x) = \ln(1+x)$ y las aproximaciones de Padé.	19
1.5.	Gráfico de $f(x) = \text{sen}(\pi x)$ y las aproximaciones polinomiales.	21
1.6.	Gráfico de $f(x) = \text{sech}(x)$ y las aproximaciones polinomiales.	22
1.7.	Ubicación de los puntos de interpolación con el criterio de Chebyshev ($n = 5$).	22
1.8.	Gráfico de $f(x) = \text{sech}(x)$ y las aproximaciones polinomiales.	23
1.9.	Gráfico de $L_5(x)$ para distintas distribuciones con $n = 8$.	23
1.10.	Teorema de Rolle.	27
1.11.	Teorema de Rolle generalizado ($n = 3$).	27
1.12.	Ajuste por segmentos lineales para $f(x) = 1.5e^{-x^2/2} + 0.4\cos(\pi x)$ en $[0, 3]$.	30
1.13.	Ajuste por segmentos cúbicos para $f(x) = 1.5e^{-x^2/2} + 0.4\cos(\pi x)$ en $[0, 3]$.	31
1.14.	Ajuste por segmentos cúbicos del número de chirridos/min en función de la temperatura.	32
2.1.	Método de bisección para la función $f(x) = 1.75x^3 - 3x - 1$.	37
2.2.	Gráfico de $f(x) = e^x - 45$ y las iteraciones del método de Newton.	38
2.3.	Gráfico de $v(\lambda)$, $T = 4500, 5000, 5500, 6000$. En línea de puntos mostramos λ_{\max} para cada T .	40
2.4.	Gráfico de $f(x) = \tanh(x) - 0.25$ y las iteraciones del método de Newton.	41
2.5.	Gráfico de $\phi_2(x)$ y las iteraciones del método de punto fijo.	44
2.6.	Rectángulo áureo de Euclides.	49
2.7.		49
3.1.	Crecimiento de la población	52
3.2.	Circuitos R-C y R-L simples.	52
3.3.	Error del método de Euler para la ecuación $\dot{x}(t) = x(t)$.	54
3.4.	Errores locales y globales.	55
3.5.	En gris, el rectángulo Q que contiene a $(t, x(t))$. Los puntos indican la solución aproximada (t_n, x_n) .	57
3.6.	Error local.	57
3.7.	Comparación de los errores de los métodos de Euler y Euler modificado.	59
3.8.	Errores de los métodos de Runge-Kutta para $\dot{x} = x$, $x(0) = 1$, $t \in [0, 2]$.	61
3.9.	Diagrama de fases	62
3.10.	Circuitos RC en cascada.	62
3.11.	Amplitud y fase de la solución periódica.	63
3.12.	Circuito R-L-C.	63
3.13.	Soluciones del circuito R-L-C.	64
3.14.	Diagrama de fases de las soluciones del circuito R-L-C.	64
3.15.	Partícula en un campo central de fuerzas.	65

3.16. Evolución de las poblaciones de lince canadienses y liebres raqueta de nieve. . .	66
3.17. Órbitas del sistema (3.2) para $\alpha = 0.25$, $\beta = 1.0$, $\gamma = \delta = 0.01$	67
3.18. Presa y predador correspondiente a los datos de la Figura 3.16.	67
3.19. Encuentro entre liebre y lince (© Canadian Museum of Nature).	68
3.20. Comportamiento del sistema (3.3): potencial vs. tiempo.	69
3.21. Comportamiento del sistema (3.3): diagrama de fases.	69
3.22. R. FitzHugh y la computadora analógica.	70
3.23. Atractor de Lorenz	70
3.24. Esquema de las conexiones.	71
3.25. Modelo de transmisión de señales a través de fibras nerviosas.	72
3.26.	77
4.1. Transiciones del semáforo.	81
4.2. Conjunto de estados \mathcal{P}_d	82
4.3. Interpretación probabilística de las matrices de Markov.	82
4.4.	84
4.5.	85
4.6. Sistema de tres nodos.	86
4.7.	87
4.8. Subcadena absorbente.	88
4.9.	88
4.10. Simulación del sistema.	89
4.11. Modelos de evolución de mutaciones.	90
4.12. El laberinto se abre unos pocos segundos cada hora.	91
4.13. Evolución de los sistemas.	92
5.1. Recta de ajuste $y = \alpha + \beta x$	97
5.2. Proyección ortogonal sobre subespacios de dimensión 1 y 2.	97
5.3.	99
5.4. Órbita elíptica y semieje mayor.	100
5.5. Gráfico de Z en función de p	105
5.6.	108
5.7.	111
5.8. Elección aleatoria de los puntos iniciales y diagrama de Voronoi.	112
5.9. Primer paso.	113
5.10. Cuarto paso: partición estable.	113
5.11. \mathcal{W}_k en función del número de particiones.	114
5.12. Dendograma correspondiente a $S = \{0, 2, 5.5, 6.5, 8\}$	116
6.1.	119
6.2. Señal armónica	120
6.3. Diferentes polarizaciones.	120
6.4. Figura de Lissajous para distintos valores de p y q	121
6.5. Períodos para distintos valores de p y q	121
6.6.	123
6.7. Cálculo de la media en los intervalos $[0, T]$ y $[\tau, \tau + T]$	124
6.8. Señal triangular.	126
6.9. Aproximaciones por polinomios trigonométricos de la señal triangular.	127
6.10. Señal cuadrada $x(t)$ y la aproximación $x_n(t)$	127
6.11. Fenómeno de Gibbs.	128
6.12.	131

6.13. Conversión de señales continuas en discretas.	135
6.14. Muestras de $x(t)$, $y(t)$ para $N = 8$	139
6.15. Raíces de la unidad para $N = 32$	142
6.16. Núcleo de Fejér para $N = 4, 6, 8$	143
6.17. Obtención de la señal discreta \mathbf{x}	144
6.18. Detección del período por el espectro de potencia.	144
6.19. Diagramas de Bode del circuito de la Figura 6.12(a) para $\tau = 1$ ms.	145
6.20.	145
6.21.	146
6.22.	146
6.23.	148
6.24. Función delta de Dirac y su transformada.	149
A.1. Conjunto \mathbf{F}_+ correspondiente al Ejemplo A.1.	155
A.2.	155
A.3. Mapa de bits para la representación doble precisión del estándar IEEE 754.	156

Índice de tablas

1.1.	Aproximaciones de Padé $Q_{m,n}$ de $f(x) = \ln(1+x)$ en $x_0 = 0$	18
1.2.	Errores $ \ln(1.5) - Q_{m,n}(1.5) $	19
1.3.	Número medio de chirridos/min. para individuos machos (n. ensiger).	31
2.1.	Iteraciones del método de bisección para $f(x) = 1.75x^3 - 3x - 1$	36
2.2.	Iteraciones del método de Newton para $e^x = 45$	38
2.3.	Iteraciones de la ecuación $(x - 2\pi)(\cos(x) - 1) = 0$	41
2.4.	Iteraciones del método de Newton para $\tanh(x) = 0.25$	41
2.5.	Iteraciones de $y = \phi_2(x)$	43
2.6.	Iteraciones de $\mathbf{x}_n = \phi(\mathbf{x}_{n-1})$ con $\mathbf{x}_0 = (0., 0., 1.)$	45
2.7.	Iteraciones del método de secantes para $e^x = 45$	47
3.1.	Error en $t = 0.1, 0.2, \dots, 1$ para $h = 0.1$ y $h = 0.01$	54
3.2.	Error para $h = 0.1$ y la estimación con $h = 0.05$	58
4.1.	Clasificación de las ciudades según el tamaño.	79
4.2.	Evolución de la población migrante en cada año.	80
4.3.	87
5.1.	96
5.2.	98
5.3.	Modelos no lineales y las linealizaciones.	99
5.4.	Valores del semieje mayor a y el período orbital τ	100
5.5.	101
5.6.	Datos de p y Z correspondientes a $T = 100$ K.	104
5.7.	Mediciones de las variables x_1, x_2, \dots, x_p, y	105
5.8.	106
5.9.	107
5.10.	108
5.11.	109
5.12.	Posibles particiones de $S = \{-2, -1, 1/5, 1, 2\}$	111
5.13.	115
5.14.	Distancia entre elementos (izquierda) y distancia entre clases (derecha).	116
6.1.	126
6.2.	Aproximación de los coeficientes de Fourier por DFT.	140
A.1.	Números de punto flotante con $t = 3$, $e_{\min} = -1$ y $e_{\max} = 3$	155

CAPÍTULO 1

Aproximación de Funciones

1.1. Polinomios de Taylor. Vamos a recordar la teoría de polinomios de Taylor. Como en todo el apunte, supondremos, salvo indicación contraria, que $f(x)$ es una función indefinidamente derivable en toda la recta real o en un intervalo de interés. El polinomio de Taylor no es otra cosa que la generalización de la noción de recta tangente a orden mayor. La función lineal tangente es, en un sentido que vamos a precisar, la mejor aproximación lineal de $f(x)$ cerca de x_0 . Geométricamente esto quiere decir que entre todas las rectas, la recta tangente es la que realiza el mayor contacto con el gráfico de la función cerca del punto. En forma analítica esto se expresa de la siguiente forma:

(a) $P_1(x_0) = f(x_0)$.

(b) $\lim_{x \rightarrow x_0} \frac{f(x) - P_1(x)}{x - x_0} = 0$.

Se puede ver que $P_1(x) = f(x_0) + f'(x_0)(x - x_0)$ es la única función lineal que verifica estas dos condiciones. En efecto, si no valiera la primera condición, el límite de la segunda sería ∞ . Para cualquier otra función lineal que pasa por el punto $(x_0, f(x_0))$, el límite es una constante, lo que quiere decir que la distancia entre la función y su aproximación lineal es (casi) proporcional a la distancia entre x y x_0 . Sólo $P_1(x)$ tiene la propiedad de acercarse a $f(x)$ más rápido que x a x_0 . Como resultado de estas condiciones se obtiene $P'_1(x_0) = f'(x_0)$.

Queremos extender esta noción a funciones cuadráticas, cúbicas, etc. La generalización es sencilla, vamos a buscar una función polinomial $P_n(x)$ de grado n (o menor), llamado polinomio de Taylor de orden n , tal que la distancia entre $f(x)$ y $P_n(x)$ decrezca más rápido que $(x - x_0)^n$, es decir

$$\lim_{x \rightarrow x_0} \frac{f(x) - P_n(x)}{(x - x_0)^n} = 0$$

Aplicando la regla de L'Hôpital reiteradamente, se puede ver que la condición anterior es equivalente a $P'_n(x_0) = f'(x_0)$, \dots , $P_n^{(n)}(x_0) = f^{(n)}(x_0)$. A partir de estas igualdades, podemos calcular el polinomio de Taylor de orden n alrededor del punto x_0 :

$$P_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \dots + \frac{1}{n!}f^{(n)}(x_0)(x - x_0)^n,$$

donde $n! = 1 \times 2 \times \dots \times n$. Si definimos $0! = 1$ y $f^{(0)}(x) = f(x)$, podemos escribir

$$P_n(x) = \sum_{k=0}^n \frac{1}{k!} f^{(k)}(x_0)(x - x_0)^k.$$

La idea es que a medida que aumentamos el orden n , $P_n(x)$ aproxime cada vez mejor a $f(x)$. Esto es cierto en muchos casos de interés, pero no siempre. Esto es discutido en una serie de

ejemplos sencillos que nos van a permitir entender las diferentes situaciones. La teoría completa de polinomios y series de Taylor excede el alcance de este apunte.

Ejemplo 1.1 (función exponencial). Si $f(x) = e^x$ y $x_0 = 0$, entonces $f^{(k)}(0) = 1$ y por lo tanto

$$P_n(x) = 1 + x + \cdots + \frac{1}{n!}x^n$$

Tomemos un valor de x cercano $x_0 = 0$, por ejemplo $x = 0.5$. El valor exacto tomando 11 decimales es $f(0.5) = 1.6487212707$. En la siguiente tabla mostramos los errores que se cometen con los polinomios de Taylor de distintos ordenes:

n	$P_n(x)$	$ f(x) - P_n(x) $
1	1.5000000000	0.149
2	1.6250000000	0.237×10^{-1}
3	1.6458333333	0.289×10^{-2}
4	1.6484375000	0.284×10^{-3}
5	1.6486979167	0.234×10^{-4}
6	1.6487196181	0.165×10^{-5}
7	1.6487211682	0.103×10^{-6}
8	1.6487212650	0.566×10^{-8}
9	1.6487212704	0.282×10^{-9}
10	1.6487212707	0.128×10^{-10}

Nota: Puede resultar poco intuitivo el significado de obtener un valor, mediante una medición o un cálculo, con un error (relativo) menor a 10^{-10} . Este error nos dice que el valor aproximado coincide con el valor exacto en los primeros 10 dígitos significativos. Para hacernos una idea concreta de lo que esto representa podríamos pensar la siguiente comparación: si midiéramos la distancia entre un punto de la ciudad de Buenos Aires y uno de la ciudad de Mendoza (aproximadamente 1000 km) con una precisión de 10 dígitos, el error sería menor que 100 μm , una longitud similar al grosor de un cabello humano.

1.1.1. Estimación del error. Para que cualquier aproximación que hagamos a un valor desconocido sea de utilidad, es necesario contar con una idea del error que se está cometiendo. En el ejemplo anterior, conocíamos los primeros 11 dígitos del valor exacto, pero esto es una situación artificial, en un problema real obviamente uno no conoce el valor que quiere calcular. Dicho resultado se obtuvo mediante algún otro procedimiento y la garantía de precisión solo pudo ser dada mediante una estimación del error.

Vamos a obtener una expresión para el resto de Taylor definido por $r_n(x) = f(x) - P_n(x)$, que nos permite estimar el error que se comete al evaluar P_n como aproximación de f . Para $x_1 > x_0$ (el caso $x_1 < x_0$ se estudia de forma similar), podemos escribir

$$f(x_1) = f(x_0) + \int_{x_0}^{x_1} f'(x)dx,$$

El segundo término del lado derecho lo reescribimos usando la fórmula de integración por partes de una forma un poco artificial, dado que $f'(x) = -(f'(x)(x_1 - x))' + f''(x)(x_1 - x)$, tenemos

$$\begin{aligned} f(x_1) &= f(x_0) - f'(x)(x_1 - x) \Big|_{x_0}^{x_1} + \int_{x_0}^{x_1} (x_1 - x)f''(x)dx \\ &= f(x_0) + f'(x_0)(x_1 - x_0) + \int_{x_0}^{x_1} f''(x)(x_1 - x)dx = P_1(x_1) + R_1(x_1). \end{aligned}$$

Análogamente, usando la igualdad

$$f''(x)(x_1 - x) = -\frac{d}{dx} \left(\frac{1}{2}f''(x)(x_1 - x)^2 \right) + \frac{1}{2}f'''(x)(x_1 - x)^2,$$

vemos que

$$\begin{aligned} f(x_1) &= P_2(x_1) + R_2(x_1) = f(x_0) + f'(x_0)(x_1 - x_0) + \frac{1}{2}(x_1 - x_0)^2 f''(x_0) \\ &\quad + \frac{1}{2} \int_{x_0}^{x_1} f'''(x)(x_1 - x)^2 dx. \end{aligned}$$

Inductivamente, podemos ver $f(x_1) = P_n(x_1) + r_n(x_1)$, donde

$$r_n(x_1) = \int_{x_0}^{x_1} \frac{1}{n!} f^{(n+1)}(x)(x_1 - x)^n dx.$$

Por el teorema de valor medio para integrales, existe $\xi \in [x_0, x_1]$ que verifica

$$f(x_1) = P_n(x_1) + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x_1 - x_0)^{n+1}.$$

El último término, se conoce como fórmula de Lagrange del resto $r_n(x) = f(x) - P_n(x)$. De esta representación obtenemos una acotación del error que se comete aproximando el valor de $f(x)$ por $P_n(x)$:

$$(1.1) \quad |f(x) - P_n(x)| \leq \frac{1}{(n+1)!} \max_{\xi \in [x_0, x]} |f^{(n+1)}(\xi)| (x - x_0)^{n+1}.$$

En la notación de Landau: $f(x) - P_n(x) = O((x - x_0)^{n+1})$, cuando $x \rightarrow x_0$. La desigualdad (1.1) nos muestra que si las derivadas están acotadas (como en el Ejemplo 1.1), $P_n(x) \rightarrow f(x)$ cuando $n \rightarrow \infty$.

En los siguientes ejemplos, analizamos el comportamiento del error para distintos valores de n .

Ejemplo 1.2 (función trigonométrica). Tomemos $f(x) = \sin(x)$ y $x_0 = 0$, las derivadas de orden par $f^{(2k)}(0)$ se anulan, por lo tanto los polinomios de orden par $n = 2k$ tienen grado impar $n - 1$ y vale $P_n(x) = P_{n-1}(x)$. Mostramos los primeros polinomios de grado impar:

$$\begin{aligned} P_1(x) &= x, \\ P_3(x) &= x - \frac{1}{6}x^3, \\ P_5(x) &= x - \frac{1}{6}x^3 + \frac{1}{120}x^5, \\ P_7(x) &= x - \frac{1}{6}x^3 + \frac{1}{120}x^5 - \frac{1}{5040}x^7, \\ P_9(x) &= x - \frac{1}{6}x^3 + \frac{1}{120}x^5 - \frac{1}{5040}x^7 + \frac{1}{362880}x^9, \\ P_{11}(x) &= x - \frac{1}{6}x^3 + \frac{1}{120}x^5 - \frac{1}{5040}x^7 + \frac{1}{362880}x^9 - \frac{1}{39916800}x^{11}. \end{aligned}$$

Tomando $x = 0.5$, tenemos $f(0.5) = 0.47942553860$. A continuación mostramos los errores al evaluar con los polinomios de Taylor de orden impar:

n	$P_n(x)$	$ f(x) - P_n(x) $
1	0.500000000000	0.206×10^{-1}
3	0.479166666667	0.259×10^{-3}
5	0.47942708333	0.154×10^{-5}
7	0.47942553323	0.537×10^{-8}
9	0.47942553862	0.122×10^{-10}

Como vemos en la Figura 1.1, el intervalo donde $P_n(x)$ es una buena aproximación de $f(x)$ aumenta con el orden del polinomio. Esto es consecuencia de la fórmula de error, siendo que todas las derivadas de $f(x)$ están acotadas por 1, para $x \in [-b, b]$ el error verifica

$$\max_{x \in [-b, b]} |r_n(x)| \leq \frac{b^n}{(n+1)!}$$

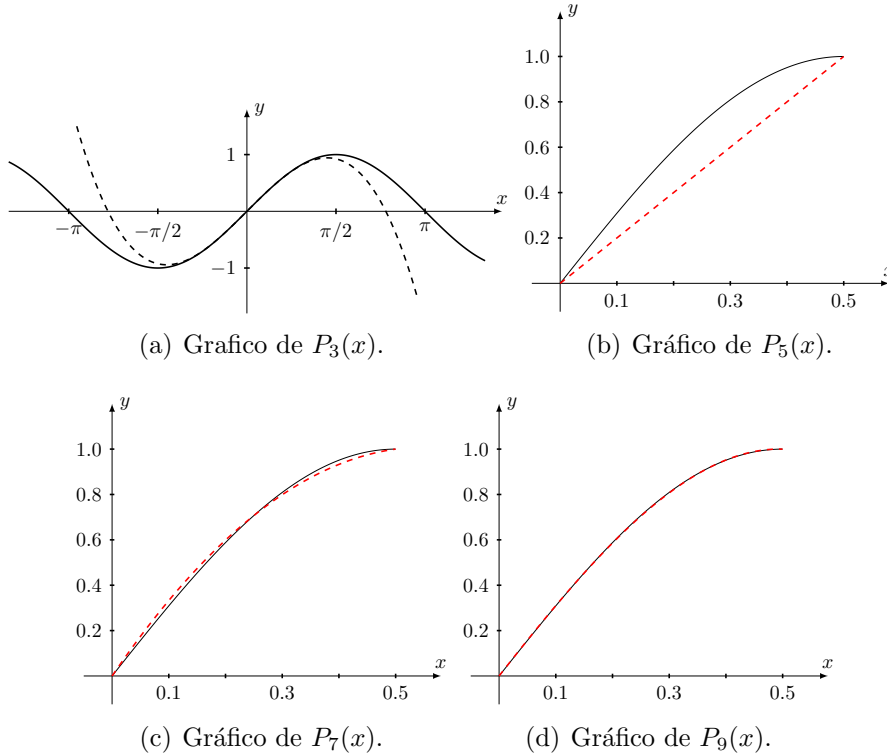


Fig. 1.1: Gráfico de $f(x) = \sin(x)$ y las aproximaciones de Taylor de orden $n = 3, 5, 7, 9$.

En los casos anteriores, la convergencia al verdadero valor es relativamente rápida, cuando aumentamos el orden del polinomio. En el siguiente ejemplo vemos que la convergencia de los polinomios de Taylor es muy lenta, lo que los vuelve ineficiente como método de aproximación de la función.

Ejemplo 1.3 (función logarítmica). Consideramos $f(x) = \ln(1+x)$ y $x_0 = 0$, podemos ver el polinomio de Taylor de orden n es

$$P_n(x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots - (-1)^n \frac{x^n}{n}$$

Una vez más consideramos las aproximaciones de $f(0.5) \cong 0.40546510811$ realizadas con los

sucesivos polinomios de Taylor

n	$P_n(x)$	$ f(x) - P_n(x) $
1	0.5000000	0.945×10^{-1}
2	0.3750000	0.305×10^{-1}
3	0.4166667	0.112×10^{-1}
4	0.4010417	0.442×10^{-2}
5	0.4072917	0.187×10^{-2}
6	0.4046875	0.778×10^{-3}
7	0.4058036	0.338×10^{-3}
8	0.4053153	0.150×10^{-3}
9	0.4055324	0.672×10^{-4}
10	0.4054346	0.305×10^{-4}

Podemos ver que, en este caso, la convergencia es mucho más lenta que en los ejemplos anteriores. La función tiene una asíntota vertical en $x = -1$, por lo que no podemos esperar convergencia de los polinomios en un intervalo que contenga a ese punto. Pero como observamos en la Figura 1.2), tampoco convergen bien cerca de $x = 1$. Esto se debe a que la región de convergencia de los polinomios de Taylor a la función siempre es un intervalo simétrico con centro x_0 , la cual no puede contener ninguna singularidad de f .

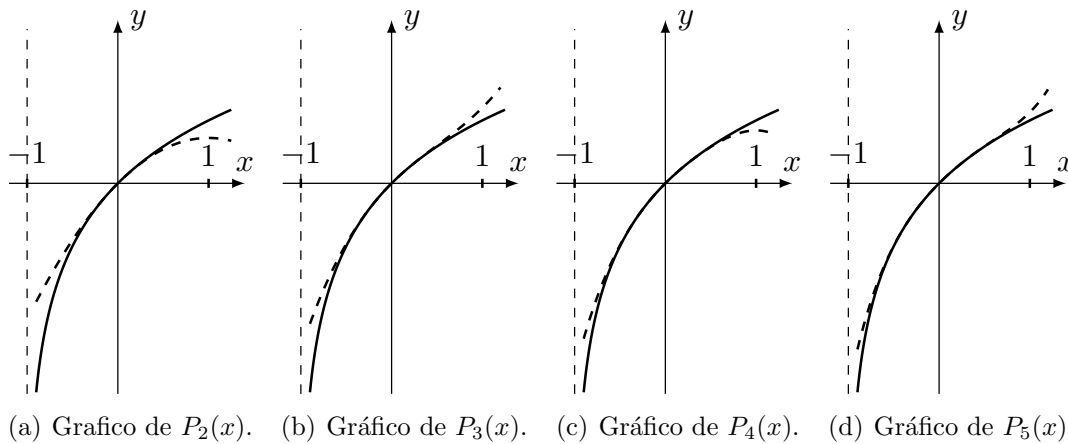


Fig. 1.2: Gráfico de $f(x) = \ln(1+x)$ y $P_n(x)$ con $n = 2, 3, 4, 5$.

Ejemplo 1.4 (función sech). En el ejemplo anterior claramente la función tiene un punto singular en $x = -1$. Esto produce que la convergencia de $P_n(x)$ a $f(x)$ sea lenta en los puntos cercanos a los extremos del intervalo simétrico $(-1, 1)$. Estudiemos el ejemplo $f(x) = \text{sech}(x)$ que está perfectamente definida en toda la recta real. Sin embargo, como vemos en la Figura 1.3 las aproximaciones de Taylor no convergen más allá de un intervalo simétrico alrededor de $x_0 = 0$. Para entender la razón de este comportamiento deberíamos considerar las extensiones al plano complejo de las funciones, lo que está fuera del alcance de este apunte. A modo de cierre del ejemplo, diremos que la función $f(x) = \text{sech}(x)$ tiene singularidades en los puntos $x = \pm i\pi/2$, lo cuales están a distancia $\pi/2$ del origen.

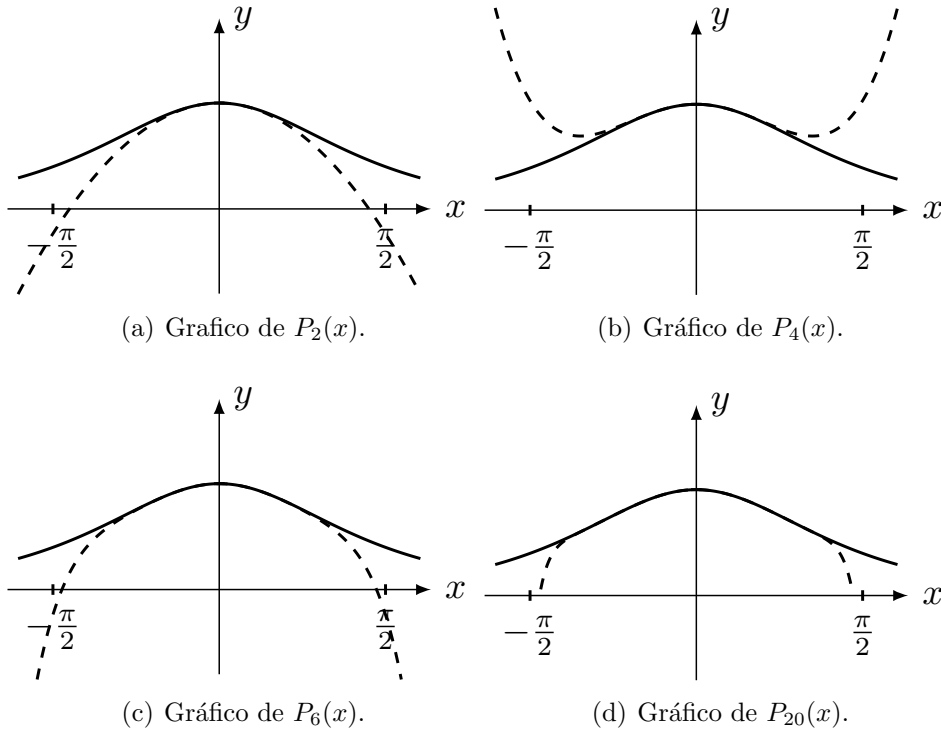


Fig. 1.3: Gráfico de $f(x) = \text{sech}(x)$ y $P_n(x)$ con $n = 2, 4, 6, 20$.

1.1.2. Aproximación de Padé*. La gran ventaja que tienen los polinomios sobre otra clase de funciones es que pueden ser evaluados a partir de las dos operaciones básicas de los números reales. En los ejemplos discutidos anteriormente, las evaluaciones de $P_n(x)$ pueden realizarse con lápiz y papel (y mucha paciencia). En la introducción planteamos que evaluar funciones racionales tiene complejidad similar a evaluar polinomios. Padé propuso aproximar una función por la mejor función racional, en el sentido de Taylor. Es decir $Q_{m,n}(x)$ es la aproximación de Padé de orden (m, n) de $f(x)$, siendo

$$Q_{m,n}(x) = \frac{a_0 + a_1(x - x_0) + \cdots + a_m(x - x_0)^m}{1 + b_1(x - x_0) + \cdots + b_n(x - x_0)^n}$$

que verifica $f^{(k)}(x_0) = Q_{m,n}^{(k)}(x_0)$ para $k = 0, \dots, m+n$. Estas condiciones determinan en forma unívoca los coeficientes de $Q_{m,n}$. Como ejemplo, tomamos $f(x) = \ln(1+x)$ y $x_0 = 0$. Las aproximaciones de Padé, $Q_{m,n}$, se muestran en la tabla 1.1.

(m, n)	1	2	3
1	$\frac{2x}{x+2}$	$\frac{12x}{-x^2+6x+12}$	$\frac{24x}{x^3-2x^2+12x+24}$
2	$\frac{x(x+6)}{4x+6}$	$\frac{3x(x+2)}{x^2+6x+6}$	$-\frac{3x(19x+30)}{x^3-21x^2-102x-90}$
3	$-\frac{x(x^2-6x-24)}{6(3x+4)}$	$\frac{x(x^2+21x+30)}{9x^2+36x+30}$	$\frac{x(11x^2+60x+60)}{3(x^3+12x^2+30x+20)}$

Tabla 1.1: Aproximaciones de Padé $Q_{m,n}$ de $f(x) = \ln(1+x)$ en $x_0 = 0$.

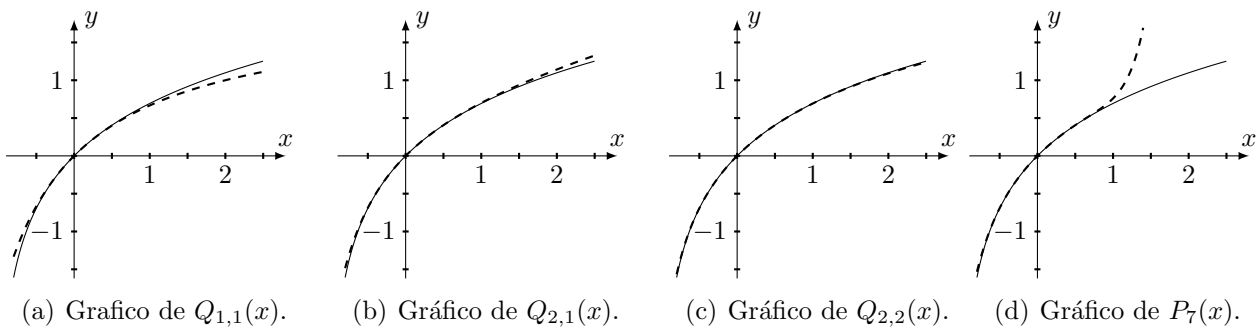
Los errores de calcular $\ln(1.5) \cong 0.47942553860$ mediante $Q_{m,n}$ se muestran en la tabla 1.2. En este ejemplo vemos que se logra una aproximación mejor con el mismo grado de complejidad.

(m, n)	1	2	3	4	5
1	0.547×10^{-2}	0.132×10^{-2}	0.402×10^{-3}	0.140×10^{-3}	0.526×10^{-4}
2	0.785×10^{-3}	0.597×10^{-4}	0.966×10^{-5}	0.207×10^{-5}	0.527×10^{-6}
3	0.162×10^{-3}	0.753×10^{-5}	0.627×10^{-6}	0.873×10^{-7}	0.155×10^{-7}
4	0.408×10^{-4}	0.127×10^{-5}	0.745×10^{-7}	0.651×10^{-8}	0.836×10^{-9}
5	0.117×10^{-4}	0.260×10^{-6}	0.113×10^{-7}	0.747×10^{-9}	0.671×10^{-10}

Tabla 1.2: Errores $|\ln(1.5) - Q_{m,n}(1.5)|$.

Por ejemplo, para evaluar $Q_{3,3}(x)$ tenemos que realizar $3 + 3$ multiplicaciones y 1 división, comparable en complejidad, a evaluar $P_7(x)$, pero el error obtenido es mucho menor.

En la Figura 1.4 se comparan los gráficos de $f(x) = \ln(1+x)$ con las aproximaciones de Padé $Q_{1,1}$, $Q_{2,1}$, $Q_{2,2}$ y el polinomio P_7 .

Fig. 1.4: Gráfico de $f(x) = \ln(1+x)$ y las aproximaciones de Padé.

1.2. Interpolación polinomial. En muchos problema científicos o en aplicaciones tecnológicas es frecuente disponer de un cierto número de puntos obtenidos por mediciones o cálculos auxiliares, y se busca construir una función que ajuste a estos datos. Este problema general, se va a tratar en las sucesivas secciones con distintos enfoques. Otro problema que se puede abordar con interpolación es de aproximar una función complicada por una más simple. Si tenemos una función cuyo cálculo resulta costoso, podemos partir de un cierto número de sus valores e interpolar dichos datos construyendo una función más simple. En general, obtendremos los valores de la función con un cierto error, pero dependiendo de las características del problema, la ganancia en simplicidad puede compensar el error cometido. La aproximación que se logra con polinomios de Taylor es local. Con esto queremos decir que es una muy buena aproximación cerca del punto donde se desarrolla, pero no lo es necesariamente en puntos alejados.

En esta sección vamos a estudiar la aproximación de funciones mediante polinomios en un intervalo acotado. El Teorema de Aproximación de Weierstrass¹ nos dice que las funciones reales continuas $f(x)$ definidas en un intervalo cerrado y acotado $[a, b]$ pueden ser aproximadas tanto como se quiera por una función polinomial $p(x)$, es decir

$$\max_{x \in [a, b]} |f(x) - p(x)| < \varepsilon.$$

Este resultado sólo afirma que la aproximación existe, pero no como hallarla. Nos proponemos construir polinomios que aproximen a $f(x)$ con un error apropiado al problema de interés.

¹No confundir con los muchos otros teoremas conocidos como Teorema de Weierstrass

Para esto vamos a construir lo que se conoce como un polinomio interpolador de Lagrange. La idea es simple, una función polinomial de grado n , $p(x) = a_0 + a_1x + \cdots + a_nx^n$, tiene $n + 1$ coeficientes, si conocemos el valor de la función $f(x)$ en $n + 1$ puntos, x_0, x_1, \dots, x_n , podemos determinar el valor de los coeficientes planteando $y_k = p(x_k) = f(x_k)$, o en forma equivalente, resolviendo el sistema de ecuaciones lineales:

$$(1.2) \quad \begin{aligned} y_0 &= a_0 + a_1x_0 + \cdots + a_nx_0^n, \\ y_1 &= a_0 + a_1x_1 + \cdots + a_nx_1^n, \\ &\vdots \\ y_n &= a_0 + a_1x_n + \cdots + a_nx_n^n. \end{aligned}$$

Se puede mostrar la existencia y unicidad del polinomio interpolador a partir del sistema lineal (1.2), dado que la matriz

$$V(x_0, \dots, x_n) = \begin{pmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^n \end{pmatrix},$$

llamada matriz de Vandermonde, siempre es inversible asumiendo que $x_j \neq x_k$ si $j \neq k$. Pero nosotros vamos a seguir un camino distinto para mostrar la existencia sin tener que resolver el sistema de ecuaciones.

Matriz de Vandermonde Queremos ver que la matriz de Vandermonde es inversible. Consideramos la función

$$w(x) = \det \begin{pmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x & \cdots & x^n \end{pmatrix},$$

desarrollando por la última fila vemos que $w(x)$ es un polinomio de grado menor o igual que n . Podemos ver que el coeficiente de x^n es

$$\begin{pmatrix} 1 & x_0 & \cdots & x_0^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n-1} & \cdots & x_{n-1}^{n-1} \end{pmatrix},$$

inductivamente vemos que es no nulo. Además, $w(x_0) = 0, \dots, w(x_{n-1}) = 0$, por lo tanto $w(x) \neq 0$ para todo $x \neq x_0, \dots, x_{n-1}$, en particular $w(x_n) \neq 0$. Esto muestra que $V(x_0, \dots, x_n)$ es inversible.

1.2.1. Forma de Lagrange. Dados los puntos x_0, \dots, x_n definimos los polinomios interpoladores de Lagrange $L_0(x), \dots, L_n(x)$ como

$$L_j(x) = \frac{(x - x_0) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_n)}{(x_j - x_0) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_n)} = \frac{\prod_{k \neq j} (x - x_k)}{\prod_{k \neq j} (x_j - x_k)},$$

es fácil ver que $L_j(x_j) = 1$ y $L_j(x_k) = 0$ si $k \neq j$. Ahora podemos construir $p_n(x)$ de la forma

$$p_n(x) = f(x_0)L_0(x) + \cdots + f(x_n)L_n(x) = \sum_{j=0}^n f(x_j)L_j(x).$$

Es claro que muchos problemas, por ejemplo cuando utilizamos datos adquiridos con anterioridad, no tenemos posibilidad de elegir los puntos x_0, \dots, x_n . Pero cuando podemos, surgen varias cuestiones:

- (I) ¿Cómo elegir los puntos x_0, \dots, x_n ?
- (II) ¿Existe una forma óptima de hacerlo?
- (III) ¿Si $n \rightarrow \infty$, $p_n(x) \rightarrow f(x)$?

Como $f(x)$ y $p_n(x)$ coinciden en x_0, \dots, x_n , parece razonable tomar puntos equidistante en el intervalos $[a, b]$, es decir $x_j = (1 - j/n)a + j/n b$. Veamos que ocurre con algunos ejemplos.

Ejemplo 1.5. Consideremos nuevamente $f(x) = \sin(\pi x)$ en el intervalo $[a, b] = [0, 0.5]$, para distintos valores de n mostramos las diferencias máximas entre la función y los polinomios interpoladores, tomando los puntos equidistantes:

n	máx $ f(x) - p_n(x) $
2	2.1051×10^{-1}
3	2.3537×10^{-2}
4	2.3932×10^{-3}
5	2.1533×10^{-4}
6	1.7105×10^{-5}
7	1.2085×10^{-6}
8	7.6645×10^{-8}
9	4.4015×10^{-9}
10	2.3075×10^{-10}

En la figura 1.5 se muestran los gráficos de $p_n(x)$ para distintos valores de n

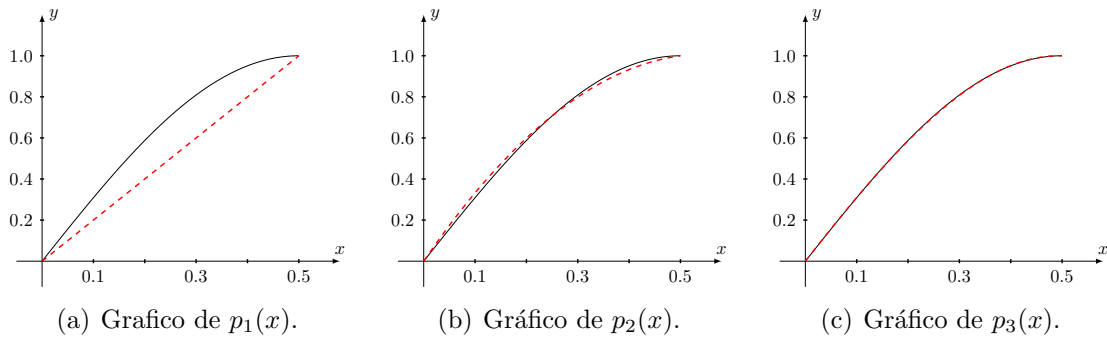


Fig. 1.5: Gráfico de $f(x) = \sin(\pi x)$ y las aproximaciones polinomiales.

Ejemplo 1.6 (función hiperbólica). Hagamos lo mismo que en el ejemplo anterior para $f(x) = \text{sech}(x)$ y el intervalo $[a, b] = [-5, 5]$, para distintos valores de n mostramos las diferencias máximas entre la función y los polinomios interpoladores:

n	máx $ f(x) - p_n(x) $
2	9.8652×10^{-1}
3	5.9306×10^{-1}
4	5.9135×10^{-1}
5	3.9335×10^{-1}
6	2.5965×10^{-1}
7	4.4200×10^{-1}
8	1.7085×10^{-1}
9	5.6791×10^{-1}
10	2.2243×10^{-1}
11	7.7654×10^{-1}

En la Figura 1.6 mostramos los gráficos de $p_n(x)$ para $n = 8, 9, 10$. Podemos observar el comportamiento oscilatorio de los polinomios interpolantes cerca de los extremos del intervalo.

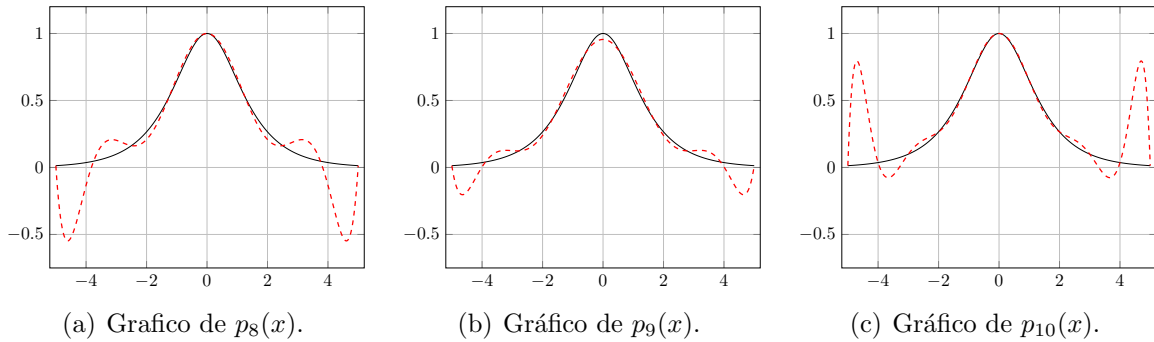


Fig. 1.6: Gráfico de $f(x) = \text{sech}(x)$ y las aproximaciones polinomiales.

Como muestra el ejemplo anterior, los polinomios interpolantes pueden ser una muy mala aproximación. Una posible solución consiste en distribuir en forma no uniforme los puntos dentro del intervalo. Si creemos que una mayor concentración de puntos mejora la aproximación, podemos espaciarlos en el medio del intervalo y acercarlos en las puntas. Esto se relaciona con la pregunta sobre como elegir en forma óptima los puntos x_0, \dots, x_n . La respuesta fue dada por Chebyshev y es la siguiente: supongamos que el intervalo es $[-1, 1]$, tomamos los puntos

$$x_0 = \cos(\theta), x_1 = \cos(3\theta), \dots, x_n = \cos((2n+1)\theta),$$

donde $\theta = \pi/(2n+2)$. Para el caso $n = 5$, mostramos los puntos en la Figura 1.7.

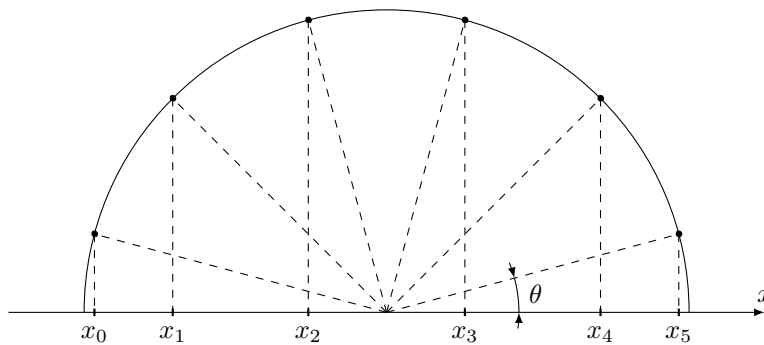
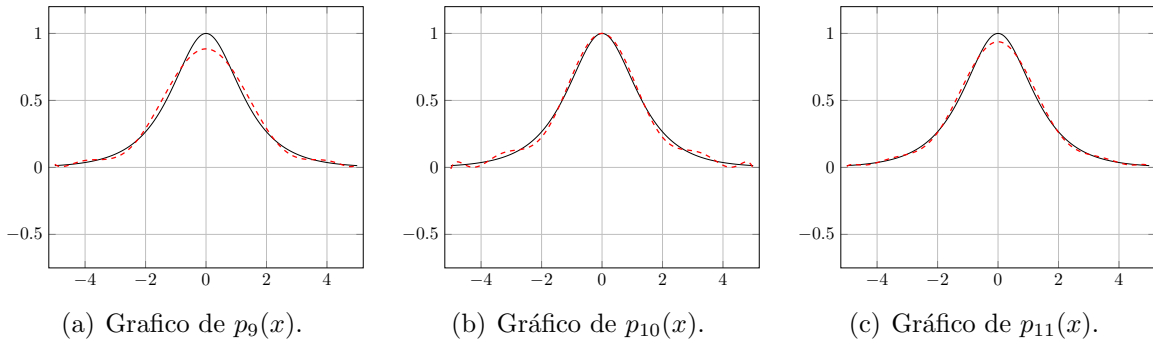


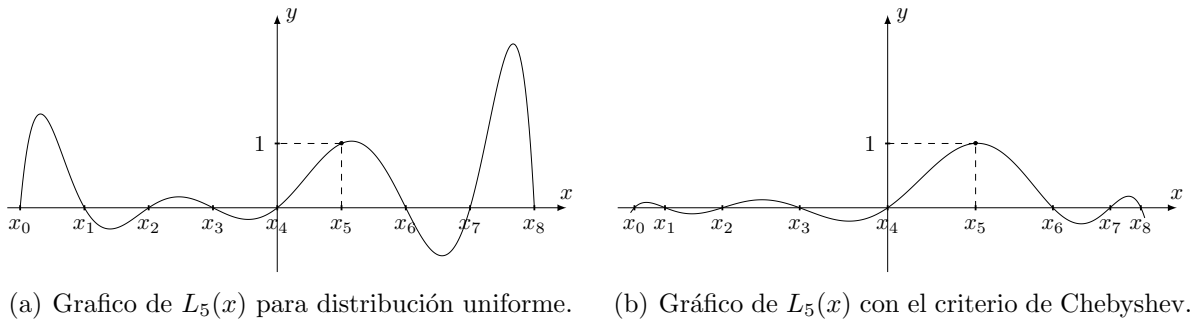
Fig. 1.7: Ubicación de los puntos de interpolación con el criterio de Chebyshev ($n = 5$).

Para un intervalo $[a, b]$ arbitrario, podemos considerar el cambio de variable $h(t) = (1-t)a/2 + (1+t)b/2$. Si $\tilde{f}(t) = f(h(t))$ y $\tilde{p}_n(t)$ el polinomio interpolador de $\tilde{f}(t)$, entonces $p_n(x) = \tilde{p}_n(h^{-1}(x))$.

Volviendo al Ejemplo 1.6, en la Figura 1.8 mostramos los polinomios que se obtienen con los puntos distribuidos con el criterio de Chebyshev. Vemos que las oscilaciones son mucho menores.

Fig. 1.8: Gráfico de $f(x) = \text{sech}(x)$ y las aproximaciones polinomiales.

Se puede ver la diferencia entre los polinomios $L_k(x)$ para la distribución uniforme y la distribución dada por el criterio de Chebyshev. En la Figura 1.9, mostramos $L_5(x)$ para el caso $n = 8$.

Fig. 1.9: Gráfico de $L_5(x)$ para distintas distribuciones con $n = 8$.

1.2.2. Forma de Newton. Vamos a considerar un procedimiento distinto para determinar los polinomios interpoladores. La idea es similar a la construcción de los polinomios de Taylor, a partir del polinomio de grado k , que verifica $p_k(x_0) = f(x_0), \dots, p_k(x_{k-1}) = f(x_{k-1})$, obtener el polinomio $p_{k+1}(x)$ sumándole al anterior un término de grado $k+1$ que no modifique lo anterior y además satisfaga $p_{k+1}(x_k) = f(x_k)$. Comencemos por el caso $n = 1$, la función lineal $p_1(x)$ que satisface $p_1(x_0) = f(x_0)$ y $p_1(x_1) = f(x_1)$ está dada por

$$p_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0).$$

Para que $p_2(x) = p_1(x) + \text{término cuadrático}$, verifique las condiciones anteriores, el término cuadrático debe anularse en x_0 y en x_1 , por lo tanto se escribe como $a_2(x - x_0)(x - x_1)$. De la condición $f(x_2) = p_2(x_2)$ obtenemos

$$f(x_2) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1),$$

por lo tanto a_2 queda determinado por

$$a_2 = \frac{\frac{f(x_2) - f(x_0)}{x_2 - x_0} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_1} = \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0}.$$

Si definimos las diferencias divididas

$$\begin{aligned} f[x_0] &= f(x_0), \\ f[x_0, x_1] &= \frac{f(x_1) - f(x_0)}{x_1 - x_0}, \\ f[x_0, x_1, x_2] &= \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}, \end{aligned}$$

tenemos que

$$(1.3) \quad p_2(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1).$$

Ejemplo 1.7. Queremos hallar la función cuadrática $p_2(x)$ interpoladora de la función $f(x)$ con $f(-1) = 1$, $f(1) = -5$ y $f(3) = 5$. Calculando las diferencias divididas obtenemos:

$$\begin{array}{l} f[-1] = \mathbf{1} \\ f[1] = -5 \\ f[3] = 5 \end{array} \quad \begin{array}{l} \rightarrow f[-1, 1] = \frac{-5 - 1}{1 - (-1)} = \mathbf{-3} \\ \rightarrow f[1, 3] = \frac{5 - (-5)}{3 - 1} = 5 \end{array} \quad \rightarrow f[-1, 1, 3] = \frac{5 - (-3)}{3 - (-1)} = \mathbf{2},$$

entonces $p_2(x) = 1 - 3(x + 1) + 2(x + 1)(x - 1) = 2x^2 - 3x - 4$.

La expresión anterior no es otra cosa que el método de Gauss–Jordan aplicado al sistema lineal (1.2). En efecto partiendo de este sistema de ecuaciones lineales y aplicando operaciones de fila, tenemos:

- Restamos la segunda fila a la tercera y la primera a la segunda:

$$\left(\begin{array}{ccc|c} 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & -5 \\ 1 & 3 & 9 & 5 \end{array} \right) \rightsquigarrow \left(\begin{array}{ccc|c} 1 & -1 & 1 & 1 \\ 0 & 2 & 0 & -6 \\ 0 & 2 & 8 & 10 \end{array} \right)$$

- Dividimos la segunda y tercer fila por 2

$$\left(\begin{array}{ccc|c} 1 & -1 & 1 & 1 \\ 0 & 2 & 0 & -6 \\ 0 & 4 & 8 & 10 \end{array} \right) \rightsquigarrow \left(\begin{array}{ccc|c} 1 & -1 & 1 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 1 & 4 & 5 \end{array} \right)$$

- Restamos la segunda fila a la tercer fila

$$\left(\begin{array}{ccc|c} 1 & -1 & 1 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 1 & 4 & 5 \end{array} \right) \rightsquigarrow \left(\begin{array}{ccc|c} 1 & -1 & 1 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 4 & 8 \end{array} \right)$$

- Dividimos la tercer fila por 4

$$\left(\begin{array}{ccc|c} 1 & -1 & 1 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 1 & 2 \end{array} \right) \rightsquigarrow \left(\begin{array}{ccc|c} 1 & -1 & 1 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 1 & 2 \end{array} \right)$$

por lo tanto $a_2 = 2$, $a_1 = -3$ y $a_0 = -4$. En forma general, tenemos

$$\begin{pmatrix} 1 & x_0 & x_0^2 & f[x_0] \\ 1 & x_1 & x_1^2 & f[x_1] \\ 1 & x_2 & x_2^2 & f[x_2] \end{pmatrix} \rightsquigarrow \begin{pmatrix} 1 & x_0 & x_0^2 & f[x_0] \\ 0 & x_1 - x_0 & x_1^2 - x_0^2 & f[x_1] - f[x_0] \\ 0 & x_2 - x_0 & x_2^2 - x_0^2 & f[x_2] - f[x_0] \end{pmatrix} \rightsquigarrow \\ \begin{pmatrix} 1 & x_0 & x_0^2 & f[x_0] \\ 0 & 1 & x_0 + x_1 & f[x_0, x_1] \\ 0 & 1 & x_0 + x_2 & f[x_0, x_2] \end{pmatrix} \rightsquigarrow \begin{pmatrix} 1 & x_0 & x_0^2 & f[x_0] \\ 0 & 1 & x_0 + x_1 & f[x_0, x_1] \\ 0 & 0 & 1 & f[x_0, x_1, x_2] \end{pmatrix},$$

de donde podemos despejar

$$\begin{aligned} a_2 &= f[x_0, x_1, x_2], \\ a_1 &= f[x_0, x_1] - (x_0 + x_1)f[x_0, x_1, x_2], \\ a_0 &= f[x_0] - x_0f[x_0, x_1] + x_0x_1f[x_0, x_1, x_2], \end{aligned}$$

y por lo tanto

$$p_2(x) = a_0 + a_1x + a_2x^2 = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1).$$

Esto se puede generalizar a cualquier número de puntos, como vemos a continuación.

1.2.3. Diferencias divididas. Vamos a definir las diferencias divididas en general en forma inductiva: $f[x_0] = f(x_0)$,

$$(1.4) \quad f[x_0, \dots, x_k] = \frac{f[x_1, \dots, x_k] - f[x_0, \dots, x_{k-1}]}{x_k - x_0}.$$

Se puede ver que el polinomio interpolador está dado por

$$(1.5) \quad p_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + \dots + f[x_0, \dots, x_n] \prod_{j=0}^{n-1} (x - x_j).$$

Las diferencias divididas se pueden obtener mediante el Algoritmo 1.1: x_list contiene la lista de puntos, x_0, \dots, x_n y f_list los valores $f(x_0), \dots, f(x_n)$. En df_list obtenemos las diferencias divididas $f[x_0], f[x_0, x_1], \dots, f[x_0, \dots, x_n]$.

Algoritmo 1.1: Diferencias divididas.

Data: n, x_list, f_list

Result: df_list

$df_list = f_list;$

for $k = 1$ **to** n **do**

for $j = n$ **to** k **do**

$df_list(j) =$

$(df_list(j) - df_list(j - 1)) / (x_list(j) - x_list(j - k));$

end

end

return $df_list;$

Para mostrar que $p_n(x_j) = f(x_j)$ para $j = 0, \dots, n$ usamos un argumento inductivo. Es claro que (1.5) es el polinomio interpolador para el caso $n = 1$. Supongamos que se verifica para $n - 1$ puntos, entonces los polinomios de orden $n - 1$ definidos como

$$\begin{aligned} p_{n-1}(x) &= f[x_0] + f[x_0, x_1](x - x_0) + \dots + f[x_0, \dots, x_{n-1}] \prod_{j=0}^{n-2} (x - x_j), \\ q_{n-1}(x) &= f[x_1] + f[x_1, x_2](x - x_1) + \dots + f[x_1, \dots, x_n] \prod_{j=1}^{n-1} (x - x_j), \end{aligned}$$

verifican $p_{n-1}(x_j) = f(x_j)$ para $j = 0, \dots, n-1$ y $q_{n-1}(x_k) = f(x_k)$ para $k = 1, \dots, n$. Siendo que

$$p_n(x) = p_{n-1}(x) + f[x_0, \dots, x_n] \prod_{j=0}^{n-1} (x - x_j),$$

tenemos que $p_n(x_j) = f(x_j)$ para $j = 0, \dots, n-1$. De la definición (1.4) tenemos

$$f[x_1, \dots, x_k] = f[x_0, \dots, x_{k-1}] + f[x_0, \dots, x_k](x_k - x_0),$$

por lo tanto

$$\begin{aligned} q_{n-1}(x_n) &= (f[x_0] + f[x_0, x_1](x_1 - x_0)) + (f[x_0, x_1] + f[x_0, x_1, x_2](x_2 - x_0))(x_n - x_1) \\ &\quad + (f[x_0, x_1, x_2] + f[x_0, x_1, x_2, x_3](x_3 - x_0))(x_n - x_1)(x_n - x_2) \\ &\quad + \dots + (f[x_0, \dots, x_{n-1}] + f[x_0, \dots, x_n](x_n - x_0)) \prod_{j=1}^{n-1} (x_n - x_j), \end{aligned}$$

reacomodando los términos tenemos

$$\begin{aligned} q_{n-1}(x_n) &= f[x_0] + f[x_0, x_1]((x_n - x_1) + (x_1 - x_0)) + f[x_0, x_1, x_2]((x_n - x_2) + (x_2 - x_0))(x_n - x_1) \\ &\quad + \dots + f[x_0, \dots, x_{n-1}](x_n - x_{n-1} + x_{n-1} - x_0) \prod_{j=1}^{n-2} (x_n - x_j) + f[x_0, \dots, x_n] \prod_{j=0}^{n-1} (x_n - x_j) \\ &= f[x_0] + f[x_0, x_1](x_n - x_0) + f[x_0, x_1, x_2](x_n - x_0)(x_n - x_1) \\ &\quad + \dots + f[x_0, \dots, x_{n-1}] \prod_{j=0}^{n-2} (x_n - x_j) + f[x_0, \dots, x_n] \prod_{j=0}^{n-1} (x_n - x_j) \end{aligned}$$

por lo tanto $p_n(x_n) = q_{n-1}(x_n) = f(x_n)$. Esto muestra que $p_n(x)$ es el polinomio interpolador.

El polinomio $p_n(x)$ se puede evaluar mediante la forma Horner:

$$\begin{aligned} p_n^{(0)}(x) &= f[x_0, \dots, x_n], \\ p_n^{(1)}(x) &= f[x_0, \dots, x_{n-1}] + (x - x_{n-1})p_n^{(0)}(x), \\ p_n^{(2)}(x) &= f[x_0, \dots, x_{n-2}] + (x - x_{n-2})p_n^{(1)}(x), \\ &\vdots \\ p_n(x) &= p_n^{(n)}(x) = f[x_0] + (x - x_0)p_n^{(n-1)}(x), \end{aligned}$$

que requiere n productos para su evaluación. Podemos ver el cálculo en el Algoritmo 1.2.

Algoritmo 1.2: Polinomio interpolador.

Data: n, df_list, x_list, x

Result: p

$p = df_list(n);$

for $k = n - 1$ **to** 0 **do**

$p = df_list(k) + (x - x_list(k)) * p;$

end

return $p;$

1.3. Fórmula de error de interpolación. En esta sección vamos a estudiar el error $r_n(x) = f(x) - p_n(x)$ que se comete al aproximar una función por el polinomio interpolador de Lagrange. La fórmula es similar a la obtenido para el polinomio de Taylor, para todo x en el intervalo, existe ξ tal que

$$r_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} W(x),$$

donde $W(x) = (x - x_0)(x - x_1) \dots (x - x_n)$. En la sección siguiente mostramos que ambos son casos particulares de un resultado más general.

Necesitamos una versión general del Teorema de Rolle. Recordemos el resultado original: si una función toma en dos puntos el mismo valor, $h(x_0) = h(x_1)$, entonces existe un punto $\xi \in (x_0, x_1)$ donde vale $h'(\xi) = 0$ (ver Figura 1.10).

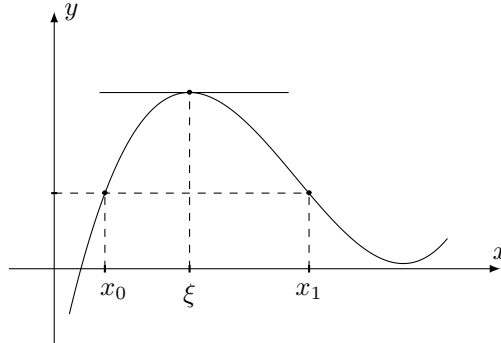


Fig. 1.10: Teorema de Rolle.

Supongamos ahora que existen $n + 1$ puntos $x_0 < x_1 < \dots < x_n$ para los cuales se verifica $h(x_0) = h(x_1) = \dots = h(x_n)$, afirmamos que existe un punto $\xi \in (x_0, x_1)$. Empezamos por un ejemplo, en la Figura 1.11 mostramos una función $h(x)$ que se anula en cuatro puntos, x_0, x_1, x_2, x_3 , en cada intervalo existe un punto ζ_k donde se anula $h'(x)$. Como $h'(\zeta_0) = h'(\zeta_1) = h'(\zeta_2)$ (que valgan cero es irrelevante), existen dos puntos, $\eta_0 \in (\zeta_0, \zeta_1)$ y $\eta_1 \in (\zeta_1, \zeta_2)$, donde se anula $h''(x)$. Aplicando por última vez el Teorema de Rolle vemos que existe ξ donde se verifica $h'''(\xi) = 0$.

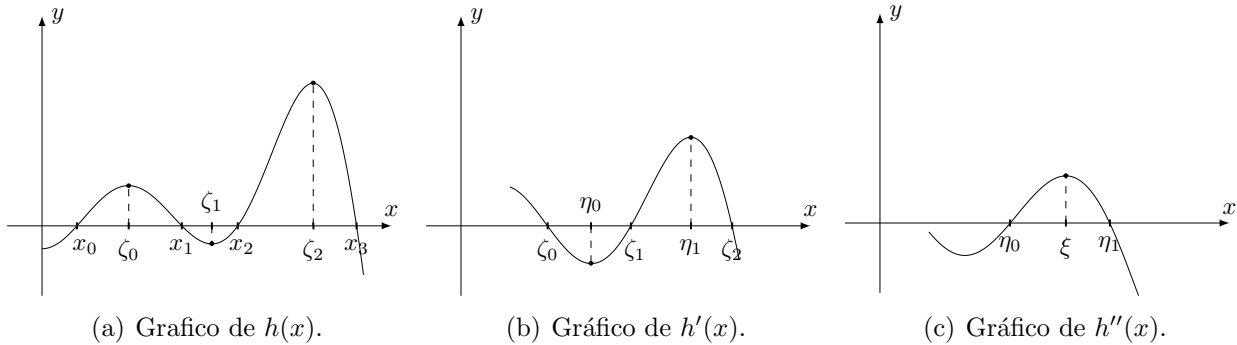


Fig. 1.11: Teorema de Rolle generalizado ($n = 3$).

La demostración general se basa en el principio de inducción. Por el Teorema de Rolle existe $\zeta_{k-1} \in (x_{k-1}, x_k)$ donde $h'(\zeta_{k-1}) = 0$, es decir que la función $h'(x)$ se anula en n puntos $\zeta_0, \dots, \zeta_{n-1}$. Aplicando en forma inductiva el resultado a la función derivada $h'(x)$ para n puntos, vemos que existe ξ tal que $0 = (h')^{(n-1)}(\xi) = h^{(n)}(\xi)$.

Ahora vamos a obtener cotas del error de las aproximaciones de interpolación. Dados $n + 1$ puntos en el intervalo (a, b) , $a \leq x_0 < x_1 < \dots < x_n \leq b$, definimos el polinomio de grado $n + 1$ que se anula en esos puntos: $W(x) = (x - x_0)(x - x_1) \dots (x - x_n)$. Fijado un punto $x \neq x_j$ ($j = 0, \dots, n$), consideramos la función auxiliar $h(t)$ dada por

$$h(t) = f(t) - p_n(t) - r_n(x) \frac{W(t)}{W(x)},$$

es fácil ver que $h(x_j) = 0$ y $h(x) = 0$. Como $h(t)$ se anula en $n + 2$ puntos, entonces existe $\xi \in [a, b]$ que verifica $h^{(n+1)}(\xi) = 0$. Por lo tanto

$$0 = f^{(n+1)}(\xi) - p_n^{(n+1)}(\xi) - r_n(x) \frac{W^{(n+1)}(\xi)}{W(x)}.$$

Ejemplo 1.8. Queremos hallar la función polinomial $p_5(x)$ interpoladora de la función $f(x)$ con $f(-1) = -3$, $f'(-1) = -1$, $f(0) = 0$, $f'(0) = 2$, $f''(0) = -6$ y $f(1) = -1$. Las diferencias divididas obtenemos:

x_k	$f(x_k)$
-1.	-3.
	-1.
-1.	-3.
	4.
	3.
	-5.
0.	0.
	-1.
	3.
	2.
	-2.
	-1.
0.	0.
	-3.
	1.
	2.
	0.
0.	0.
	-3.
	-1.
1.	-1.

$$\begin{aligned}
 p_5(x) &= -3 - (x+1) + 4(x+1)^2 - 5(x+1)^2x + 3(x+1)^2x^2 - (x+1)^2x^3 \\
 &= -3 + (x+1)(-1 + (x+1)(4 + x(5 + x(3 - x))))
 \end{aligned}$$

La información sobre la función se puede dar en la forma de una lista de listas $\{l_0, \dots, l_m\}$ donde $l_j = \{x_j, y_{j,0}, \dots, y_{j,q_j-1}\}$. Si armamos los vectores

$$\begin{aligned}
 x_list &= (x_0 \dots x_0 \ x_1 \dots x_1 \dots x_m \dots x_m) \\
 y_list &= (y_{0,0} \dots y_{0,0} \ y_{1,0} \dots y_{1,0} \dots y_{m,0} \dots y_{m,0})
 \end{aligned}$$

Algoritmo 1.3: Diferencias divididas.

Data: n, x_list, y_list

Result: df_list

$df_list = f_list;$

for $k = 1$ **to** n **do**

for $j = n$ **to** k **do**

if $x_list(j) == x_list(j - k)$ **then**

$df_list(j) = l(j, k);$

else

$df_list(j) = (df_list(j) - df_list(j - 1)) / (x(j) - x(j - k));$

end

end

return $df_list;$

1.5. Interpolación lineal y cúbica segmentada. Si queremos aproximar una función en un número grande de nodos, el polinomio interpolador resulta de grado muy alto. Esto presenta dos problemas: el costo de evaluar el polinomio y la inestabilidad numérica que presentan. Una alternativa es separar el intervalo $[a, b]$ en subintervalos y realizar la interpolación en cada uno de ellos. En cada subintervalo tendremos un polinomio interpolador diferente, pero al momento de evaluar, solo debemos decidir a que subintervalo pertenece el valor de x y evaluar la expresión correspondiente. Esto muestra que la complejidad no aumenta con el números de nodos.

1.5.1. Interpolación lineal. El problema de esta técnica es la falta de suavidad de la función partida interpolante. Lo mínimo que le pedimos es que sea continua, en ese caso basta considerar en cada intervalo funciones lineales. En forma más precisa, si $a = x_0 < \dots < x_n = b$, definimos $s_n(x)$ como la función partida

$$s_n(x) = \begin{cases} f(x_0) \frac{x_1 - x}{x_1 - x_0} + f(x_1) \frac{x - x_0}{x_1 - x_0} & , \text{ si } x \in [x_0, x_1], \\ \vdots & \vdots \\ f(x_{n-1}) \frac{x_n - x}{x_n - x_{n-1}} + f(x_n) \frac{x - x_{n-1}}{x_n - x_{n-1}} & , \text{ si } x \in [x_{n-1}, x_n]. \end{cases}$$

En la Figura 1.12 se grafica la función $f(x) = 1.5e^{-x^2/2} + 0.4 \cos(\pi x)$ y su aproximación mediante funciones lineales a trozos en el intervalo $[0, 3]$ con $n = 4$.

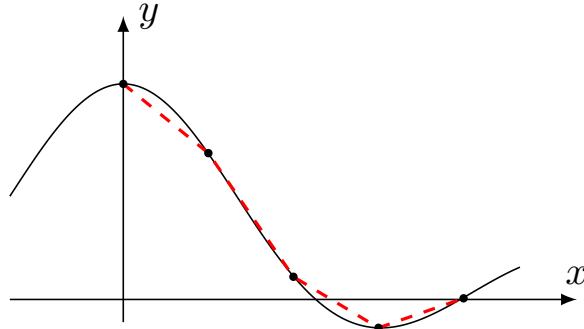


Fig. 1.12: Ajuste por segmentos lineales para $f(x) = 1.5e^{-x^2/2} + 0.4 \cos(\pi x)$ en $[0, 3]$.

Usando (1.6), podemos ver que para $x \in [x_{j-1}, x_j]$ se verifica

$$|f(x) - s_n(x)| \leq \frac{1}{2} \max_{\xi \in [a, b]} |f''(\xi)| (x_j - x)(x - x_{j-1}).$$

Si los puntos son equidistantes, $x_j - x_{j-1} = \frac{b-a}{n}$, entonces $(x_j - x)(x - x_{j-1}) \leq \frac{(b-a)^2}{4n^2}$ y por lo tanto

$$|f(x) - s_n(x)| \leq \max_{\xi \in [a, b]} |f''(\xi)| \frac{(b-a)^2}{8n^2}.$$

1.5.2. Interpolación cúbica. Si queremos que la función interpolante sea derivable, debemos aumentar el grado de los polinomios en cada subintervalo. Si por ejemplo, buscamos que sea dos veces derivable, podemos usar funciones polinomiales cúbicas. Para lograr esto, planteamos la continuidad de $s_n(x)$, $s'_n(x)$ y $s''_n(x)$ en los puntos x_k donde cambia la definición de $s_n(x)$:

$$\begin{aligned} f(x_0) &= s_n(x_0), \\ f(x_1) &= s_n(x_1^-) = s_n(x_1^+), \quad s'_n(x_1^-) = s'_n(x_1^+), \quad s''_n(x_1^-) = s''_n(x_1^+), \\ f(x_2) &= s_n(x_2^-) = s_n(x_2^+), \quad s'_n(x_2^-) = s'_n(x_2^+), \quad s''_n(x_2^-) = s''_n(x_2^+), \\ &\vdots \\ f(x_k) &= s_n(x_k^-) = s_n(x_k^+), \quad s'_n(x_k^-) = s'_n(x_k^+), \quad s''_n(x_k^-) = s''_n(x_k^+), \\ &\vdots \\ f(x_n) &= s_n(x_n). \end{aligned}$$

Como son n polinomios de cúbicos, el número de coeficientes indeterminados es $4n$. Por otro lado, de $s_n(x_k) = f(x_k)$ obtenemos $2n$ condiciones y de la continuidad de $s'_n(x)$ y $s''_n(x)$ resultan $2n-2$ condiciones más. Por lo tanto tenemos $4n-2$ condiciones y $4n$ coeficientes indeterminados, eso nos deja 2 condiciones libres. Existen distintas formas de elegir las, de acuerdo al objetivo buscado. Se puede considerar $s''_n(x_0) = s''_n(x_n) = 0$, conocidas como condiciones naturales. Por ejemplo, si f es periódica, buscamos que s_n también lo sea. Como $f(x_0) = f(x_n)$, se verifica $s_n(x_0) = s_n(x_n)$, entonces podemos tomar como condiciones adicionales: $s'_n(x_0) = s'_n(x_n)$ y $s''_n(x_0) = s''_n(x_n)$. Se puede probar que la solución de este sistema de ecuaciones siempre existe.

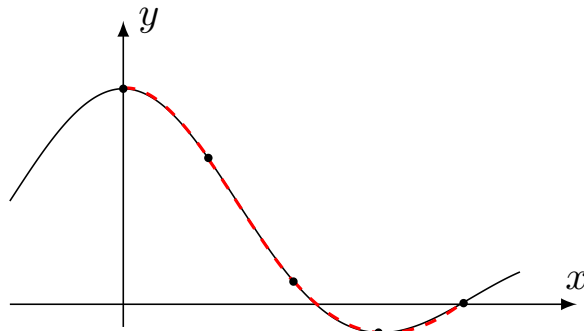


Fig. 1.13: Ajuste por segmentos cúbicos para $f(x) = 1.5e^{-x^2/2} + 0.4\cos(\pi x)$ en $[0, 3]$.

1.6. Aplicación.

1.6.1. Temperatura ambiente. En el artículo [6], Hubert Frings y Mable Frings estudiaron la influencia de la temperatura sobre el número de chirridos por minuto de grillos (*neoconocephalus ensiger*) machos. En la Tabla 1.3 se muestran los siguientes valores (Tabla 1 en [6]). En la Figura 1.14 se muestra el ajuste hecho con segmentos cúbicos.

T ($^{\circ}\text{C}$)	8	9	14	17	18	19	20.5	21.5	23	24	25	26
N (min^{-1})	264	285	346	417	438	495	524	540	643	693	744	780

Tabla 1.3: Número medio de chirridos/min. para individuos machos (n. ensiger).

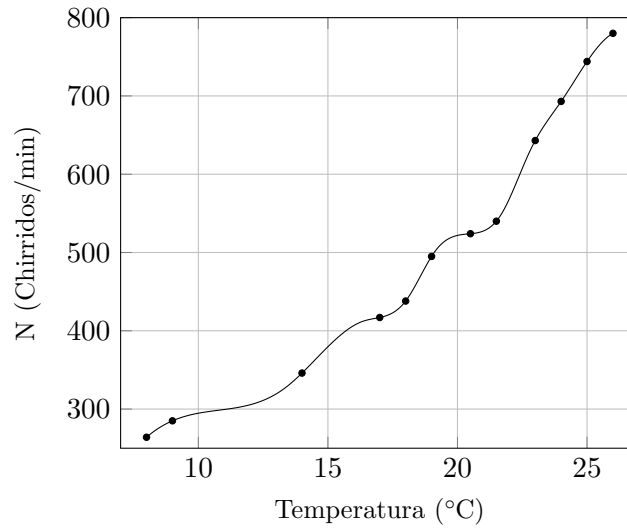


Fig. 1.14: Ajuste por segmentos cúbicos del número de chirridos/min en función de la temperatura.

1.6.2. Integración. Como aplicación de interpolación segmentada, podemos estudiar métodos de cuadratura, es decir cálculos aproximados del valor de la integral de f en el intervalo a $[a, b]$, tomando la integral de la función aproximante. Si consideramos la aproximación lineal a trozos obtenemos

$$\begin{aligned} I &= \int_a^b f(x)dx \cong \int_a^b s_n(x)dx = \int_{x_0}^{x_1} s_n(x)dx + \cdots + \int_{x_{n-1}}^{x_n} s_n(x)dx \\ &= \frac{f(x_0) + f(x_1)}{2}(x_1 - x_0) + \cdots + \frac{f(x_{n-1}) + f(x_n)}{2}(x_n - x_{n-1}). \end{aligned}$$

Si suponemos $x_j - x_{j-1} = (b - a)/n$, nos queda el método de los trapecios

$$I \cong T_n = \frac{b - a}{2n}(f(x_0) + 2f(x_1) + \cdots + 2f(x_{n-1}) + f(x_n)).$$

Usando la acotación del error se obtiene

$$(1.7) \quad I - T_n = \frac{f''(\xi_1)}{2} \int_{x_0}^{x_1} (x - x_1)(x - x_0)dx + \cdots + \frac{f''(\xi_n)}{2} \int_{x_{n-1}}^{x_n} (x - x_n)(x - x_{n-1})dx,$$

de donde obtenemos

$$|I - T_n| \leq \frac{\max |f''(\xi)|}{12} \frac{(b - a)^3}{n^2}.$$

Ejemplo 1.9. Tomemos el problema de integrar $f(x) = e^{-x^2/2}$ en $[0, 3]$, con $I \cong 1.24993044474$

n	T_n	$ I - T_n $	$n^2(I - T_n)$
2	1.2453104484	0.46200×10^{-2}	1.8480×10^{-2}
4	1.2484545573	0.14759×10^{-2}	2.3614×10^{-2}
8	1.2495453664	0.38508×10^{-3}	2.4645×10^{-2}
16	1.2498331500	0.97295×10^{-4}	2.4907×10^{-2}
32	1.2499060568	0.24388×10^{-4}	2.4973×10^{-2}
64	1.2499243437	0.61010×10^{-5}	2.4990×10^{-2}
128	1.2499289192	0.15255×10^{-5}	2.4994×10^{-2}
256	1.2499300633	0.38140×10^{-6}	2.4995×10^{-2}

Observemos que $n^2(I - T_n)$ converge a una constante C , esto permite dos cosas importantes. La primera es una estimación del error, tengamos en cuenta no conocemos I , pero usando

$$I - T_n \cong \frac{C}{n^2},$$

$$I - T_{2n} \cong \frac{C}{4n^2},$$

podemos despejar $I - T_n \cong \frac{4}{3}(T_{2n} - T_n)$. Con esta estimación, podemos saber si el error está en valores aceptables, o si por el contrario, debemos refinar nuestra partición. Inclusive podemos hacerlo en algunos subintervalos, aquellos donde la función presente mayores dificultades (comportamiento oscilatorio, singularidades, etc.) Por otro lado, de las estimaciones anteriores podemos despejar $I \cong r_n = \frac{4}{3}T_{2n} - \frac{1}{3}T_n$. Este método se conoce como extrapolación de Richardson. Para este ejemplo obtenemos

n	r_n	$ I - r_n $	$n^4 I - r_n $
2	1.24950259352	0.42785×10^{-3}	6.84562×10^{-3}
4	1.24990896941	0.21475×10^{-4}	5.49769×10^{-3}
8	1.24992907793	0.13668×10^{-5}	5.59845×10^{-3}
16	1.24993035902	0.85723×10^{-7}	5.61793×10^{-3}
32	1.24993043938	0.53620×10^{-8}	5.62245×10^{-3}
64	1.24993044441	0.33519×10^{-9}	5.62359×10^{-3}
128	1.24993044472	0.20949×10^{-10}	5.62418×10^{-3}

Se puede probar que

$$\lim_{n \rightarrow \infty} n^2(I - T_n) = -\frac{(b-a)^2}{12} \int_a^b f''(x)dx.$$

En efecto, como

$$\int_{x_{j-1}}^{x_j} (x - x_j)(x - x_{j-1})dx = \frac{(x_j - x_{j-1})^3}{6} = -\frac{(b-a)^2}{6n^2}(x_j - x_{j-1}),$$

usando (1.7) obtenemos

$$n^2(I - T_n) = -\frac{(b-a)^2}{12} (f''(\xi_1)(x_1 - x_0) + \cdots + f''(\xi_n)(x_n - x_{n-1})),$$

interpretando el lado derecho como una suma de Riemann, obtenemos el resultado. En el ejemplo anterior tenemos

$$-\frac{(b-a)^2}{12} \int_a^b f''(x)dx = \frac{3}{4} \int_0^3 e^{-x^2/2}(1-x^2)dx \cong 0.024995.$$

1.7. Ejercicios.

▣ **Ejercicio 1.1.** Dada la función $f(x) = \tanh(x)$.

- Mostrar las diferencias $f(x) - P_6(x)$ y $(f(x) - P_6(x))/x^7$, con $P_6(x) = x - x^3/3 + 2x^5/15$ polinomio de Talor de orden $n = 6$ en $x_0 = 0$.
- Repetir reemplazando $P_6(x)$ por la aproximación de Padé $P_{3,2}(x)$ dada por

$$P_{3,2}(x) = \frac{x + x^3/15}{1 + 2x^2/5}.$$

(c) Graficar las funciones $f(x)$, $P_6(x)$ y $P_{3,2}(x)$ en el intervalo $[-2, 2]$.

▢ **Ejercicio 1.2.** Calcular las diferencias divididas y el polinomio interpolador para $f(x)$ en los puntos x_0, \dots, x_n :

(a) $f(x) = \sqrt{x}$, $x_0 = 4.0, x_1 = 5.0, x_2 = 6.0, x_3 = 7.0, x_4 = 8.0$.

(b) $f(x) = e^{-x}$, $x_0 = 0.0, x_1 = 1.0, x_2 = 2.0, x_3 = 3.0, x_4 = 4.0$.

🔗 **Ejercicio 1.3*** Si $f(x)$ es una función infinitamente derivable, probar que son equivalentes las siguientes afirmaciones:

I. $f(x)$ es un polinomio de grado n .

II. $f[x_0, x_1, \dots, x_n]$ es constante para todo $x_0, x_1, \dots, x_n \in \mathbb{R}$.

III. $f[x_0, x_1, \dots, x_{n+1}] = 0$ para todo $x_0, x_1, \dots, x_{n+1} \in \mathbb{R}$.

▢ **Ejercicio 1.4.** Construir el polinomio interpolador en el intervalo $[8, 26]$, correspondiente a los datos de la Tabla 1.3. Graficar.

▢ **Ejercicio 1.5.** Graficar los polinomios de Lagrange en el intervalo $[-1, 1]$: $L_0(x), \dots, L_8(x)$, en los casos

(a) $-1 = x_0 < \dots < x_8 = 1$ equidistantes ($x_j = j/4 - 1$).

(b) x_j elegidos con el criterio de Chebyshev ($x_j = \cos((j + 1/2)\pi/9)$).

(c) x_j elegidos al azar en el intervalo $[-1, 1]$.

🔗 **Ejercicio 1.6*** Si $f(x)$ es una función infinitamente derivable,

$$f[x_0, x_0 + h, \dots, x_0 + kh] = \frac{f^{(k)}(x_0)}{k!} + o(h)$$

🔗 **Ejercicio 1.7.** Obtener la expresión de r_n como suma pesada de las evaluaciones de f , es decir $r_n = \beta_0 f(x_0) + \dots + \beta_{2n} f(x_{2n})$. Comparar con la regla de Simpson.

🔗 **Ejercicio 1.8.** Calcular $\lim_{n \rightarrow \infty} n^4(I - r_n)$.

CAPÍTULO 2

Resolución de Ecuaciones no Lineales

”Desde los seis años sentí el impulso de dibujar las formas de las cosas. Hacia los cincuenta, expuse una colección de dibujos; pero nada de lo ejecutado antes de los setenta me satisface. Sólo a los setenta y tres años pude intuir, siquiera aproximadamente la verdadera forma y naturaleza de las aves, peces y plantas. Por consiguiente, a los ochenta años habré hecho grandes progresos; a los noventa habré penetrado la esencia de todas las cosas; a los cien, habré seguramente ascendido a un estado más alto, indescriptible, y si llego a ciento diez años, todo, cada punto y cada línea, vivirá.”

Katsushika Hokusai

2.1. Método de bisección. El teorema de Bolzano nos dice que una función continua en un intervalo que cambia de signo, es decir que toma valores positivos y negativos, tiene que tener puntos donde se anula. En general, si toma dos valores reales y_1, y_2 , debe tomar todos los valores intermedios. Como muchos resultados de la matemática, el teorema de Bolzano es un caso típico de teorema de existencia: asegura que lo que buscamos existe, en este caso una solución de la ecuación, pero no dice si es único, ni tampoco como encontrarlo. El método de bisección trata de hallar una solución aproximada basado en este teorema. La idea es sencilla: partimos de un intervalo $[a_0, b_0]$ tal que las evaluaciones en los extremos tengan distinto signo, es decir $f(a_0)f(b_0) < 0$, tomamos un punto intermedio $c_0 \in (a_0, b_0)$. Si $f(c_0) = 0$, entonces $x_* = c_0$ es una solución. Si $f(c_0)f(a_0) > 0$, definimos un nuevo intervalo $[a_1, b_1] = [c_0, b_0]$. En el caso $f(c_0)f(a_0) < 0$, tomamos el intervalo $[a_1, b_1] = [a_0, c_0]$. Tomemos como ejemplo la ecuación $f(x) = 0$ para la función $f(x) = 1.75x^3 - 3x - 1$. En el intervalo $[-2, 2]$ se verifica $f(-2) < 0 < f(2)$. Más aún, están las tres raíces del polinomio $f(x)$, $x = -1.09114, -0.360711, 1.45185$. Las aplicaciones del método de bisección de muestran en la Tabla 2.1. Los cuatro primeros pasos se muestran en la Figura 2.1. Observemos que no tenemos control a cual de la soluciones converge el algoritmo.

2.1.1. Algoritmo. El Algoritmo 2.1 realiza la búsqueda del cero de $f(x)$ a partir de un intervalo donde hay un cambio de signo. En cada paso, la longitud del intervalo se reduce a la mitad manteniendo el cambio de signo. Después de n pasos, el programa devuelve el último intervalo y el punto medio.

n	a_n	b_n	c_n
0	-2.00000	2.00000	0.00000
1	0.00000	2.00000	1.00000
2	1.00000	2.00000	1.50000
3	1.00000	1.50000	1.25000
4	1.25000	1.50000	1.37500
5	1.37500	1.50000	1.43750
6	1.43750	1.50000	1.46875
7	1.43750	1.46875	1.45313
8	1.43750	1.45313	1.44531
9	1.44531	1.45313	1.44922
10	1.44922	1.45313	1.45117
11	1.45117	1.45313	1.45215
12	1.45117	1.45215	1.45166

Tabla 2.1: Iteraciones del método de bisección para $f(x) = 1.75x^3 - 3x - 1$.

Algoritmo 2.1: Método de bisección.

Data: f, a, b, n
Result: a, b, c
 $y_a = f(a);$
 $y_b = f(b);$
for $i = 0$ **to** n **do**
 $c = 0.5 * (a + b);$
 print $i, a, b, c;$
 $y_c = f(c);$
 if $y_c == 0$ **then**
 $a = c;$
 $b = c;$
 else if $y_a * y_c > 0$ **then**
 $a = c;$
 $y_a = y_c;$
 else
 $b = c;$
 $y_b = y_c;$
 end
end
return $a, b, c;$

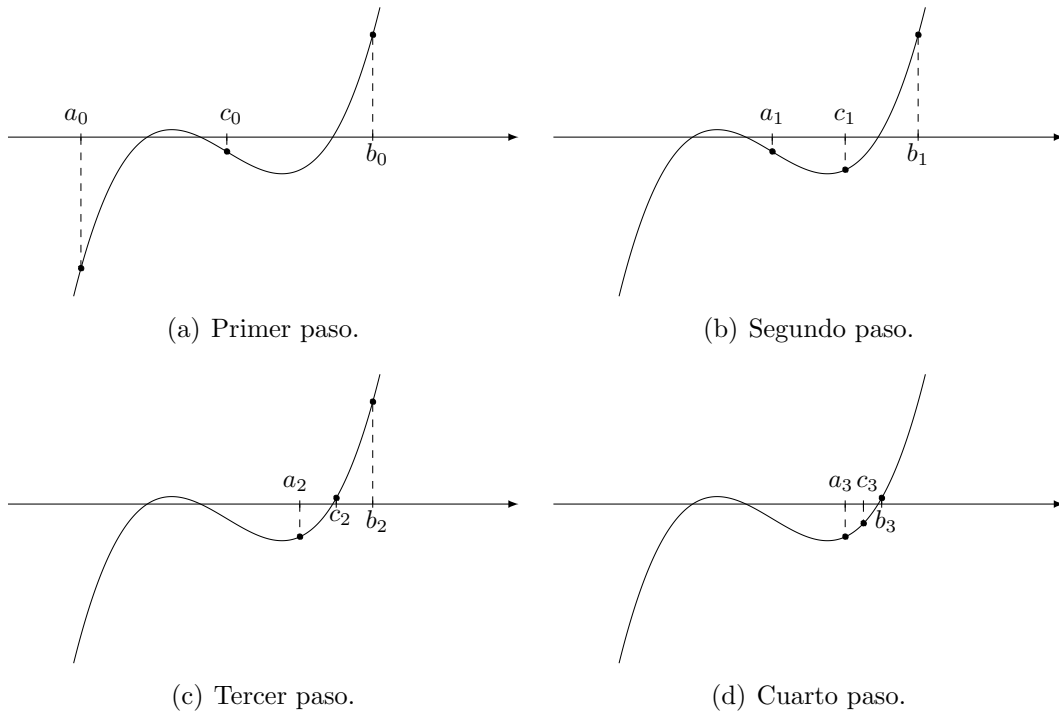


Fig. 2.1: Método de bisección para la función $f(x) = 1.75x^3 - 3x - 1$.

2.1.2. Fórmula del error. Es fácil ver que la longitud del intervalo se reduce a la mitad en cada paso, por lo que vemos que $b_n - a_n = 2^{-n}(b_0 - a_0)$, es decir $b_n - a_n \cong 10^{-3}(b_0 - a_0)$ para $n = 10$. Cada paso requiere una evaluación de la función.

2.2. Método de Newton. El método de Newton es un procedimiento iterativo que permite hallar la solución de la ecuación $f(x) = 0$. El método consiste en aproximar la función $f(x)$ por la recta tangente en un punto inicial y buscar los ceros de la ecuación aproximada, $f(x_0) + f'(x_0)(x - x_0) = 0$. El punto hallado x_1 reemplaza al punto anterior y se comienza nuevamente. Se obtiene una sucesión x_0, x_1, x_2, \dots que, en caso de ser convergente, tiene como límite una solución del problema. Concretamente, a partir de x_0 se construye recursivamente la sucesión

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}.$$

Observemos que se necesita evaluar la función y su derivada, esto puede ser un problema cuando $f(x)$ no tiene una expresión conocida, sino que es el resultado de otros cálculos. La convergencia no está garantizada (ver Figura 2.4), pero es fácil ver bajo condiciones si la sucesión converge a un punto x_* , entonces $f(x_*) = 0$. En efecto, como $f(x_{n-1}) = -f'(x_{n-1})(x_n - x_{n-1})$, si $x_n \rightarrow x_*$ entonces $x_n - x_{n-1} \rightarrow 0$ y $f'(x_{n-1}) \rightarrow f'(x_*)$, por lo tanto $f(x_{n-1}) \rightarrow 0$.

2.2.1. Interpretación geométrica. Para ilustrar el método vamos a considerar diferentes ejemplos. Estudiamos primero la ecuación $e^x = 45$ cuya solución es $x = \ln 45 \cong 3.8066624898$. El problema es equivalente a la ecuación $f(x) = 0$, donde $f(x) = e^x - 45$, la sucesión se obtiene recursivamente de la forma

$$x_n = x_{n-1} - \frac{e^{x_{n-1}} - 45}{e^{x_{n-1}}} = x_{n-1} - 1 + 45 e^{-x_{n-1}}$$

En la Tabla 2.2 mostramos las iteraciones partiendo de $x_0 = 6$. En la segunda columna vemos como disminuyen los errores, primero lentamente y después en forma acelerada. Eso se refleja

n	x_n	ϵ_n	$\epsilon_n/\epsilon_{n-1}^2$
0	6.0000000000	-2.193	
1	5.1115438479	-1.305	-0.271
2	4.3827485573	-0.576	-0.338
3	3.9448426224	-0.138	-0.416
4	3.8157844150	-0.912×10^{-2}	-0.478
5	3.8067039683	-0.415×10^{-4}	-0.498
6	3.8066624906	-0.860×10^{-9}	-0.500

Tabla 2.2: Iteraciones del método de Newton para $e^x = 45$.

en la tercera columna donde se compara ϵ_n con ϵ_{n-1}^2 . En la Figura 2.2 mostramos los primeros puntos y las rectas tangentes

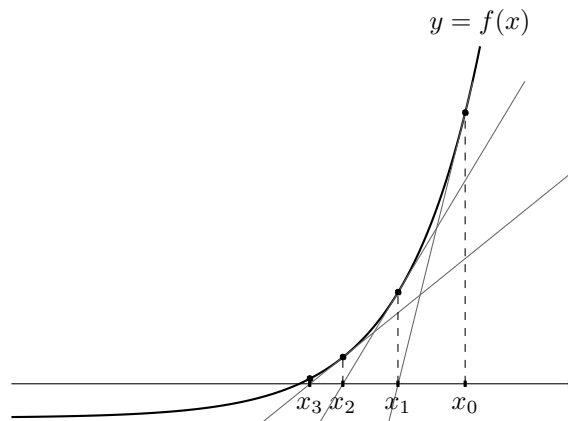


Fig. 2.2: Gráfico de $f(x) = e^x - 45$ y las iteraciones del método de Newton.

2.2.2. Algoritmo de Newton. En el código 2.2 tenemos el método de Newton. El programa llama a las rutinas que evalúan $f(x)$ y $f'(x)$. Devuelve el valor calculado de x y un código de error: code que toma el valor cero si el programa termina correctamente. Se consideran dos criterios de parada: cuando dos iteraciones distan menos que δ , lo que podría interpretarse como convergencia a una solución, o cuando el número de iteraciones supera un valor máximo $N_{\text{máx}}$, en ese caso se considera que el método no alcanza un resultado, posiblemente porque el método no es convergente.

Algoritmo 2.2: Método de Newton.

Input: $f, f', x_0, \delta, N_{\max}$
Output: x , code
 $n = 1$;
 $x_{\text{old}} = x_0$;
 $x_{\text{new}} = x_{\text{old}} - f(x_{\text{old}})/f'(x_{\text{old}})$;
while $|x_{\text{new}} - x_{\text{old}}| > \delta \wedge n \leq N_{\max}$ **do**
 $n = n + 1$;
 $x_{\text{old}} = x_{\text{new}}$;
 $x_{\text{new}} = x_{\text{old}} - f(x_{\text{old}})/f'(x_{\text{old}})$;
end
if $n \leq N_{\max}$ **then**
 code = 0;
else
 code = 1;
end
return x_{new} , code;

Observemos que estos controles no pueden considerarse válidos en todos los casos. Por ejemplo, la ecuación $f(x) = 2 + x^3 + \tanh(a(x - 1)) = 0$ con $a = 5 \times 10^3$, tiene solución $x_* \cong -1$, pero partiendo del punto $x_0 = 4.6149$ obtenemos $x_1 = 3.02965, x_2 = 1.91082, x_3 = 0.999997, x_4 = 0.9994$. Vemos que $|x_4 - x_3| < 6 \times 10^{-4}$. Por lo tanto, si $\delta = 10^{-3}$, el programa da como resultado $x = x_4$, muy lejos del resultado correcto. Por otro lado, si $x_0 = 0.001$ podemos ver que $|x_n - x_{n-1}| < 10^{-3}$ recién para $n = 36$, con $x_{35} = -1.00002$ y $x_{36} \cong -1$. Si $N_{\max} < 36$, el programa entregaría un valor de x incorrecto.

Ejemplo 2.1. La ley de Planck establece la radiación electromagnética emitida por un cuerpo negro en equilibrio térmico a una temperatura T definida. La densidad de radiación en la longitud de onda λ está dada por

$$v(\lambda, T) = \frac{8\pi hc}{\lambda^5 \left(\exp\left(\frac{hc}{kT\lambda}\right) - 1 \right)}$$

donde $h = 6.626 \times 10^{-34}$ J s es la constante de Planck, $c = 2.998 \times 10^8$ m s⁻¹ es la velocidad de la luz en el vacío y $k = 1.381 \times 10^{-23}$ J K⁻¹ es la constante de Boltzmann. En la Figura 2.3 se grafica la densidad $v(\lambda)$ (en kJ m⁻³ nm⁻¹) en función λ (en nm). La longitud de onda donde la densidad es máxima se obtiene resolviendo

$$\frac{\partial v}{\partial \lambda}(\lambda, T) = \frac{8\pi h c e^x}{\lambda^6 (e^x - 1)^2} (x + 5e^{-x} - 5) = 0,$$

con $x = hc/(kT\lambda)$. Para hallar la solución de la ecuación $x + 5e^{-x} - 5 = 0$, con $x > 0$, usamos el método de Newton

$$x_n = 5 \frac{e^{x_{n-1}} - 1 - x_{n-1}}{e^{x_{n-1}} - 1}$$

obteniendo la solución $x = 4.96511$. Por lo tanto vale la relación

$$\lambda_{\max} = \frac{hc}{4.96511kT} = \frac{0.289805 \times 10^{-2} \text{ K m}}{T},$$

conocida como ley de Wien. A la temperatura del sol, $T \cong 6000$ K, $\lambda_{\max} = 483$ nm.

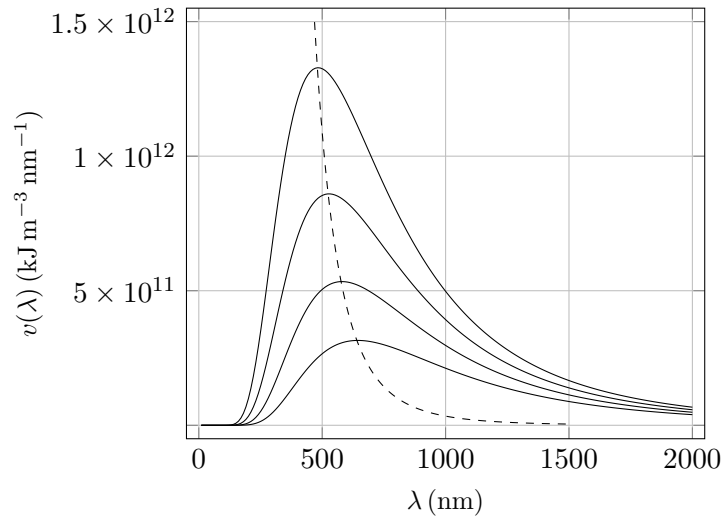


Fig. 2.3: Gráfico de $v(\lambda)$, $T = 4500, 5000, 5500, 6000$.
En línea de puntos mostramos λ_{\max} para cada T .

2.2.3. Fórmula del error para el método de Newton. Vamos a ver como evoluciona el error $\epsilon_n = x_* - x_n$. Si tomamos el polinomio de Taylor de segundo orden de f alrededor de x_n , tenemos

$$0 = f(x_*) = f(x_{n-1}) + f'(x_{n-1})\epsilon_{n-1} + \frac{f''(\xi_{n-1})}{2}\epsilon_{n-1}^2,$$

para algún ξ_{n-1} en el intervalo de extremos x_*, x_{n-1} . Como $0 = f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1})$, usando que $\epsilon_{n-1} = x_n - x_{n-1}$ obtenemos $0 = f'(x_{n-1})\epsilon_n + \frac{1}{2}f''(\xi_{n-1})\epsilon_{n-1}^2$, asumiendo que $f'(x_{n-1}) \neq 0$, tenemos

$$(2.1) \quad \epsilon_n = -\frac{f''(\xi_{n-1})}{2f'(x_{n-1})}\epsilon_{n-1}^2.$$

Si en un intervalo $I = [x_* - r, x_* + r]$, $f'(x)$ no se anula entonces $c_1 = \min_{x \in I} |f'(x)| > 0$. Definamos $c_2 = \max_{x \in I} |f''(x)|$, entonces si $\epsilon_0 = x_* - x_0$ verifica $|\epsilon_0| < r$ y $c_2/c_1 |\epsilon_0| = \gamma < 1$, entonces por (2.1) tenemos $|\epsilon_1| < c_2/(2c_1) \epsilon_0^2 < \gamma |\epsilon_0| < r$, por lo tanto $x_1 \in I$. Inductivamente, se obtiene $x_n \in I$ y $|\epsilon_n| < \gamma^{-n} r \rightarrow 0$, lo que implica $x_n \rightarrow x_*$. Entonces $\xi_n \rightarrow x_*$, de donde se obtiene $\epsilon_n/\epsilon_{n-1}^2 \rightarrow -f''(x_*)/(2f'(x_*))$.

Vimos que si $f'(x_*) \neq 0$, el error se comporta en forma cuadrática, es decir $\epsilon_n/\epsilon_{n-1}^2 \rightarrow c$. Vamos a analizar el error cuando $f'(x_*) = 0$, tomemos como ejemplo $f(x) = a(x - x_*)^m$. Claramente $f'(x_*) = 0$, los errores se relacionan de la forma

$$\epsilon_n = x_* - x_n = x_* - \left(x_{n-1} - \frac{a(x_{n-1} - x_*)^m}{m a(x_{n-1} - x_*)^{m-1}} \right) = \left(1 - \frac{1}{m} \right) \epsilon_{n-1}.$$

Vemos que $\epsilon_n \rightarrow 0$, pero el error decrece linealmente, más lentamente a medida que m es mayor. En el caso general, si $f(x_*) = 0, \dots, f^{(m-1)}(x_*) = 0$, cerca de x_* la función $f(x) \cong a(x - x_*)^m$, por lo que su comportamiento asintótico será similar al caso anterior.

Por ejemplo, la ecuación $f(x) = (x - 2\pi)(\cos(x) - 1) = 0$, el punto $x_* = 2\pi$ es un cero de tercer orden, es decir $f(x_*) = 0, f'(x_*) = 0, f''(x_*) = 0$. Tomamos $x_0 = 4$, en la Tabla 2.3 mostramos las iteraciones y los errores, vemos que $\epsilon_n/\epsilon_{n-1} \rightarrow 2/3$.

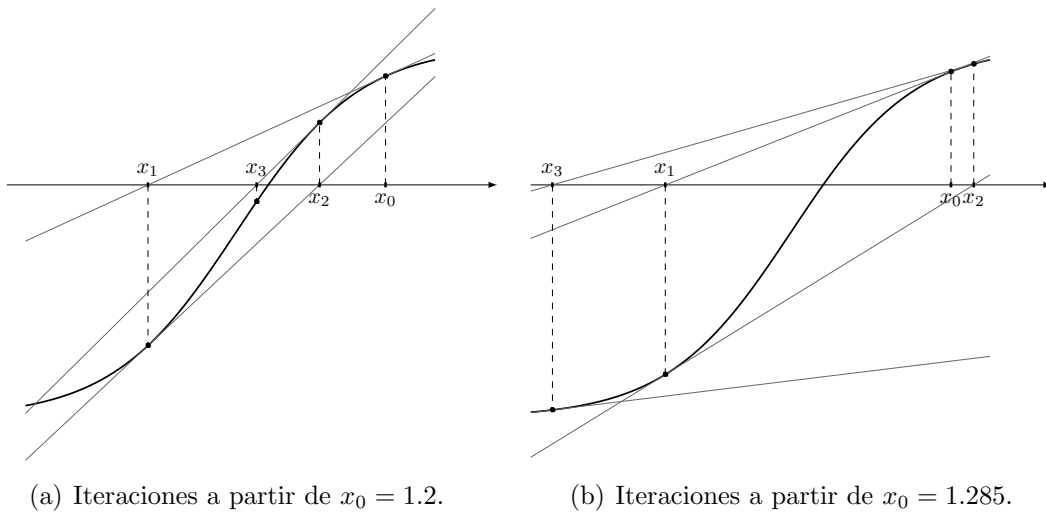
Convergencia para funciones no analíticas El análisis anterior asume que alguna derivada es diferente de cero, existen funciones que no verifican esto, por ejemplo $f(x) = xe^{-1/x^2}$ ($f(0) = 0$). Es posible mostrar que $f^{(m)}(0) = 0$ para $m \geq 0$. El único cero de la función es $x_* = 0$. Numéricamente podemos ver que la convergencia es muy lenta, si $x_0 = 1$, obtenemos $x_{100} = 0.0989013$, $x_{1000} = 0.0315785$, $x_n = 0.00141421$ para $n = 5 \times 10^5$. Podemos ver que $x_n = 2x_{n-1}/(2 + x_{n-1}^2)$, de donde resulta $\epsilon_n/\epsilon_{n-1} \rightarrow 1$.

n	x_n	ϵ_n	$\epsilon_n/\epsilon_{n-1}$
0	4.00000	2.28319	
1	5.11652	1.16667	0.51098
2	5.53803	0.74516	0.63871
3	5.79440	0.48879	0.65595
4	5.95953	0.32366	0.66216
5	6.06805	0.21514	0.66471
6	6.13994	0.14324	0.66581

Tabla 2.3: Iteraciones de la ecuación $(x - 2\pi)(\cos(x) - 1) = 0$.

2.2.4. Resultados de convergencia global. Consideremos la ecuación $\tanh(x) = 0.25$, cuya solución $x = 0.255413$. Veamos que pasa con diferentes puntos iniciales, si comenzamos con el punto $x_0 = 1.2$, vemos que la sucesión converge a la solución como se ve en la Tabla 2.4. Pero si el punto inicial es $x_0 = 1.285$, el comportamiento cambia. Los primeros términos de la sucesión son $x_1 = -1.01592$, $x_2 = 1.4683$, $x_3 = -1.92482$, $x_4 = 12.8762$, $x_5 = -2.86493 \times 10^{10}$, lo que muestra la no convergencia del método de Newton. En la Figura 2.4 mostramos gráficamente el comportamiento del método para los distintos punto iniciales.

n	x_n	ϵ_n	$\epsilon_n/\epsilon_{n-1}^2$
0	1.200000	-0.944587	
1	-0.713496	0.968909	1.085920
2	0.668420	-0.413007	-0.439938
3	0.161694	0.937186×10^{-1}	0.549427
4	0.253760	0.165250×10^{-2}	0.188144
5	0.255412	0.679682×10^{-6}	0.248899
6	0.255413	0.115519×10^{-12}	0.250058

Tabla 2.4: Iteraciones del método de Newton para $\tanh(x) = 0.25$.Fig. 2.4: Gráfico de $f(x) = \tanh(x) - 0.25$ y las iteraciones del método de Newton.

Veamos que hay hipótesis sobre $f(x)$ que garantizan la convergencia del método de Newton en forma incondicional. Supongamos que $f''(x) > 0$, es decir $f(x)$ es estrictamente convexa, si

x_* no es mínimo de $f(x)$, entonces $x_n \rightarrow x_*$ (si x_0 tampoco es un punto mínimo). En efecto, consideremos $f'(x_*) > 0$, si $x_0 > x_*$, es decir $\epsilon_0 < 0$, entonces $f'(x_0) > 0$ y $f(x_0) > 0$. Entonces $x_1 < x_0$ y por la ecuación (2.1), $\epsilon_1 < 0$. En conclusión, $x_* < x_1 < x_0$ y en general $x_* < x_n < x_{n-1} < \dots < x_0$. Siendo x_n una sucesión decreciente y acotada, es convergente y el límite tiene que ser una solución. Si $x_0 < x_*$ y $f'(x_0) > 0$, como $f''(x) > 0$, tenemos $x_1 > x_*$.

Fast inverse square root: Dado $x > 0$, se quiere calcular $y = x^{-1/2}$, o en forma equivalente $1/y^2 - x = 0$. La iteración del método de Newton es

$$y_{n+1} = y_n \frac{3 - y_n^2 x}{2} = y_n (1.5 - 0.5 x y_n^2).$$

El algoritmo *fast inverse square root* se basa en hallar un punto y_0 inicial cercano a y_* , lo que permite en pocas iteraciones (una o dos) del método de Newton obtener una buena aproximación. Si partimos de la representación en punto flotante $x = (1+m)2^e$, tenemos $\log_2(x) = e + \log_2(1+m) \cong e + m + \sigma$. Se puede ver que la mejor elección de σ es

$$\sigma = \frac{\ln(2) - 1 - \ln(\ln(2))}{2 \ln(2)} \cong 0.0430357,$$

en el sentido que minimiza la distancia $\max_{m \in [0,1]} |\log_2(1+m) - m - \sigma|$. La clave consiste en considerar a la representación en punto flotante como un número entero. Si definimos el entero $I(x) = b_0(x) + \dots + b_{62}(x) \times 2^{62}$ ($b_{63}(x) = 0$ dado que $x > 0$), tenemos

$$I(x) = m \times 2^{52} + (e + 1023) \times 2^{52} = (e + m + \sigma + 1023 - \sigma) \times 2^{52} \cong \log_2(x) + (1023 - \sigma) \times 2^{52}.$$

Usando que $\log_2(y_*) = -1/2 \log_2(x)$, obtenemos

$$\begin{aligned} I(y) &\cong -\frac{1}{2} \log_2(x) + (1023 - \sigma) \times 2^{52} \cong -\frac{1}{2} (I(x) - (1023 - \sigma) \times 2^{52}) + (1023 - \sigma) \times 2^{52} \\ &= \frac{3}{2} (1023 - \sigma) \times 2^{52} - \frac{1}{2} I(x) \end{aligned}$$

En https://github.com/id-Software/Quake-III-Arena/blob/master/code/game/q_math.c#L552 se puede encontrar el código en lenguaje C del video juego Quake-III-Arena, donde se muestra la implementación del algoritmo *Fast inverse square root*. En la línea 560 se considera al número y en punto flotante como un entero largo i . Se usa una sola iteración del método de Newton como se ve en 563-564. En 561 se observa la constante expresada en hexadecimal y la división de i por 2 mediante un desplazamiento de los bits hacia la izquierda. Por los comentarios se puede deducir que el programador no comprendía completamente el algoritmo.

```

552 float Q_rsqrt( float number )
553 {
554     long i;
555     float x2, y;
556     const float threehalfs = 1.5F;
557
558     x2 = number * 0.5F;
559     y = number;
560     i = * ( long * ) &y; // evil floating point bit level
        hacking
561     i = 0x5f3759df - ( i >> 1 ); // what the fuck?
562     y = * ( float * ) &i;
563     y = y * ( threehalfs - ( x2 * y * y ) ); // 1st iteration
564 // y = y * ( threehalfs - ( x2 * y * y ) ); // 2nd iteration, this can be
        removed
565
566     return y;
567 }
```

2.3. Método de punto fijo. El métodos de Newton es caso particular de los llamados métodos de punto fijo, estos métodos consisten en transformar la ecuación original en hallar la solución de $\phi(x) = x$ y buscarla mediante la sucesión $x_1 = \phi(x_0), x_2 = \phi(x_1), x_3 = \phi(x_2), \dots$, a

partir de un punto x_0 apropiado. Vale la pena aclarar que los problemas de punto fijo tienen interés en si mismo, muchas veces la evolución de un sistema se modela como la aplicación de una función al estado actual. En este caso, un punto fijo representa un estado estacionario, es decir un punto de equilibrio.

En general la forma de transformar el problema original en la ecuación de punto fijo no es única y no todas son apropiadas para obtener una solución aproximada. El siguiente ejemplo es una muestra de esta observación.

Ejemplo 2.2. Consideremos el problema $x^2 - 4 = 0$ cuyas soluciones son $x = \pm 2$. Si sumamos x a ambos miembros obtenemos la ecuación de punto fijo $x = \phi_1(x) = x^2 + x - 4$, podemos ver que la sucesión que se obtiene con ϕ_1 diverge para cualquier punto inicial, salvo si $x_0 = \pm 2$. Tomemos $x_0 = 1.9$, las iteraciones sucesivas resultan $x_1 = 1.51$, $x_2 = -0.2099$, $x_3 = -4.16584$, $x_4 = 9.1884$, $x_5 = 89.615$, $x_6 = 8116.47$. Obtendríamos algo similar partiendo de otros puntos iniciales. Esta forma de buscar las soluciones como una ecuación de punto fijo es completamente inútil.

Una manera alternativa de convertir el problema original en una ecuación de punto fijo, consiste en hacer las siguientes transformaciones: dividiendo por 2 obtenemos la ecuación $x^2/2 - 2 = 0$, sumando luego $x^2/2$ en ambos miembros resulta $x^2 - 2 = x^2/2$, por último despejamos la variable x del lado derecho como $x = \phi_2(x) = (x^2/2 + 2)^{1/2}$. Observemos que siendo $\phi_2(x) > 0$, las iteraciones no pueden converger a la solución negativa $x = -2$. En la Tabla 2.5 mostramos solamente las iteraciones pares por cuestiones de espacio para el punto inicial $x_0 = 8$.

n	x_n	ϵ_n	$\epsilon_n/\epsilon_{n-1}$
0	8.00000	-6.00000	
2	4.35890	-2.35890	0.615747
4	2.78388	-0.78388	0.563472
6	2.22205	-0.22205	0.523897
8	2.05776	-0.57759×10^{-1}	0.506923
10	2.01460	-0.14595×10^{-1}	0.501805
12	2.00366	-0.36588×10^{-2}	0.500456
14	2.00092	-0.91532×10^{-3}	0.500114
16	2.00023	-0.22887×10^{-3}	0.500029
18	2.00006	-0.57219×10^{-4}	0.500007
20	2.00001	-0.14305×10^{-4}	0.500002

Tabla 2.5: Iteraciones de $y = \phi_2(x)$.

Observemos que a medida que nos acercamos a la solución $x_* = 2$, los errores disminuyen con una tasa fija, 0.5 en este caso. Podemos ver en forma gráfica de ver las sucesivas iteraciones, conocido un punto x_n sobre el eje x subimos verticalmente hasta el gráfico de $y = \phi_2(x)$, esa altura corresponde a $x_n = \phi_2(x_{n-1})$. Para ubicar el punto sobre el eje de las abscisas, trazamos una recta horizontal hasta la intersección con la recta $y = x$ y luego verticalmente hasta el eje x . En la Figura 2.5 mostramos las primeras iteraciones correspondientes al punto inicial $x_0 = 8$.

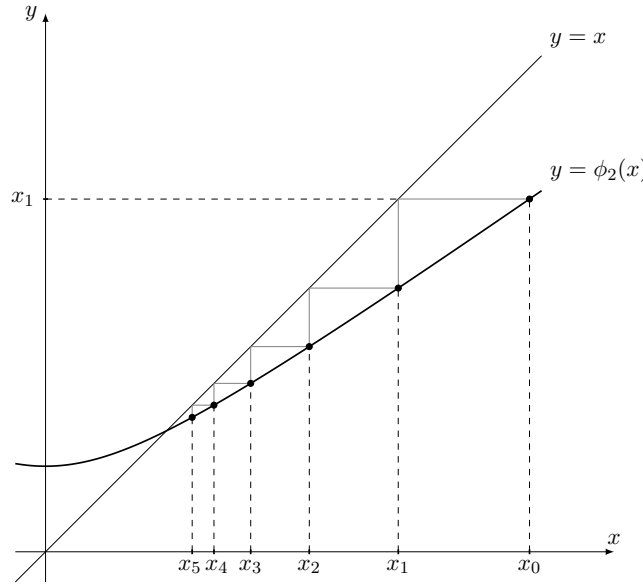


Fig. 2.5: Gráfico de $\phi_2(x)$ y las iteraciones del método de punto fijo.

2.3.1. Análisis de convergencia para funciones derivables. Vamos a estudiar ahora el comportamiento de las iteraciones en el Ejemplo 2.2. En general, buscamos condiciones que permitan asegurar la convergencia a una solución del problema. Si x_* es la solución, entonces $x_* = \phi(x_*)$, usando la recurrencia $x_n = \phi(x_{n-1})$, tenemos


$$(2.2) \quad \epsilon_n = x_* - x_n = \phi(x_*) - \phi(x_{n-1}) = \phi'(\xi_{n-1})(x_* - x_{n-1}) = \phi'(\xi_{n-1})\epsilon_{n-1},$$


donde ξ_{n-1} es un punto del intervalo determinado por x_* y x_{n-1} . Supongamos que un intervalo $I = [x_* - r, x_* + r]$ se verifica $|\phi'(x)| \leq \gamma < 1$, entonces por (2.2) para un punto inicial $x_0 \in I$ se verifica $|\epsilon_1| \leq \gamma |\epsilon_0| < |\epsilon_0|$, lo que muestra que $x_1 \in I$. De la misma forma, $|\epsilon_2| \leq \gamma |\epsilon_1| \leq \gamma^2 |\epsilon_0|$ y en general $x_n \in I$, $|\epsilon_n| = \gamma^n |\epsilon_0|$. Por lo tanto $|\epsilon| \rightarrow 0$ o en forma equivalente $x_n \rightarrow x_*$. Dado que $\xi_n \rightarrow x_*$, vemos que $\epsilon_n/\epsilon_{n-1} \rightarrow |\phi'(x_*)|$. En el Ejemplo 2.2, se verifica $\phi'(2) = 0.5$ lo que observa en la tasa de convergencia $\epsilon_n/\epsilon_{n-1}$.

Como ya vimos, el método de Newton es una ecuación de punto fijo con $\phi(x) = x - f(x)/f'(x)$

$$\phi'(x) = \frac{f(x)f''(x)}{f'(x)^2},$$

por lo tanto $\phi'(x_*) = 0$ y $\phi'(x) \cong -(x_* - x)f''(x_*)/f'(x_*)$.

 **Ejercicio 2.1.** Mostrar que para cualquier intervalo acotado I , existe $0 \leq \gamma < 1$ tal que $|\phi'_2(x)| \leq \gamma < 1$ para $x \in I$. Probar que el método de punto fijo converge a una solución en forma incondicional, es decir para cualquier punto inicial.

 **Ejercicio 2.2.** Estudiar numéricamente las iteraciones de $\phi(x) = \lambda x(1 - x)$ para $x_0 = 0.3$, en los casos $\lambda = 0.9, 2.5, 3.56, 3.9$.

2.4. Métodos de punto fijo para sistemas de ecuaciones. Consideremos el sistema de ecuaciones, con d ecuaciones y d incógnitas

$$(2.3) \quad \begin{aligned} 0 &= f_1(x_1, x_2, \dots, x_d), \\ 0 &= f_2(x_1, x_2, \dots, x_d), \\ &\vdots \\ 0 &= f_d(x_1, x_2, \dots, x_d). \end{aligned}$$

Igual que en el caso escalar, se puede transformar el sistema de ecuaciones (2.3) en un problema de punto fijo $\mathbf{x} = \phi(\mathbf{x})$, con $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^d$. Igual que en el caso escalar, podemos plantear un método iterativo $\mathbf{x}_n = \phi(\mathbf{x}_{n-1})$. Si la sucesión es convergente, el límite tiene que ser un punto fijo. Por ejemplo, consideramos el sistema

$$\begin{cases} 3x_1 + x_2 + x_3 = 8, \\ -x_1 + 4x_2 - x_3 = 9, \\ x_1 - 3x_2 + 5x_3 = 2, \end{cases}$$

siendo un sistema lineal, se resuelve por los métodos usuales (Gauss-Jordan) obteniendo la solución $\mathbf{x}_* = (1, 3, 2)$. Pero también podemos transformarlo en un problema de punto fijo mediante el método de Jacobi, que consiste en despejar una de variable diferente de cada ecuación:

$$\begin{cases} x_1 = (8 - x_2 - x_3)/3, \\ x_2 = (9 + x_1 + x_3)/4, \\ x_3 = (2 - x_1 + 3x_2)/5. \end{cases}$$

Si definimos $\phi : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ como

$$\phi(x_1, x_2, x_3) = ((8 - x_2 - x_3)/3, (9 + x_1 + x_3)/4, (2 - x_1 + 3x_2)/5)$$

podemos plantear el método iterativo $\mathbf{x}_n = \phi(\mathbf{x}_{n-1})$. En la Tabla 2.6 mostramos los resultados y el error $|\epsilon_n| = |\mathbf{x}_* - \mathbf{x}_n|$. Igual que en el caso escalar, la convergencia no está asegurada.

n	\mathbf{x}_n	$ \epsilon_n $	$ \epsilon_n / \epsilon_{n-1} $
0	(0.00000, 0.00000, 1.00000)	3.317	
1	(2.33333, 2.50000, 0.40000)	2.142	0.646
2	(1.70000, 2.93333, 1.43333)	0.903	0.421
3	(1.21111, 3.03333, 1.82000)	0.279	0.309
4	(1.04889, 3.00778, 1.97778)	0.543×10^{-1}	0.194
5	(1.00481, 3.00667, 1.99489)	0.968×10^{-2}	0.178
6	(0.99948, 2.99993, 2.00304)	0.308×10^{-2}	0.318
7	(0.99901, 3.00063, 2.00006)	0.117×10^{-2}	0.381
8	(0.99977, 2.99977, 2.00058)	0.662×10^{-3}	0.564
\vdots	\vdots	\vdots	\vdots
33	1.00000 , 3.00000 , 2.00000)	0.111×10^{-11}	0.455
34	1.00000 , 3.00000 , 2.00000)	0.506×10^{-12}	0.455

Tabla 2.6: Iteraciones de $\mathbf{x}_n = \phi(\mathbf{x}_{n-1})$ con $\mathbf{x}_0 = (0., 0., 1.)$.

Otras formas de transformar el sistema original en un problema de punto fijo pueden dar lugar a métodos no convergentes. Queremos obtener un criterio para determinar esto: linealizando alrededor del punto fijo tenemos $\phi(\mathbf{x}) = \mathbf{x}_* + \phi'(\mathbf{x}_*)(\mathbf{x} - \mathbf{x}_*) + O(|\mathbf{x} - \mathbf{x}_*|^2)$, por lo tanto

$$\epsilon_n = \mathbf{x}_* - \phi(\mathbf{x}_{n-1}) = -\phi'(\mathbf{x}_*)\epsilon_{n-1} + O(|\epsilon_{n-1}|^2),$$

donde $\phi'(\mathbf{x})$ es la matriz jacobiana. El criterio de convergencia para puntos iniciales \mathbf{x}_0 cercanos a \mathbf{x}_* consiste en ver que el radio espectral $\varrho(\mathbf{x}_*)$ de $\phi'(\mathbf{x}_*)$ sea menor que 1. El radio espectral es el máximo de los módulos de los autovalores, es decir

$$\varrho(\mathbf{x}_*) = \max \{|\lambda| : \lambda \text{ autovalor de } \phi'(\mathbf{x}_*)\}.$$

En el caso anterior, los autovalores de la matriz

$$\phi'(\mathbf{x}_*) = \begin{pmatrix} 0 & -1/3 & -1/3 \\ 1/4 & 0 & 1/4 \\ -1/5 & 3/5 & 0 \end{pmatrix}$$

son $\lambda_1 = -0.4546$, $\lambda_{2,3} = 0.2273 \pm i0.1472$, por lo tanto $\varrho = |\lambda_1| = 0.4546 > 0.271 = |\lambda_{2,3}|$. Observemos en la Tabla 2.6 que el comportamiento de los errores responde aproximadamente a la relación $|\epsilon_n| \cong \varrho(\mathbf{x}_*) |\epsilon_{n-1}|$.

Se podría pensar que resolver un problema lineal mediante métodos iterativos es ineficiente dado que se puede hallar la solución exacta con métodos directos, como por ejemplo Gauss–Jordan. La primera observación es que los métodos directos dejan de ser exactos si las operaciones se realizan en punto flotante. Por otro lado, el número de operaciones que deben hacerse para resolver un sistema de ecuaciones por el método de Gauss–Jordan es $O(d^3)$, en cambio los métodos de punto fijo requieren n multiplicaciones de matriz por vector es decir nd^2 las operaciones en punto flotante, que puede ser sustancialmente menor si $n \ll d$. Hay diferentes formas de transformar un sistema de ecuaciones lineales, entre los más conocidos tenemos el método de Jacobi y Gauss-Seidel.

¿Si la condición sobre el radio espectral se verifica en todo \mathbb{R}^n , $\varrho(\mathbf{x}) < 1$, el método es incondicionalmente convergente?

2.5. Método de la secante. El método de la secante es similar al método de Newton, la diferencia consiste en que en lugar de usar la recta tangente por el punto donde estamos actualmente, tomamos la recta secante que pasa por los últimos dos puntos obtenidos por el método. Concretamente, si x_{n-2} , x_{n-1} son los puntos obtenidos con el método en pasos sucesivos, planteamos el punto siguiente x_n como el resultado de resolver la ecuación $p_1(x) = 0$, donde $p_1(x)$ es el polinomio interpolador de orden 1 en x_{n-1} y x_{n-2} , es decir

$$p_1(x) = f[x_{n-1}] + f[x_{n-1}, x_{n-2}](x - x_{n-1}) = f(x_{n-1}) + \frac{f(x_{n-1}) - f(x_{n-2})}{x_{n-1} - x_{n-2}}(x - x_{n-1}),$$

que luego de despejar obtenemos

$$x_n = \frac{x_{n-2}f(x_{n-1}) - x_{n-1}f(x_{n-2})}{f(x_{n-1}) - f(x_{n-2})}.$$

La implementación se muestra en el Algoritmo 2.3. Cada paso requiere una única evaluación de la función. El algoritmo se detiene cuando la primera iteración que da puntos muy cercanos entre sí o cuando se supera el número máximo de iteraciones.

Algoritmo 2.3: Método de la secante.

Data: $f, x_0, x_1, \delta, N_{\max}$
Result: x, code
 $n = 0;$
 $x_{\text{old}} = x_0;$
 $x_{\text{new}} = x_1;$
 $y_{\text{old}} = f(x_{\text{old}});$
 $y_{\text{new}} = f(x_{\text{new}});$
while $|x_{\text{new}} - x_{\text{old}}| > \delta \wedge n \leq N_{\max}$ **do**
 $n = n + 1;$
 $x_{\text{aux}} = (x_{\text{old}} * y_{\text{new}} - x_{\text{new}} * y_{\text{old}}) / (y_{\text{new}} - y_{\text{old}});$
 $x_{\text{old}} = x_{\text{new}};$
 $y_{\text{old}} = y_{\text{new}};$
 $x_{\text{new}} = x_{\text{aux}};$
 $y_{\text{new}} = f(x_{\text{new}});$
end
if $n \leq N_{\max}$ **then**
 $\text{code} = 0;$
else
 $\text{code} = 1;$
end
return $x_{\text{new}}, \text{code};$

n	x_n	ϵ_n	$ \epsilon_n / \epsilon_{n-1} ^\varphi$
0	6.0000000000	-2.193	
1	5.0000000000	-1.193	0.335
2	4.5944830624	-0.788	0.592
3	4.1524045162	-0.346	0.509
4	3.9199653449	-0.113	0.632
5	3.8248131138	-0.182×10^{-1}	0.615
6	3.8076683978	-0.101×10^{-2}	0.660
7	3.8066715896	-0.910×10^{-5}	0.644
8	3.8066624943	-0.458×10^{-8}	0.656
9	3.8066624898	-0.209×10^{-13}	0.650

Tabla 2.7: Iteraciones del método de secantes para $e^x = 45$.**Ejemplo 2.3.**

2.5.1. Fórmula del error para el método de las secante. Vamos a estudiar el error $\epsilon_n = x_* - x_n$ al aplicar el método de la secante. Si tomamos el polinomio de interpolador de primer orden de f en los puntos x_{n-2}, x_{n-1} , de la fórmula del error de interpolación obtenemos

$$0 = f(x_*) = f(x_{n-1}) + f[x_{n-1}, x_{n-2}]\epsilon_{n-1} + \frac{f''(\xi_{n-1})}{2}\epsilon_{n-1}\epsilon_{n-2},$$

para algún ξ_{n-1} en el intervalo que contiene a los puntos x_*, x_{n-1}, x_{n-2} . Como la sucesión es obtenida por el método de la secante, se verifica

$$0 = f(x_{n-1}) + f[x_{n-1}, x_{n-2}](x_n - x_{n-1}),$$

usando que $\epsilon_{n-1} = \epsilon_n + x_n - x_{n-1}$ obtenemos $0 = f[x_{n-1}, x_{n-2}]\epsilon_n + \frac{1}{2}f''(\xi_{n-1})\epsilon_{n-1}\epsilon_{n-2}$, asumiendo que $f[x_{n-1}, x_{n-2}] \neq 0$, tenemos

$$(2.4) \quad \epsilon_n = -\frac{f''(\xi_{n-1})}{2f[x_{n-1}, x_{n-2}]} \epsilon_{n-1}\epsilon_{n-2}.$$

Si en un intervalo $I = [x_* - r, x_* + r]$, $f'(x)$ no se anula entonces

$$\min_{x, x' \in I} |f[x, x']| = c_1 > 0.$$

Definamos $c_2 = \max_{x \in I} |f''(x)|$, entonces si se verifica $|\epsilon_0|, |\epsilon_1| < r$ y $c_2/(2c_1)r = \gamma < 1$, por (2.4) tenemos

$$|\epsilon_2| < \frac{c_2}{2c_1} |\epsilon_0| |\epsilon_1| < \gamma |\epsilon_1| < r,$$

por lo tanto $x_2 \in I$. Inductivamente, se obtiene $x_n \in I$ y $|\epsilon_n| < \gamma^{-n}r \rightarrow 0$, lo que implica $x_n \rightarrow x_*$. Entonces $\xi_n \rightarrow x_*$, de donde se obtiene

$$(2.5) \quad \frac{\epsilon_n}{\epsilon_{n-1}\epsilon_{n-2}} \rightarrow -\frac{f''(x_*)}{2f'(x_*)}.$$

Proposición 2.1. Si la sucesión de números positivos $\{e_0, e_1, \dots\}$ verifica

$$\frac{e_n}{e_{n-1}e_{n-2}} \rightarrow c > 0,$$

para $n \geq 2$, entonces $e_n e_{n-1}^{-\varphi} \rightarrow c^{\varphi-1}$, con $\varphi = (1 + \sqrt{5})/2$.

Lema 2.1. Dada una sucesión de números reales $\{\lambda_1, \lambda_2, \dots\}$, son equivalentes las siguientes afirmaciones

(1) Existe $0 < r < 1$ tal que $\lambda_n + r\lambda_{n-1} \rightarrow 0$.

(2) $\lambda_n \rightarrow 0$.

(3) $\lambda_n + r\lambda_{n-1} \rightarrow 0$ para todo $r \in \mathbb{R}$.

Demostración. Vamos a mostrar que (1) implica (2). Probaremos que para $\varepsilon > 0$, existe $n_0 \in \mathbb{N}$ tal que $|\lambda_n| < \varepsilon$ si $n \geq n_0$. Por (1), existe $n_1 \in \mathbb{N}$ tal que $|\lambda_n + r\lambda_{n-1}| < \varepsilon(1-r)/2$ si $n \geq n_1$, por lo tanto

$$|\lambda_n| \leq r|\lambda_{n-1}| + |\lambda_n + r\lambda_{n-1}| < r|\lambda_{n-1}| + \frac{\varepsilon(1-r)}{2}.$$

Inductivamente obtenemos

$$|\lambda_{n_1+k}| \leq r^k |\lambda_{n_1}| + \frac{\varepsilon(1-r)}{2} (1 + r + \dots + r^{k-1}) < r^k |\lambda_{n_1}| + \frac{\varepsilon}{2},$$

por lo tanto existe $k_1 \in \mathbb{N}$ tal que $r^k |\lambda_{n_1}| < \varepsilon/2$ si $k \geq k_1$, por lo tanto $|\lambda_n| < \varepsilon$ para $n \geq n_0 = n_1 + k_1$.

Las afirmaciones (2) implica (3) y (3) implica (1) son inmediatas. \square

Demostración de la Proposición 2.1. Si definimos $\lambda_n = \ln(e_n) - \varphi \ln(e_{n-1}) - (\varphi - 1) \ln(c)$ y $r = \varphi - 1 = \varphi^{-1}$, tenemos

$$\lambda_n + r\lambda_{n-1} = \ln \left(\frac{e_n}{c e_{n-1} e_{n-2}} \right) \rightarrow 0,$$

usando el Lema anterior vemos que $\lambda_n \rightarrow 0$ y por lo tanto $e_n e_{n-1}^{-\varphi} c^{-(\varphi-1)} = e^{\lambda_n} \rightarrow 1$. \square

Como consecuencia de (2.4), usando la Proposición 2.1 deducimos

$$\frac{|\epsilon_n|}{|\epsilon_{n-1}|^\varphi} \rightarrow \left| \frac{f''(x^*)}{2f'(x^*)} \right|^{\varphi-1}$$

En la última columna de la Tabla 2.7, vemos que $|\epsilon_n|/|\epsilon_{n-1}|^\varphi \rightarrow 0.5^{0.618} = 0.6515$
El número de oro

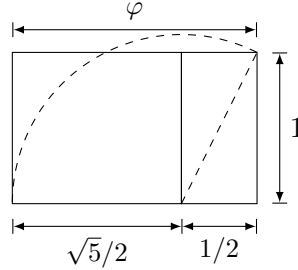


Fig. 2.6: Rectángulo áureo de Euclides.

2.6. Aplicaciones.

2.6.1. Red de resistores. Muchas ocasiones es útil considerar sistemas que involucran grandes números de resistencias en diversas geometrías, a veces de dimensión infinita. Hay muchas aplicaciones prácticas que utilizan este tipo de redes, por ejemplo, en la exploración geofísica, en la prospección de agua y petróleo. También en ingeniería es muy útil conocer el funcionamiento de redes, como los son las redes eléctricas o de distribución de agua o gas en una ciudad. Nuestro objetivo aquí es presentar una serie de modelos simples de redes que pueden ser estudiadas en el laboratorio. Consideramos una red larga de resistores como se muestra en la Figura 2.7.

Un sistema simple e interesante de analizar es el llamado red en escalera que se ilustra en la Figura 2.7. Esta red consiste en n mallas idénticas, consistente cada una de ellas en un par de resistores de valores r_i y r_m . Para calcular la resistencia equivalente de este sistema se puede usar el siguiente procedimiento recursivo. Llamaremos R_1 a la resistencia equivalente de la primera malla, R_2 a la de la siguiente y así sucesivamente.

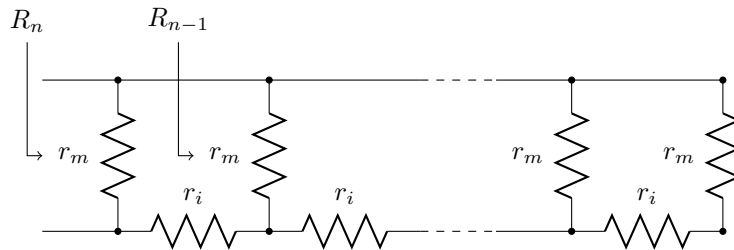


Fig. 2.7

La resistencia que se observa desde cada elemento de la red se relaciona por la ecuación

$$R_n = \phi(R_{n-1}) = \frac{r_m(r_i + R_{n-1})}{r_m + r_i + R_{n-1}},$$

con $R_0 = r_m$. Los únicos puntos fijos de ϕ son $R = r_i \left(-1/2 \pm \sqrt{1/4 + \mu} \right)$, donde $\mu = r_m/r_i$. Pero como las resistencias son positivas, el único punto fijo con significado físico es

$$R_* = r_i \left(-\frac{1}{2} + \sqrt{\frac{1}{4} + \mu} \right).$$

Derivando obtenemos

$$\phi'(R) = \frac{4\mu^2}{(2\mu + \sqrt{4\mu + 1} + 1)^2},$$

que verifica $0 < \phi'(R) < 1$, para $R > 0$, por lo tanto $r_n \rightarrow R_*$ cuando $n \rightarrow \infty$.

CAPÍTULO 3

Resolución Numérica de Ecuaciones Diferenciales

“En la crisis de 1972 el presidente Nixon, anunció que la tasa de incremento de la inflación estaba descendiendo. Esa fue la primera vez que un presidente usó la tercera derivada como argumento para su reelección.”

Hugo Rossi

3.1. Problemas de valores iniciales. En este capítulo vamos a estudiar métodos de resolución de ecuaciones diferenciales. Las ecuaciones y sistemas diferenciales aparecen en casi todas las áreas de la física. El tema es muy amplio y cubre muchísimas técnicas de la física-matemática. Nos restringiremos a ecuaciones diferenciales ordinarias, es decir estudiaremos aquellos sistemas cuyo estado se describe por una función (escalar o vectorial) que depende de una sola variable, que en general podemos pensar que es el tiempo. Concretamente, $x(t)$ representa uno o varios números que determinan el estado del sistema bajo estudio en el instante t . Dentro de esta clase de problemas, solo consideraremos el problema de valores iniciales:

$$(3.1) \quad \begin{cases} \dot{x}(t) = f(t, x(t)), \\ x(t_0) = x_0, \end{cases}$$

donde $f(t, x)$ es una función suave en ambas variables. Sabemos que este problema tiene solución única definida en algún intervalo que contiene a t_0 . Además, depende en forma continua de los datos iniciales (t_0, x_0) .

A continuación damos algunos ejemplos de problemas escalares ($x(t) \in \mathbb{R}$) de valores iniciales, relacionados con distintas aplicaciones.

3.1.1. Crecimiento poblacional. Los modelos clásicos de crecimiento del número de individuos de una especie son:

- I. Crecimiento exponencial (Malthus 1798): $\dot{N} = rN$.
- II. Crecimiento logístico (Verhulst, 1838): $\dot{N} = r(1 - N/K)N$.

En el primer caso la solución es $N(t) = N_0 e^{rt}$, el segundo tiene como respuesta:

$$N(t) = \frac{N_0 K}{N_0 + (K - N_0)e^{-rt}}.$$

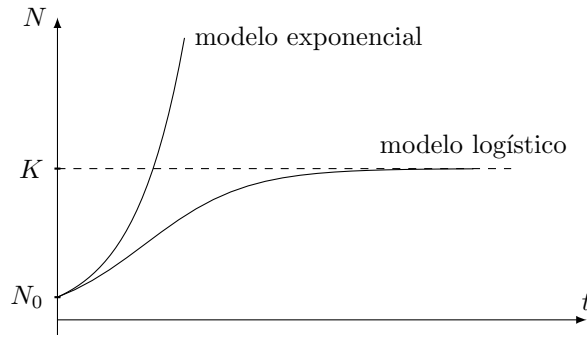


Fig. 3.1: Crecimiento de la población

Sin embargo, modelos más generales de la forma $\dot{N} = g(N)N$ fueron planteados en diferentes situaciones. Por ejemplo, un modelo usado para describir el crecimiento de tumores se basa en la ley de Gompertz.

3.1.2. Circuitos R-C y R-L. Se denomina circuito R-C a un circuito eléctrico compuesto de varios resistores y capacitores interconectados. El más simple de los circuitos R-C consiste de un sólo resistor y un capacitor, como se muestra en la Figura 3.2(a).

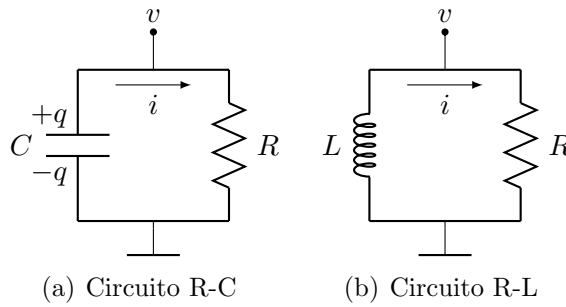


Fig. 3.2: Circuitos R-C y R-L simples.

Si el capacitor tiene carga inicialmente, al estar conectado al resistor, circulará una corriente eléctrica a través del circuito. Las placas del capacitor perderán su carga y la energía eléctrica se transformará en calor (ley de Joule). La ley de Ohm establece que la intensidad de la corriente i que circula por un resistor es proporcional a la diferencia de potencial v aplicada en sus extremos, es decir $v = Ri$, siendo R su resistencia. Por otro lado, sabemos que la relación entre la carga almacenada en un capacitor es proporcional a la diferencia de potencial entre sus conductores, $q = Cv$. Por conservación de la carga, y asumiendo que la carga sólo se acumula en el capacitor, tenemos $i = -\dot{q} = -C\dot{v}$, de donde se deduce $RC\dot{v} + v = 0$. Si el potencial en el tiempo t_0 es v_0 , la evolución en el tiempo del potencial eléctrico está dado por $v(t) = v_0 e^{-(t-t_0)/\tau}$, donde $\tau = RC$. Observemos que si la resistencia R se mide en Ω (ohmios) y la capacidad C en F (faradios), τ se mide en s (segundos). La ecuación diferencial de v refleja la conservación de la energía. En efecto, la energía eléctrica acumulada en el capacitor está dada por $\mathcal{E}_{\text{elec}} = qv/2 = Cv^2/2$, por lo tanto la variación en el tiempo de dicha energía vale $\dot{\mathcal{E}}_{\text{elec}} = Cv\dot{v}$. Por otro lado, la ley de Joule establece que la energía transformada en calor por unidad de tiempo, está dada por $\dot{\mathcal{E}}_{\text{cal}} = iv = v^2/R$. Vale entonces

$$\dot{\mathcal{E}}_{\text{elec}} + \dot{\mathcal{E}}_{\text{cal}} = Cv\dot{v} + \frac{v^2}{R} = 0.$$

Analizamos ahora el circuito presentado en la Figura 3.2(b), formado por un resistor y un inductor. Igual que en el caso anterior, de la ley de Ohm obtenemos $v = Ri$. La ley de Faraday

establece la relación entre la corriente que circula por el inductor y la diferencia de potencial entre sus extremos, $v = -L di/dt$, donde L es el coeficiente de autoinducción.

De ambas relaciones, obtenemos la ecuación diferencial $\dot{v} + R/Lv = 0$, cuya solución es similar a la del caso anterior, con $\tau = L/R$. Si el valor resistencia R está dado en Ω y el de la inductancia L en H (henrios), τ queda expresado en s.

3.1.3. Caída libre. Se denomina caída libre al movimiento de un cuerpo bajo la acción de un campo gravitatorio. Si se considera la fuerza de rozamiento fluidodinámico, la ecuación de movimiento se puede escribir en términos la altura z :

$$m\dot{v} = -mg + f_r,$$

donde g es la aceleración de la gravedad y f_r es la fuerza de rozamiento que depende de v . Para velocidades altas (régimen turbulento), f_r está dada por

$$f_r = \frac{1}{2}C_d\rho A v^2,$$

donde C_d es el coeficiente aerodinámico de resistencia al avance, ρ la densidad del fluido y A es el área transversal a la dirección del movimiento. Si definimos $\gamma = 1/2 C_d\rho A$, la ecuación diferencial para v es $\dot{v} = -g + \gamma/mv^2$. Si suponemos que el cuerpo inicialmente se encuentra en reposo, $v(0) = 0$, la solución está dada por $v(t) = -g\tau \tanh(t/\tau)$, con $\tau = \sqrt{m/(g\gamma)}$. Vemos que, a diferencia de lo ocurre cuando no hay rozamiento, la velocidad se estabiliza en un valor final $v_\infty = g\tau$. Para más detalles, ver el Ejercicio 3.11.

3.2. Método de Euler. En algunos problemas, como los ejemplos anteriores, se pueden obtener soluciones analíticas. Sin embargo, esto es raro y en la mayoría de los problemas solo podemos obtener soluciones aproximadas por medio de métodos numéricos. Inicialmente, vamos a considerar el caso escalar, es decir $x(t) \in \mathbb{R}$, la extensión al caso vectorial (o a ecuaciones de mayor orden) es inmediata. Resolveremos los problemas de valores iniciales hacia el futuro, $t \in [t_0, t_0 + T]$ (con $T > 0$). Esto no representa ninguna restricción, siendo que (3.1) es un problema reversible. En efecto, si resolvemos

$$\begin{cases} \dot{\tilde{x}}(t) = -f(-t, \tilde{x}(t)), \\ \tilde{x}(-t_0) = x_0, \end{cases}$$

en $[-t_0, -t_0 + T]$, entonces $x(t) = \tilde{x}(-t)$ es la solución del problema (3.1) en $[t_0 - T, t_0]$.

El más sencillo de los métodos de integración numérica se conoce el método de Euler explícito. Básicamente consiste en aproximar la solución mediante la recta tangente que pasa por el punto (t_0, x_0) dada por

$$x = x(t_0) + \dot{x}(t_0)(t - t_0) = x_0 + f(t_0, x_0)(t - t_0).$$

En el tiempo $t_1 = t_0 + h_0$ obtenemos $x_1 = x_0 + h_0 f(t_0, x_0)$. Repitiendo la operación a partir del punto (t_1, x_1) , obtenemos un nuevo punto $x_2 = x_1 + h_1 f(t_1, x_1)$, correspondiente al tiempo $t_2 = t_1 + h_1$ (ver Figura 3.4). Concretamente, planteamos la sucesión de puntos (t_n, x_n) definida por

$$\begin{aligned} t_n &= t_{n-1} + h_{n-1}, \\ x_n &= x_{n-1} + h_{n-1} f(t_{n-1}, x_{n-1}), \end{aligned}$$

donde h_n son valores arbitrarios (pequeños). Como la recta tangente es una aproximación de la función para pequeños incrementos en la variable independiente, $x_1 \cong x(t_1)$, siendo $x(t)$ es la solución exacta. A partir del segundo paso, el punto inicial no coincide con el verdadero. Sin embargo, por la continuidad de la soluciones con respecto a los datos iniciales, $x_n \cong x(t_n)$. En general vamos a tomar $h_n = h$ y por lo tanto $t_n = t_0 + nh$.

Ejemplo 3.1. Para ilustrar el método de Euler explícito, vamos a estudiar el problema

$$\begin{cases} \dot{x}(t) = x(t), \\ x(0) = 1, \end{cases}$$

en el intervalo $[0, 1]$, como conocemos la solución exacta, $x(t) = e^t$, podemos obtener los errores obtenidos con el método numérico para diferentes pasos temporales h . Observemos que para dos discretizaciones distintas, la solución en un instante t se alcanza en elementos diferentes de la sucesión x_n . En efecto, si para los pasos $h, \tilde{h} > 0$ tenemos las soluciones aproximadas x_n, \tilde{x}_n , tendremos que $x(t) \cong x_n$ y $x(t) \cong \tilde{x}_{\tilde{n}}$, donde $t = nh = \tilde{n}\tilde{h}$. En la Tabla 3.1 se muestran los errores $E_n = |x(nh) - x_n|$ para $h = 0.1, 0.01$. En la Figura 3.3 graficamos los errores para

t	$x(t)$	x_n ($h = 0.1$)	Error	x_n ($h = 0.01$)	Error
0.1	1.105	$x_1 = 1.100$	5.17×10^{-3}	$x_{10} = 1.105$	5.49×10^{-4}
0.2	1.221	$x_2 = 1.210$	1.14×10^{-2}	$x_{20} = 1.220$	1.21×10^{-3}
0.3	1.350	$x_3 = 1.331$	1.89×10^{-2}	$x_{30} = 1.348$	2.01×10^{-3}
0.4	1.492	$x_4 = 1.464$	2.77×10^{-2}	$x_{40} = 1.489$	2.96×10^{-3}
0.5	1.649	$x_5 = 1.611$	3.82×10^{-2}	$x_{50} = 1.645$	4.09×10^{-3}
0.6	1.822	$x_6 = 1.772$	5.06×10^{-2}	$x_{60} = 1.817$	5.42×10^{-3}
0.7	2.014	$x_7 = 1.949$	6.50×10^{-2}	$x_{70} = 2.007$	6.99×10^{-3}
0.8	2.226	$x_8 = 2.144$	8.20×10^{-2}	$x_{80} = 2.217$	8.83×10^{-3}
0.9	2.460	$x_9 = 2.358$	1.02×10^{-1}	$x_{90} = 2.449$	1.10×10^{-2}
1.0	2.718	$x_{10} = 2.594$	1.25×10^{-1}	$x_{100} = 2.705$	1.35×10^{-2}

Tabla 3.1: Error en $t = 0.1, 0.2, \dots, 1$ para $h = 0.1$ y $h = 0.01$.

$h = 10^{-1}, 10^{-2}, 10^{-3}$ en un sistema de coordenadas semi-logarítmico. Podemos observar que los gráficos son asintóticos a rectas, lo que refleja el comportamiento exponencial de los errores. Esto coincide con el análisis del error que haremos más adelante.

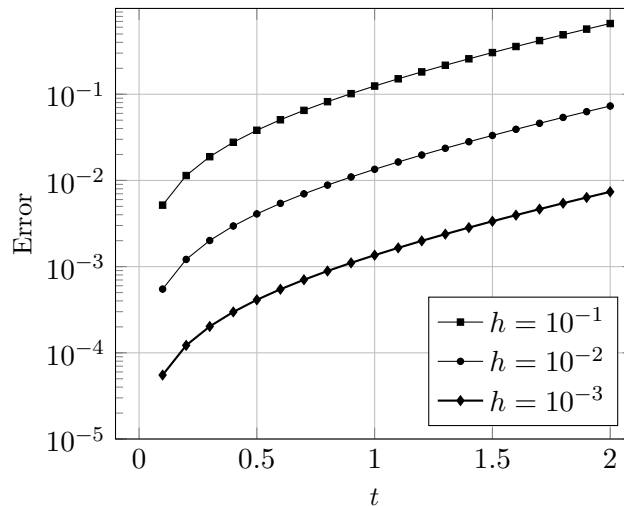


Fig. 3.3: Error del método de Euler para la ecuación $\dot{x}(t) = x(t)$.

El método anterior se describe en el código mostrado en el Algoritmo 3.1, los datos son: tiempo inicial t_0 , el valor inicial x_0 , el tiempo de integración T , el paso h temporal y la función que representa la ecuación diferencial f . La salida es una lista de tiempos t_list y otra de los estados correspondientes x_list .

Algoritmo 3.1: Método de Euler explícito.**Data:** f, t_0, T, h, x_0 **Result:** t_list, x_list $n = 0;$ $t_list(0) = t_0;$ $x_list(0) = x_0;$ **while** $t_list(n) - t_0 < T$ **do** $n = n + 1;$ $t_list(n) = t(n - 1) + h;$ $x_list(n) = x_list(n - 1) + hf(t_list(n - 1), x_list(n - 1));$ **end**

▣ **Ejercicio 3.4.** Escribir un programa que implemente el método de Euler explícito descrito en el Algoritmo 3.1 para resolver el problema (3.1).

3.2.1. Error de truncamiento del método de Euler. Vamos a acotar el error que se comete al aproximar la solución exacta usando el método de Euler. Aclaremos que por *acotar* entendemos obtener una estimación pesimista del error, es decir que vamos a suponer que todas las cantidades no conocidas toman los valores que maximizan el error. Más adelante proponemos métodos de análisis del error más precisos, los cuales permiten optimizar los parámetros del método (el paso temporal h en este caso) para lograr integradores eficientes sin superar el error tolerado.

El análisis del error se basa en calcular el número de términos del desarrollo de Taylor de la solución aproximada que coinciden con los de la solución exacta. Si la coincidencia es hasta el término n -ésimo, decimos que el método es de orden n . En el caso del método de Euler, tenemos

$$x(t_1) = x(t_0) + h\dot{x}(t_0) + \frac{h^2}{2}\ddot{x}(\tau_1).$$

Como $\dot{x}(t) = f(t, x(t))$, tenemos

$$\ddot{x}(t) = f_t(t, x(t)) + f_x(t, x(t))\dot{x}(t) = f_t(t, x(t)) + f_x(t, x(t))f(t, x(t)),$$

de donde obtenemos

$$x(t_1) = x_0 + hf(t_0, x_0) + \frac{h^2}{2}(f_t(\tau_1, x(\tau_1)) + f_x(\tau_1, x(\tau_1))f(\tau_1, x(\tau_1))).$$

Como $x_1 = x(t_0) + hf(t_0, x_0)$, obtenemos

$$\epsilon_1 = x(t_1) - x_1 = \frac{h^2}{2}(f_t(\tau_1, x(\tau_1)) + f_x(\tau_1, x(\tau_1))f(\tau_1, x(\tau_1))) = O(h^2).$$

La cantidad ϵ_1 se denomina error local de truncamiento.

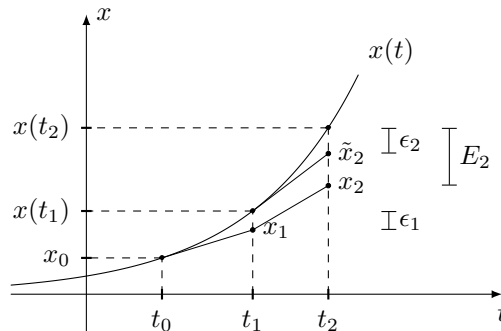


Fig. 3.4: Errores locales y globales.

3.2.2. Error global. Como vemos en el gráfico de la Figura 3.4, a partir del segundo paso, la diferencia entre la solución exacta y la aproximada no se debe solo al error de truncamiento local. Como partimos del punto (t_1, x_1) , que difiere del punto de la solución exacta $(t_1, x(t_1))$, se suma el error del paso anterior.. Queremos estudiar la acumulación de errores en cada paso, lo haremos para el método de Euler pero se aplica a todos los métodos propuestos. Definimos como error global en el paso n -ésimo a $E_n = |x(t_n) - x_n|$. Vamos a calcular una cota para E_2 y luego extenderemos el análisis al caso general. Repitiendo los argumentos utilizados anteriormente, vemos que el error local de truncamiento $\epsilon_2 = x(t_2) - \tilde{x}_2$, con $\tilde{x}_2 = x(t_1) + hf(t_1, x(t_1))$. El error en tiempo t_2 está definido como $E_2 = |x(t_2) - x_2|$, que a diferencia de lo que sucede en t_1 , no es ϵ_2 como vemos en la figura 3.4. Para calcular E_2 podemos escribir

$$E_2 \leq |x(t_2) - \tilde{x}_2| + |\tilde{x}_2 - x_2| = |\epsilon_2| + |\tilde{x}_2 - x_2|.$$

Para estimar el segundo término escribimos

$$\begin{aligned} |\tilde{x}_2 - x_2| &= |x(t_1) - x_1 + h(f(t_1, x(t_1)) - f(t_1, x_1))| \\ &\leq |x(t_1) - x_1| + h |f(t_1, x(t_1)) - f(t_1, x_1)|. \end{aligned}$$

Si suponemos que $|f(t_1, x(t_1)) - f(t_1, x_1)| \leq L |x(t_1) - x_1| = L |\epsilon_1|$, entonces

$$E_2 \leq |\epsilon_2| + (1 + Lh) |\epsilon_1|.$$

En general, vemos que $E_n \leq |\epsilon_n| + (1 + Lh)E_{n-1}$, e inductivamente podemos ver que

$$E_n \leq |\epsilon_n| + (1 + Lh) |\epsilon_{n-1}| + \cdots + (1 + Lh)^{n-1} |\epsilon_1|.$$

Si definimos $\epsilon_{\max} = \max \{|\epsilon_1|, \dots, |\epsilon_n|\}$, resulta

$$E_n \leq (1 + (1 + Lh) + \cdots + (1 + Lh)^{n-1}) \epsilon_{\max}.$$

De la expresión de la serie geométrica y usando que $1 + Lh \leq e^{Lh}$, tenemos

$$E_n \leq \frac{(1 + Lh)^n - 1}{Lh} \epsilon_{\max} \leq \frac{e^{Lnh} - 1}{Lh} \epsilon_{\max}.$$

Como $nh \leq T$, obtenemos

$$E_n \leq \frac{e^{LT} - 1}{L} \frac{\epsilon_{\max}}{h}.$$

Falta ver la condición que impusimos sobre $f(t, x)$. Por el teorema del valor medio (o de Lagrange) vemos que $f(t_1, x(t_1)) - f(t_1, x_1) = f_x(t_1, \xi)$, para algún punto ξ en el intervalo determinado por x_1 y $x(t_1)$. Por lo tanto, si $|f_x(t, x)| \leq L$, la hipótesis se verifica. En general no podemos suponer la acotación de $f_x(t, x)$ en todo el dominio, pero tampoco es necesario. Si tomamos un rectángulo $Q = [t_0, t_0 + T] \times [\xi_1, \xi_2]$ suficientemente grande para que la solución $(t, x(t))$ permanezca en Q , podemos usar la constante $L = \max \{|f_x(t, x)| : (t, x) \in Q\}$ (ver Figura 3.5). Para h pequeño, tendremos $(t_n, x_n) \in Q$ lo que termina con el argumento.

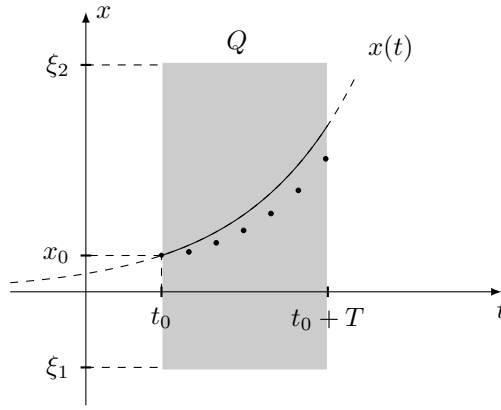


Fig. 3.5: En gris, el rectángulo Q que contiene a $(t, x(t))$. Los puntos indican la solución aproximada (t_n, x_n) .

3.2.3. Estimación del error. En 3.2.1, obtuvimos cotas de error de truncamiento con constantes que dependen de la solución exacta y estimaciones de las derivadas de $f(t, x)$, lo que vuelve difícil de estimar con precisión. Con el objetivo de mantener el error dentro de valores aceptables, buscamos fórmulas que nos permitan calcularlo con buen grado de aproximación. Como definimos en 3.2.1, el error local de truncamiento ϵ_1 es la diferencia entre la solución exacta $x(t_1)$ y x_1 . El error con paso h está dado por

$$\epsilon_1 = x(t_1) - x_1 = C_2 h^2 + O(h^3),$$

Sin en lugar de considerar un paso de tamaño h , consideramos dos pasos de tamaño $h/2$ obtenemos

$$\tilde{\epsilon}_1 = x(t_1) - \tilde{x}_1 = \frac{1}{2} C_2 h^2 + O(h^3),$$

restando las estimaciones tenemos $\tilde{x}_1 - x_1 = 1/2 C_2 h^2 + O(h^3)$, de donde podemos despejar

$$\epsilon_1 = 2(\tilde{x}_1 - x_1) + O(h^3).$$

En 3.2.1 definimos ϵ_n como la diferencia entre el valor de la solución exacta que pasa por (t_0, x_0) en el tiempo $t = t_n$, y la solución obtenida con el método desde el punto $(t_{n-1}, x(t_{n-1}))$ (Figura 3.4). Ahora, hemos considerado la diferencia entre el valor de la solución exacta que pasa por (t_{n-1}, x_{n-1}) en el tiempo $t = t_n$, que denotamos ξ_n , y la solución numérica x_n . Si suponemos que el error global es pequeño, ambos errores deben ser similares, como se puede apreciar de la comparación de las Figuras 3.4 y 3.6.

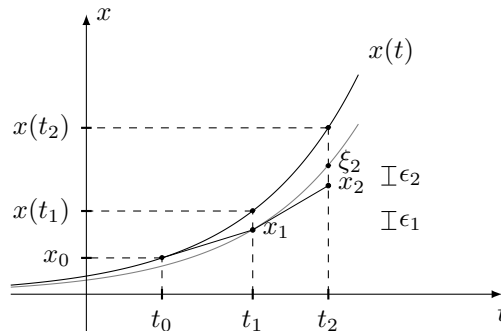


Fig. 3.6: Error local.

Analicemos el problema

$$\begin{cases} \dot{x}(t) = e^{-2t} - 2x(t), \\ x(0) = 0.1, \end{cases}$$

para $t \in [0, 1]$. La solución exacta que pasa por (t_{n-1}, x_{n-1}) está dada por

$$x(t) = (t - t_{n-1})e^{-2t} + e^{-2(t-t_{n-1})}x_{n-1},$$

entonces $\xi_n = he^{-2t_n} + e^{-2h}x_{n-1}$. En la Tabla 3.2 se muestran los valores obtenidos para $t_0 = 0$, $x_0 = 0.1$ y $h = 0.1$. Observemos que las diferencias entre los errores locales de truncamiento y

t	x_n	ξ_n	\tilde{x}_n	$\epsilon_n = \xi_n - x_n$	$2(\tilde{x}_n - x_n)$
0.1	0.1800	0.1637	0.1712	-1.625×10^{-2}	-1.752×10^{-2}
0.2	0.2259	0.2144	0.2197	-1.147×10^{-2}	-1.238×10^{-2}
0.3	0.2477	0.2398	0.2434	-7.920×10^{-3}	-8.565×10^{-3}
0.4	0.2531	0.2478	0.2502	-5.308×10^{-3}	-5.756×10^{-3}
0.5	0.2474	0.2440	0.2455	-3.405×10^{-3}	-3.708×10^{-3}
0.6	0.2347	0.2327	0.2336	-2.035×10^{-3}	-2.232×10^{-3}
0.7	0.2179	0.2168	0.2173	-1.064×10^{-3}	-1.184×10^{-3}
0.8	0.1990	0.1986	0.1987	-3.891×10^{-4}	-4.551×10^{-4}
0.9	0.1794	0.1794	0.1794	6.692×10^{-5}	3.895×10^{-5}
1.0	0.1600	0.1604	0.1602	3.632×10^{-4}	3.612×10^{-4}

Tabla 3.2: Error para $h = 0.1$ y la estimación con $h = 0.05$.

las estimaciones son del orden de $h^3 = 0.001$.

Las estimaciones del error permiten implementar métodos adaptativos, estos integradores modifican el paso h para lograr que el error local de truncamiento en cada paso se encuentre cerca de un valor preestablecido. Si fuera mayor, la solución numérica tendría un error superior al deseado. Por el contrario, si fuera mucho menor, se estarían desperdiciando recursos computacionales.

3.2.4. Extrapolación. En el párrafo anterior, mostramos como se puede estimar el error local de truncamiento comparando los resultados que se obtienen aplicando un paso del método de tamaño h y dos pasos con tamaño $h/2$. Podemos, en lugar de estimar el error, reducirlo considerando que $\xi_n = x_n + \epsilon_n$. Proponemos la modificación:

$$\bar{x}_n = x_n + 2(\tilde{x}_n - x_n) = 2\tilde{x}_n - x_n,$$

Explícitamente, tenemos

$$\begin{aligned} x_n &= x_{n-1} + hf(t_{n-1}, x_{n-1}), \\ \tilde{x}_n &= x_{n-1} + h/2f(t_{n-1}, x_{n-1}) + h/2f(t_{n-1} + h/2, x_{n-1} + h/2f(t_{n-1}, x_{n-1})) \end{aligned}$$

de donde obtenemos el método de Euler modificado:

$$\bar{x}_n = x_{n-1} + hf(t_{n-1} + h/2, x_{n-1} + h/2f(t_{n-1}, x_{n-1})).$$

Este método pertenece a la familia de métodos propuestos en el Ejercicio 3.10, cuando $\alpha = 1/2$ y $\beta = 1$.

Ejemplo 3.2. Consideremos el problema

$$\begin{cases} \dot{x}(t) = 3x(t) - 3x^2(t), \\ x(0) = 0.01, \end{cases}$$

cuya solución exacta es $x(t) = 0.01e^{3t}/(1 + 0.01(e^{3t} - 1))$. En la Figura 3.7 se muestran las soluciones obtenidas con ambos métodos. Para el método de Euler, tomamos un paso $h = 0.05$, y para el método de Euler modificado elegimos $h = 0.1$. De esta forma, el número de evaluaciones de $f(t, x)$ es el mismo. Como $x(t)$ toma valores muy diferentes ($0.01 \leq x(t) \leq 1$), consideramos el error relativo que se muestra en la Figura 3.7(b), donde vemos la superioridad del método de Euler modificado.

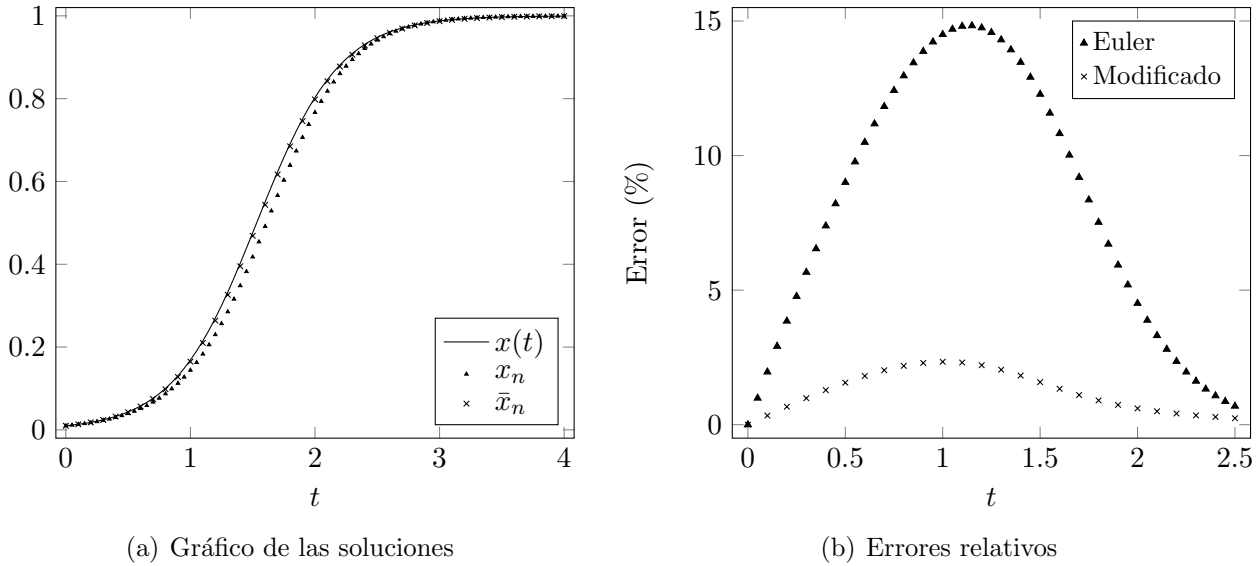


Fig. 3.7: Comparación de los errores de los métodos de Euler y Euler modificado.

3.2.5. Método de Euler implícito. Consideremos la ecuación $\dot{x}(t) = -\lambda x(t)$

3.3. Métodos de Taylor. En el método de Euler vimos que x_1 era una aproximación de primer orden del verdadero valor $x(t_1)$. Si lográramos que x_1 coincidiera con más términos del desarrollo de Taylor de $x(t)$, el orden del error de truncamiento sería mayor (y por lo tanto el del error global) y tendríamos un método más preciso. Para lograr esto desarrollamos los primeros términos de $x(t_1)$:


$$\begin{aligned} x(t_1) &= x_0 + h \dot{x}(t_0) + \frac{h^2}{2} \ddot{x}(t_0) + \frac{h^3}{6} \ddot{\ddot{x}}(\tau) \\ &= x_0 + h f(t_0, x_0) + \frac{h^2}{2} (f_t(t_0, x_0) + f_x(t_0, x_0)f(t_0, x_0)) + \frac{h^3}{6} \ddot{\ddot{x}}(\tau). \end{aligned}$$

Si tomamos

$$x_1 = x_0 + h f(t_0, x_0) + \frac{h^2}{2} (f_t(t_0, x_0) + f_x(t_0, x_0)f(t_0, x_0)),$$

obtenemos un método de segundo orden, es decir $\epsilon_n = O(h^3)$, lo que prueba (por un argumento similar al anterior) que el error global es $E_n = O(h^2)$.

 **Ejercicio 3.5.** Plantear el método de Taylor para el problema del Ejemplo 3.1.

 **Ejercicio 3.6.** Escribir un código para la resolución del problema del Ejemplo 3.1 mediante el método dado en el Ejercicio 3.5.

 **Ejercicio 3.7*.** Obtener el método de Taylor de tercer orden.

3.4. Métodos Runge-Kutta. Para implementar el método de Taylor necesitamos evaluar la función $f(t, x)$ y también sus derivadas. Esto presenta varios problemas. Por un lado tenemos que modificar todos los códigos donde se evalúan la función y sus derivadas, es decir que hay que escribir varias rutinas específicas para utilizar el integrador. Más aún, en muchos casos $f(t, x)$ no tiene una expresión analítica, sus evaluaciones son el resultado de cálculos realizados numéricamente. No podemos entonces obtener los valores de sus derivadas. Vamos a presentar en esta sección una familia de métodos de alto orden donde sólo necesitan evaluaciones de la función $f(t, x)$. Consideremos el más simple de estos integradores: tomamos $k_1 = f(t_0, x_0)$ y $k_2 = f(t_0 + h, x_0 + hk_1)$, desarrollando k_2 como función de h

$$k_2 = f(t_0, x_0) + h(f_t(t_0, x_0) + f_x(t_0, x_0)f(t_0, x_0)) + O(h^2),$$

si definimos $x_1 = x_0 + \frac{h}{2}(k_1 + k_2)$, vemos que

$$\begin{aligned} x_1 &= x_0 + hf(t_0, x_0) + \frac{h^2}{2}(f_t(t_0, x_0) + f_x(t_0, x_0)f(t_0, x_0)) + O(h^3), \\ x(t_1) &= x_0 + hf(t_0, x_0) + \frac{h^2}{2}(f_t(t_0, x_0) + f_x(t_0, x_0)f(t_0, x_0)) + O(h^3), \end{aligned}$$

restando ambas expresiones obtenemos $\epsilon_1 = x(t_1) - x_1 = O(h^3)$.

Describimos los pasos correspondiente al método en el Algoritmo 3.2.

Algoritmo 3.2: Método Runge-Kutta de orden 2.

Data: f, t_0, T, h, x_0

Result: t_list, x_list

$n = 0;$

$t_list(0) = t_0;$

$x_list(0) = x_0;$

while $t(n) - t_0 < T$ **do**

$n = n + 1;$

$t_list(n) = t_list(n - 1) + h;$

$k_1 = f(t_list(n - 1), x_list(n - 1));$

$k_2 = f(t_list(n), x_list(n - 1) + h * k_1);$

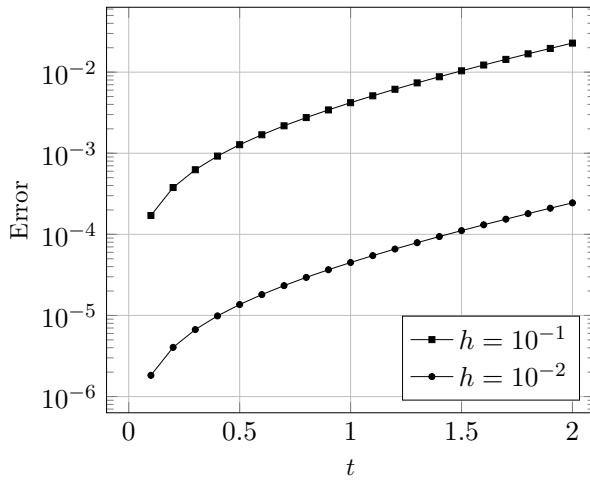
$x_list(n) = x_list(n - 1) + h * (k_1 + k_2)/2;$

end

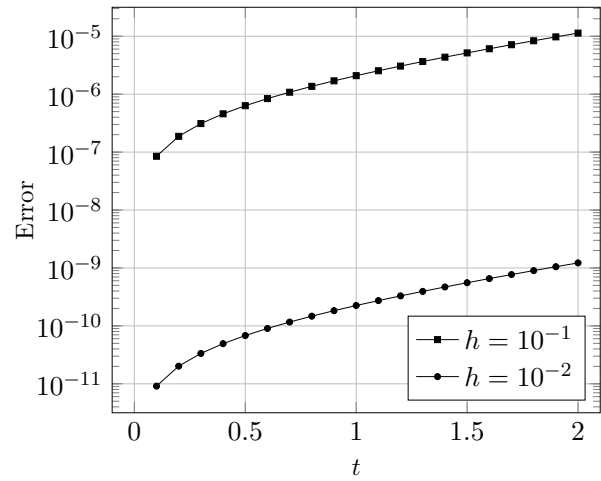
De la misma forma, podemos mostrar que el método que se obtiene calculando

$$\begin{aligned} k_1 &= f(t_0, x_0), \\ k_2 &= f(t_0 + h/2, x_0 + hk_1/2), \\ k_3 &= f(t_0 + h/2, x_0 + hk_2/2), \\ k_4 &= f(t_0 + h, x_0 + hk_3), \end{aligned}$$

y tomando $x_1 = x_0 + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$ es un método de cuarto orden. Consideramos nuevamente el problema del Ejemplo 3.1. En la Figura 3.8 graficamos los errores en función del tiempo, para los métodos Runge-Kutta estudiados con dos pasos temporales distintos: $h = 0.1, 0.01$.



(a) Runge-Kutta de segundo orden.



(b) Runge-Kutta de cuarto orden.

Fig. 3.8: Errores de los métodos de Runge-Kutta para $\dot{x} = x$, $x(0) = 1$, $t \in [0, 2]$.

Ejercicio 3.13. Escribir un programa que implemente el método de Runge-Kutta de segundo y cuarto orden para resolver ecuaciones de la forma

$$\begin{cases} \dot{x}(t) = f(t, x(t)), \\ x(t_0) = x_0, \end{cases}$$

tomando como parámetros la función f , los tiempos inicial t_0 , el intervalo de integración T , el paso h y el dato inicial x_0 ; y arrojando como resultados las listas $t_list = (t_0 \dots t_N)$ y $x_list = (x_0 \dots x_N)$. Utilizar este método para resolver nuevamente el Ejercicio 3.17. Comparar la solución con la obtenida con el método de Euler.

3.5. Problemas en dimensión mayor. Sabemos que existen muchos ejemplos de problemas donde se necesita más de una variable para describir el estado del sistema. La evolución se expresa por (3.1), pero interpretando esta ecuación en forma vectorial, es decir

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t)), \\ \mathbf{x}(t_0) = \mathbf{x}_0, \end{cases}$$

donde $\mathbf{x}(t) = (x_1(t), \dots, x_d(t)) \in \mathbb{R}^d$ representa el estado del sistema en el instante t , $\mathbf{f}(t, \mathbf{x})$ es una función suave definida en $\mathbb{R} \times \mathbb{R}^d$ con valores en \mathbb{R}^d , t_0 es el tiempo y \mathbf{x}_0 el estado inicial del sistema. Los métodos numéricos son los mismos interpretándolos vectorialmente. Los ordenes de convergencia se mantienen, así tenemos que el método de Euler tiene orden $O(h)$ y los métodos de Runge-Kutta presentados en 3.2 y 3.12 tienen orden $O(h^2)$ y $O(h^4)$ respectivamente.

En sistemas de baja dimensión (dos o tres), las soluciones se pueden representar como curvas forma paramétrica donde el parámetro es el tiempo. Los puntos estacionarios, se grafican como puntos y las soluciones periódicas como curvas cerradas. El comportamiento de las trayectorias cerca de los puntos estacionarios se puede analizar mediante el teorema de Hartman-Grobman, estudiando el sistema linealizado. Para obtener el diagrama de fases, se puede resolver numéricamente el sistema con diferentes condiciones iniciales y luego graficar las órbitas paramétricamente.

Ejemplo 3.3. Vamos a estudiar el sistema autónomo $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t))$, con $\mathbf{x} = (x_1, x_2)$ dado

por

$$\begin{cases} \dot{x}_1 = x_1(2 - x_1) - x_1x_2, \\ \dot{x}_2 = x_1x_2 - x_2. \end{cases}$$

Tenemos los puntos estacionarios $(0,0)$, $(2,0)$ y $(1,1)$. La matriz de derivadas parciales es

$$\mathbf{f}'(\mathbf{x}) = \begin{pmatrix} 2 - 2x_1 - x_2 & -x_1 \\ x_2 & x_1 - 1 \end{pmatrix},$$

por lo tanto $\mathbf{f}'(0,0)$ tiene autovalores 2 y -1 con autovectores $(1 \ 0)$ y $(0 \ 1)$, los autovalores de $\mathbf{f}'(2,0)$ son -2 y 1 con autovectores $(1 \ 0)$ y $(0 \ 1)$. Por último, $\mathbf{f}'(1,1)$ tiene autovalores $-1/2 \pm i\sqrt{3}/2$. El diagrama de fases se muestra en la Figura 3.9.

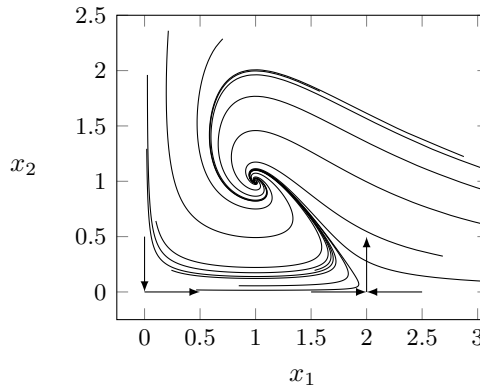


Fig. 3.9: Diagrama de fases

A continuación, mostramos un sistema lineal de ecuaciones diferenciales, donde las variables representan los potenciales eléctricos en puntos de un circuito R-C.

Ejemplo 3.4. Estudiamos un circuito formado por resistores y capacitores conectados a un generador de tensión cosenoidal. las variables representan los potenciales eléctricos en puntos del circuito, las corrientes en cada rama se deduce de la ley de Ohm y las leyes de Kirchhoff. Consideremos el circuito de la Figura 3.10.

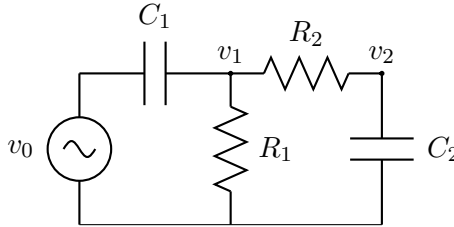


Fig. 3.10: Circuitos RC en cascada.

Aplicando la primera ley de Kirchhoff, obtenemos las ecuaciones

$$\begin{aligned} C_2 \dot{v}_2 &= \frac{v_1 - v_2}{R_2}, \\ C_1(\dot{v}_0 - \dot{v}_1) &= \frac{v_1}{R_1} + \frac{v_1 - v_2}{R_2}, \end{aligned}$$

si definimos los tiempos $\tau_1 = R_1 C_1$, $\tau_2 = R_2 C_2$, $\tau_a = R_2 C_1$, el sistema se escribe como

$$\begin{aligned}\dot{v}_1 &= -\left(\frac{1}{\tau_1} + \frac{1}{\tau_a}\right)v_1 + \frac{1}{\tau_a}v_2 + \dot{v}_0, \\ \dot{v}_2 &= \frac{1}{\tau_2}v_1 - \frac{1}{\tau_2}v_2.\end{aligned}$$

El sistema es un sistema lineal no homogéneo con coeficientes constantes. Si el generador produce una diferencia de potencial $v_0 = \cos(\omega t)$, el problema admite una única solución periódica $\mathbf{v}_{\text{per}}(t) = (v_{\text{per},1}(t) \ v_{\text{per},2}(t))^T$, con $v_{\text{per},j}(t) = \hat{v}_j(\omega) \cos(\omega t + \phi_j(\omega))$, donde $v_1(\omega), v_2(\omega) \in (0, 1)$ y $\phi(\omega) \in (-\pi, \pi]$. Podemos ver numéricamente que para cualquier condición inicial, se verifica

$$\lim_{t \rightarrow +\infty} \mathbf{v}(t) - \mathbf{v}_{\text{per}}(t) = 0.$$

Si $R_1 = 2 \text{ k}\Omega$, $R_2 = 1 \text{ k}\Omega$, $C_1 = 2 \text{ }\mu\text{F}$, $C_2 = 1 \text{ }\mu\text{F}$, obtenemos los gráficos de la Figura 3.11. El

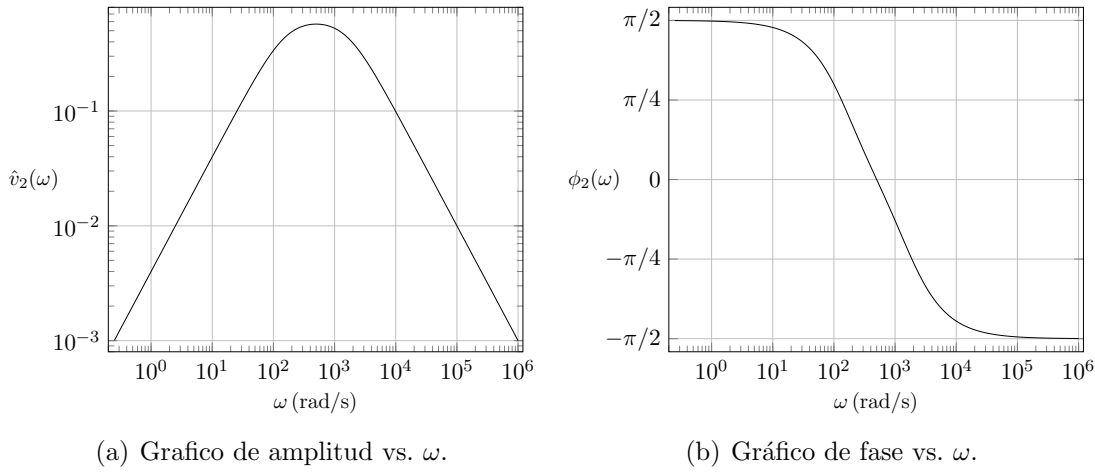


Fig. 3.11: Amplitud y fase de la solución periódica.

máximo de $\hat{v}_2(\omega)$ se alcanza en $\omega = 500 \text{ rad s}^{-1}$, para ese valor $\phi_2(\omega) = 0$.

Ejemplo 3.5 (Circuito R-L-C). Consideramos el circuito R-L-C de la Figura 3.12. Por lo que vimos en ejemplos anteriores, tenemos $Cv_C = q$, $\dot{q} = -i$, $v_L = L di/dt$ y $v_C - v_L = Ri$. De estas relaciones se deduce el sistema lineal

$$\begin{aligned}\frac{dq}{dt} &= -i, \\ \frac{di}{dt} &= \omega^2 q - 2\xi i.\end{aligned}$$

donde $\omega = 1/\sqrt{LC}$ y $\xi = R/(2L)$. En la Figura 3.13(a) mostramos las soluciones para los parámetros $\omega = 2$, $\xi = 1/2$ y las condiciones iniciales $q(0) = 1$, $i(0) = 0$.

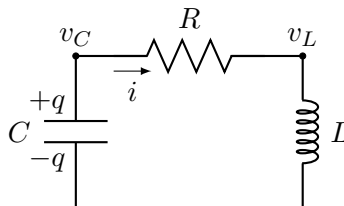
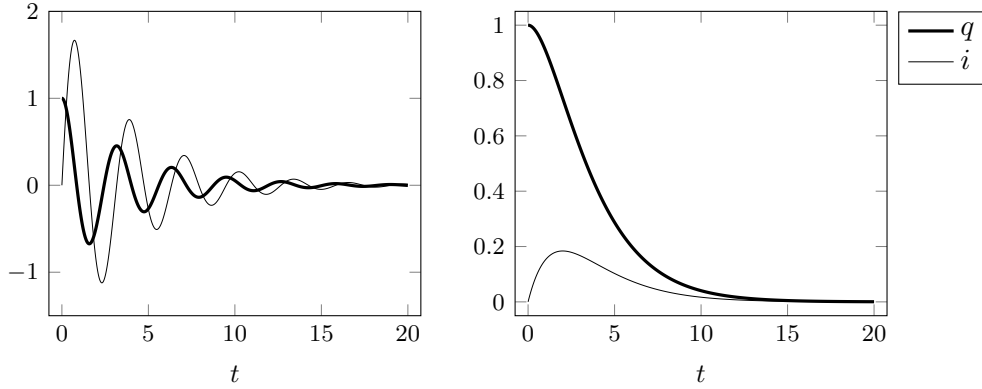


Fig. 3.12: Circuito R-L-C.

En la Figura 3.13(b) mostramos las soluciones para los parámetros $\omega = 1/2$, $\xi = 1/2$ y las mismas condiciones iniciales. Si bien las soluciones convergen asintóticamente a 0 en ambos casos, el comportamiento es diferente. En el primero, lo hacen en forma oscilatoria, mientras que en el último, en forma monótona. Estas diferencias se relacionan con los autovalores de la matriz de coeficientes que valen $\lambda_{1,2} = -\xi \pm \sqrt{\xi^2 - \omega^2}$. En el primer caso, los autovalores son complejos conjugados con parte real negativa, mientras que en el último, son ambos negativos.

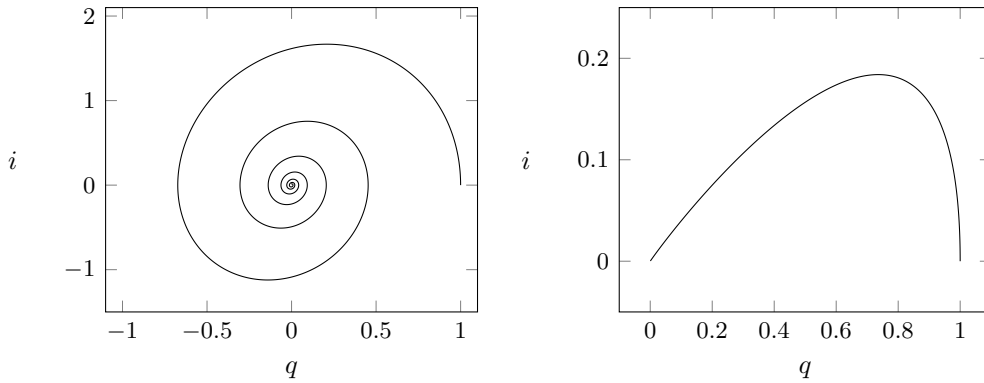


(a) Solución sub-amortiguada.

(b) Solución super-amortiguado.

Fig. 3.13: Soluciones del circuito R-L-C.

En la Figura 3.14 mostramos las trayectorias en el espacio de fases.



(a) Solución sub-amortiguada.

(b) Solución super-amortiguado.

Fig. 3.14: Diagrama de fases de las soluciones del circuito R-L-C.

3.5.1. Ecuaciones de orden superior. La evolución de un sistema físico puede estar determinado por las derivadas de orden superior respecto del tiempo, $\dot{\mathbf{x}}(t), \ddot{\mathbf{x}}(t), \dots$. Generalmente, aunque no siempre, el sistema se escribe como la derivada de mayor orden en función del tiempo y de las derivadas de orden menor:

$$\mathbf{x}^{(k)}(t) = \mathbf{f}(t, \mathbf{x}(t), \dot{\mathbf{x}}(t), \dots, \mathbf{x}^{(k-1)}(t)),$$

con condiciones iniciales $\mathbf{x}(t_0) = \mathbf{x}_0, \dot{\mathbf{x}}(t_0) = \mathbf{x}_1, \dots, \mathbf{x}^{(k-1)}(t_0) = \mathbf{x}_{k-1}$. Este problema se puede reducir al caso anterior. Si $\mathbf{x}(t) \in \mathbb{R}^d$, consideramos el vector $\mathbf{X}(t) \in \mathbb{R}^{kd}$ definido por $\mathbf{X}(t) = (\mathbf{x}(t) \dot{\mathbf{x}}(t) \dots \mathbf{x}^{(k-1)}(t))$, vale entonces

$$\dot{\mathbf{X}}(t) = (\dot{\mathbf{x}}(t) \ddot{\mathbf{x}}(t) \dots \mathbf{x}^{(k)}(t)),$$

por lo tanto $\dot{\mathbf{X}}(t) = \mathbf{F}(t, \mathbf{X}(t))$, donde la función $\mathbf{F}(t, \mathbf{X}(t))$ está dada por

$$\mathbf{F}(t, \mathbf{X}(t)) = (\dot{\mathbf{x}}(t) \ddot{\mathbf{x}}(t) \dots \mathbf{f}(t, \mathbf{x}(t), \dot{\mathbf{x}}(t), \dots, \mathbf{x}^{(k-1)}(t))).$$

Las condiciones iniciales se pueden escribir como $\mathbf{X}(t_0) = (\mathbf{x}_0 \mathbf{x}_1 \dots \mathbf{x}_{k-1})$.

Vamos a ilustrar estas ideas con algunos ejemplos.

Ejemplo 3.6. Para explicar las leyes de Kepler, Isaac Newton estudió el movimiento de una partícula sometida a una fuerza que depende de la posición de la misma. El caso considerado corresponde al de una fuerza central, es decir la fuerza está dirigida a lo largo de la recta que une un centro fijo \mathbf{o} y la posición de la partícula, la magnitud de dicha fuerza sólo depende de la distancia al punto central. A Newton le interesa conocer el movimiento de los planetas alrededor del sol, pero sus resultados cambiaron de manera radical toda la física y la matemática. Si \mathbf{r} es el vector posición de la partícula, la fuerza ejercida sobre la misma está dada por $\mathbf{f}(\mathbf{r}) = \varphi(r)\mathbf{r}$, donde $r = \sqrt{\mathbf{r} \cdot \mathbf{r}}$ y $\varphi(r)$ una función conocida. Por la segunda ley de Newton, si la partícula tiene masa m , las ecuaciones de movimiento son $m\ddot{\mathbf{r}} = \mathbf{f}(\mathbf{r})$.

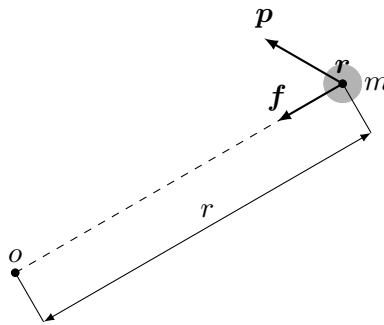


Fig. 3.15: Partícula en un campo central de fuerzas.

Si definimos $\mathbf{p} = m\dot{\mathbf{r}}$, el problema original se puede escribir de la forma

$$\begin{cases} \dot{\mathbf{r}} = \frac{1}{m}\mathbf{p}, \\ \dot{\mathbf{p}} = \mathbf{f}(\mathbf{r}), \end{cases}$$

tomando $\mathbf{X} = (r_1 \ r_2 \ r_3 \ p_1 \ p_2 \ p_3)$

3.6. Aplicaciones.

3.6.1. Modelo depredador-presa (Lotka–Volterra). Cuando estudiamos el crecimiento de una población, tuvimos en cuenta la tasa de crecimiento de la población natural, es decir bajo la condición de recursos ilimitados. También consideramos el caso de recursos que se agotan, pero siempre tomando una sola especie. Vamos a analizar ahora la evolución de poblaciones que interactúan. Para mantener la discusión lo más simple posible vamos a tomar dos especies, la primera que llamaremos depredador, la que se alimenta de la segunda especie, a la que denominaremos presa. Asumimos algunas hipótesis:

- los depredadores tienen como única fuente de alimentación a las presas,
- las presas cuentan con recursos ilimitados,
- la única amenaza de las presas son los depredadores.

Denotamos $x(t)$ e $y(t)$ el número de individuos de predadores y de presas a tiempo t , respectivamente. En ausencia de presas, $y(t) = 0$, $x(t)$ decae con tasa α ; mientras que en ausencia de predadores, $x(t) = 0$, $y(t)$ crece con tasa β . Además, los encuentros entre individuos de cada especie hacen crecer la población de los predadores, mientras que decrece la población de las presas, en forma proporcional al número de encuentros $n(t)$ por unidad de tiempo. Siendo que $n(t)$ es proporcional al número de predadores y al número de presas, vemos que $n(t) = k x(t) y(t)$. De esto modo, se obtiene el sistema Lotka-Volterra:

$$(3.2) \quad \begin{cases} \dot{x}(t) = -\alpha x(t) + \gamma x(t) y(t), \\ \dot{y}(t) = \beta y(t) - \delta x(t) y(t), \end{cases}$$

donde γ y δ poderan el efecto de $n(t)$ en la tasa de crecimiento de predadores y presas, respectivamente. Existen muchos ejemplos que encajan parcialmente en esta descripción: leones y gacelas, aves e insectos, pandas y eucaliptos. Uno de los primeros casos donde se cuentan con datos es el de las liebres raqueta de nieve, Figura 3.18(a)¹, y el lince canadiense, Figura 3.18(b)², en la Figura. De los registros sobre el comercio de pieles en Canadá llevados por la compañía *Hudson Bay Company*³, se pueden extraer datos sobre la población de ambas especies (ver [9], página 216). En la Figura 3.16 mostramos la evolución entre los años 1845 y 1935.

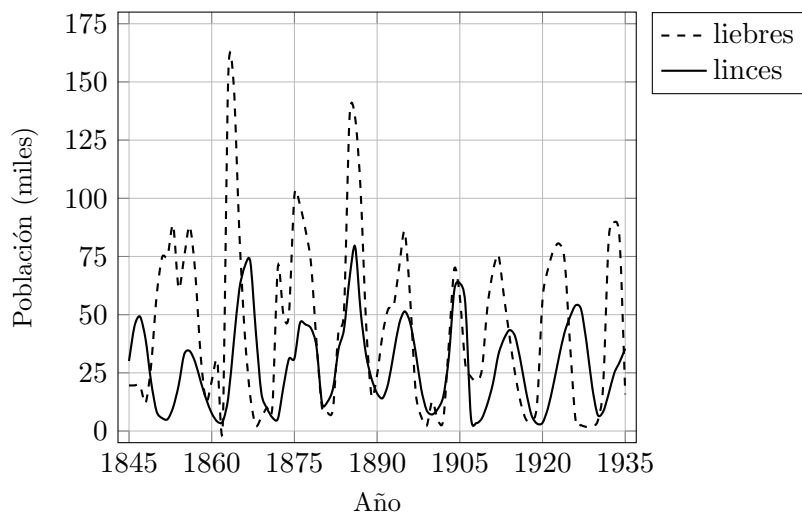


Fig. 3.16: Evolución de las poblaciones de linces canadienses y liebres raqueta de nieve.

Se puede ver analíticamente que el sistema (3.2) presenta dos puntos de equilibrio, uno de ellos es el trivial $x(t) = 0, y(t) = 0$ y el otro $x(t) = \beta/\delta, y(t) = \alpha/\gamma$. El primer cuadrante, $x, y > 0$, es invariante, es decir si $x(t_0) > 0$ e $y(t_0) > 0$, entonces $x(t) > 0$ e $y(t) > 0$ para todo t . La función $V(x, y) = \alpha \ln(y) + \beta \ln(x) - \gamma y - \delta x$ es una cantidad conservada, es decir, para cualquier solución del sistema (3.2), $V(x(t), y(t))$ es constante en el tiempo. Por lo tanto las curvas de nivel de $V(x, y)$ describen las trayectorias en el espacio de fases x, y . Se tiene entonces que las órbitas son cerradas. Dado que el punto de equilibrio $(\beta/\delta, \alpha/\gamma)$ es un máximo de $V(x, y)$, es un equilibrio estable.

En la Figura 3.17 se muestra las soluciones periódicas del sistema (3.2). Claramente, los datos mostrados en la Figura 3.16 no presentan la misma regularidad. Esto puede deberse a muchos factores; por ejemplo las hipótesis del modelo Lotka-Volterra no se verifican completamente, los linces podrían alimentarse de otros animales pequeños, la fuente de alimento de las liebres

¹© 2013, D. Gordon E. Robertson, CC BY SA 3.0.

²© 2010, Keith Williams, CC BY 2.0 <https://commons.wikimedia.org/w/index.php?curid=11394713>.

³<https://github.com/bblais/Systems-Modeling-Spring-2015-Notebooks/tree/master/data>

podrían fluctuar por razones climáticas, etc. Por otro lado, los datos poblacionales se infieren de las pieles comercializadas. Esta actividad humana esta sujeta a cambios, no necesariamente relacionados con la población. Modificaciones en las regulaciones de la caza o variaciones de la demanda⁴ de pieles podrían haber afectado la relación entre el volumen de la actividad comercial y la población de las especies.

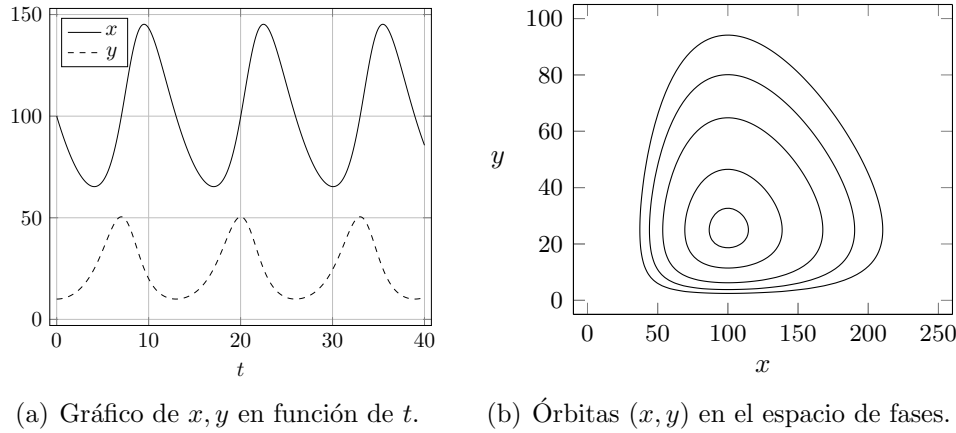


Fig. 3.17: Órbitas del sistema (3.2) para $\alpha = 0.25$, $\beta = 1.0$, $\gamma = \delta = 0.01$.



(a) Liebre raqueta de nieve.



(b) Lince canadiense.

Fig. 3.18: Presa y predador correspondiente a los datos de la Figura 3.16.

3.6.2. Modelo FitzHugh–Nagumo. En 1952, Alan Hodgkin y Andrew Huxley desarrollaron el primer modelo cuantitativo de las corrientes iónicas de Sodio (Na^+) y Potasio (K^+) en el axón gigante de la neurona de calamar, al ser estimulada con una corriente externa. El modelo consiste en un sistema de cuatro ecuaciones diferenciales no lineales, conocido como modelo de Hodgkin y Huxley (HH), el cual describe la dinámica del potencial de membrana de una neurona ante la acción de una corriente aplicada. Este modelo reproduce muchas de las propiedades observadas. Sin embargo, la alta dimensión del sistema dificulta la comprensión en forma cualitativa de los fenómenos involucrados.

En 1961, Richard FitzHugh propuso un modelo simplificado que capturaba el proceso de excitación de células nerviosas. Para esto, utilizó computadoras analógicas (Figura 3.22), es decir circuitos cuyas características eléctricas simulaban el comportamiento de las células neuronales.

⁴El período considerado incluye dos crisis mundiales: la gran depresión (1873-1896) y la crisis del 29 (1929-1939).



Fig. 3.19: Encuentro entre liebre y lince (© Canadian Museum of Nature).

Su trabajo dio como resultado un nuevo modelo simplificado pero efectivo. Al año siguiente, J. Nagumo realizó un circuito electrónico equivalente a partir de diodos tunel (conductancia diferencial negativa). El modelo FitzHugh–Nagumo es un oscilador de relajación que consiste un sistema de ecuaciones diferenciales de dimensión dos que se escriben de la forma:

$$(3.3) \quad \begin{cases} \dot{v}(t) = v(t) - v^3(t)/3 - w(t) + I, \\ \tau \dot{w}(t) = v(t) + a - b w(t), \end{cases}$$

donde v representa el potencial a través de la membrana, I la corriente aplicada (estímulo), w es una variable de recuperación del sistema sin significado biofísico. Los parámetros a, b, τ son adimensionales y positivos. El valor de τ determina la relación entre la velocidad de w en relación con la de v . Debido a que la naturaleza no lineal del sistema, no podemos obtener soluciones en forma analítica. Sin embargo, se puede el comportamiento del sistema realizando algunas simulaciones numéricas. En la Figura 3.20 se muestran las soluciones para un estímulo escalón $I(t) = I_0 H(t - t_0)$, donde I_0 la amplitud del salto, t_0 el tiempo del salto y H la función de Heaviside. Se consideran los parámetros $a = 0.7$, $b = 0.8$, $\tau = 12.5$ y $t_0 = 100$. En la Figura 3.20(a) el estímulo escalón produce un pulso único de gran amplitud casi independiente de I_0 . Cuando el estímulo supera cierto umbral, la solución se comporta en forma oscilatoria (Figura 3.20(b)). Los diagramas de fases se muestran la Figura 3.21, vemos en la Figura 3.21(a) ($I_0 = 0.3$) que la trayectoria parte del estado estable para $I = 0$ y realiza una larga excursión para luego converger al nuevo punto de equilibrio. En la Figura 3.21(b), se observa el ciclo límite atractor.

3.6.3. Atractor de Lorenz. En 1963, Edward Lorenz estudió (ver [14]) un sistema de ecuaciones diferenciales tridimensional no lineal, el cual era un modelo simplificado de la evolución temporal de flujos hidrodinámicos de la atmósfera terrestre. De la observación de los resultados obtenidos numéricamente, concluyó que el sistema no evolucionaba hacia un estado estacionario ni tampoco presentaba un comportamiento periódico. Por el contrario, su dinámica era irregular y parecía cambiar aleatoriamente, a pesar que el sistema era completamente

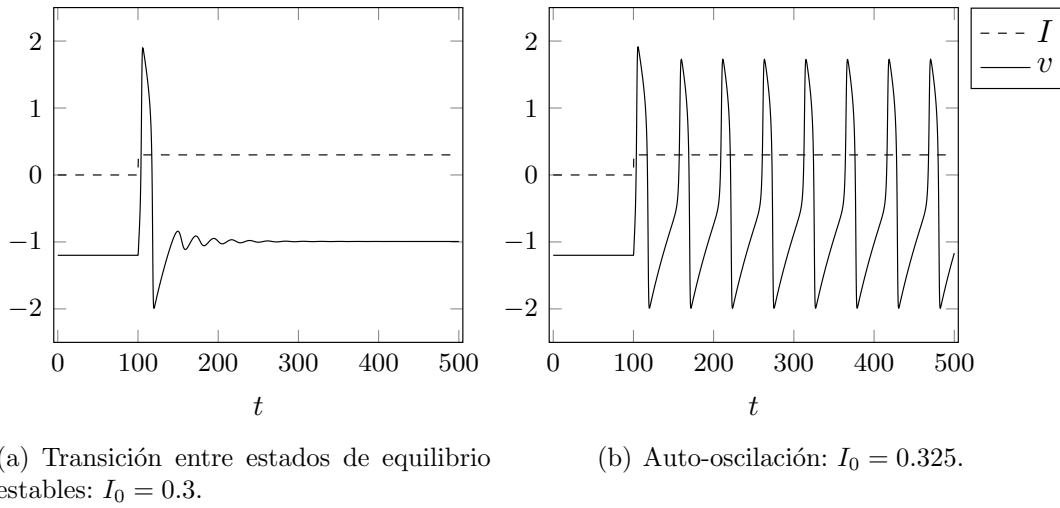


Fig. 3.20: Comportamiento del sistema (3.3): potencial vs. tiempo.

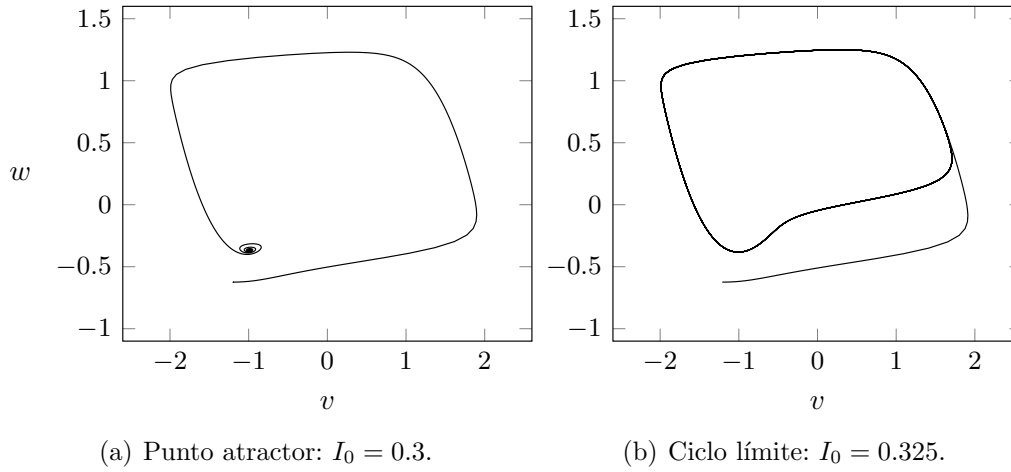


Fig. 3.21: Comportamiento del sistema (3.3): diagrama de fases.

determinístico. El sistema estudiado es

$$\begin{cases} \dot{x}(t) = \sigma x(t) - \sigma y(t), \\ \dot{y}(t) = \rho x(t) - y(t) - x(t)z(t), \\ \dot{z}(t) = -\beta z(t) + x(t)y(t), \end{cases}$$

que para ciertos valores de los parámetros presenta, lo que hoy denominamos, comportamiento caótico⁵. Las trayectorias no convergen a un punto de equilibrio, ni se acercan a un ciclo límite (o a un toro invariante en dimensión mayor que dos). El atractor de este sistema es un conjunto fractal (dimensión de Hausdorff entre dos y tres) y exhibe un comportamiento muy complicado.

3.6.4. Métodos homotópicos para ecuaciones no lineales. En el Capítulo 2 estudiamos la resolución de ecuaciones $f(x_*) = 0$. Vamos a analizar un método diferente a los tratados ahí. La idea es modificar la ecuación lentamente desde una ecuación con solución conocida hasta la ecuación a resolver. Más concretamente, consideramos el problema $f(x(t)) = (1-t)f(x_0)$,

⁵Un sistema dinámico caótico presenta, entre otras características, una alta sensibilidad a variaciones de las condiciones iniciales, es decir que pequeñas variaciones en dichas condiciones iniciales producen diferencias significativas en el comportamiento futuro. Esto impide la predicción a largo plazo (efecto mariposa).

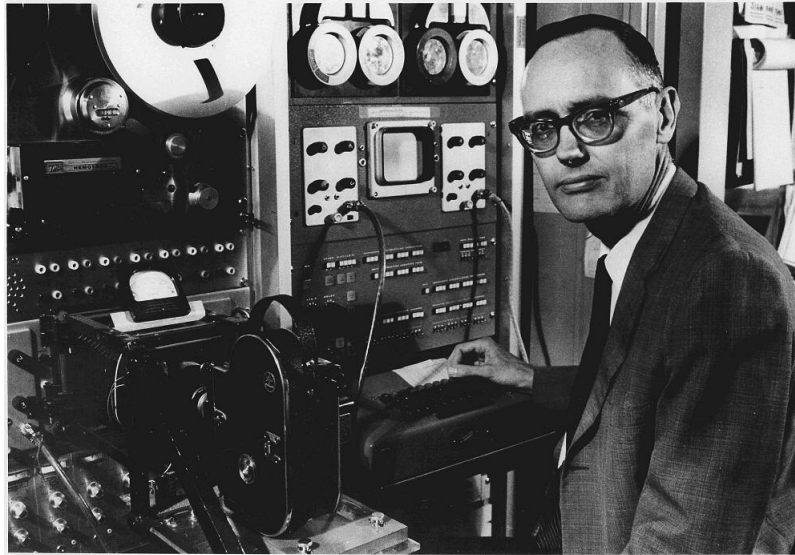


Fig. 3.22: R. FitzHugh y la computadora analógica.

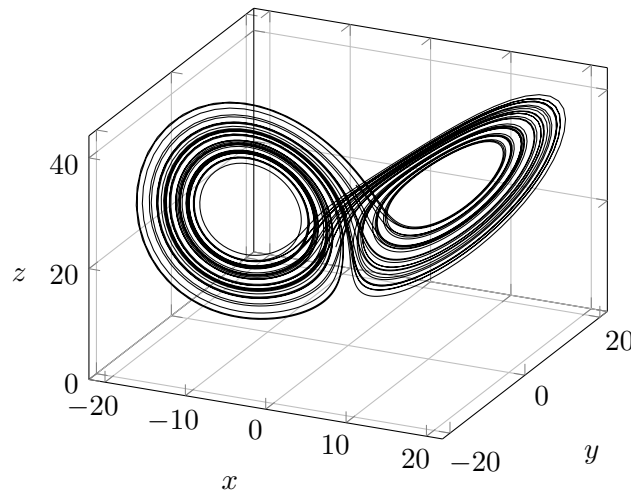


Fig. 3.23: Atractor de Lorenz

con $t \in [0, 1]$ para $t = 0$, x_0 es claramente una solución. Por otro lado, derivando la igualdad con respecto a t , obtenemos

$$\begin{cases} \dot{x}(t) = -f(x_0)/f_x(x(t)), \\ x(0) = x_0. \end{cases}$$

Aplicando los métodos de integración, obtenemos una aproximación de $x(1) = x_*$, solución de la ecuación $f(x_*) = 0$. Si tomamos el ejemplo $e^x = 45$, podemos plantear el problema para $f(x) = e^x - 45$ y $x_0 = 0$:

$$\begin{cases} \dot{x}(t) = 44e^{-x(t)}, \\ x(0) = 0. \end{cases}$$

Con el método de Runge-Kutta de orden 4 y paso $h = 0.01$, obtenemos $x(1) = 3.8066641621$, el error es $|\ln(45) - x(1)| = 1.67232 \times 10^{-6}$.

El método anterior se puede generalizar, si consideramos una función $F(t, x)$ que verifique $F(1, x) = f(x)$, $F(0, x) = g(x)$, podemos plantear $F(t, x(t)) = 0$. Si $g(x_0) = 0$ es fácil de

resolver (en forma analítica o numérica), resolviendo

$$\begin{cases} \dot{x}(t) = -\frac{F_t(t, x(t))}{F_x(t, x(t))}, \\ x(0) = x_0, \end{cases}$$

obtenemos $x(1) = x_*$ solución de $f(x_*)$.

Para el problema $1.75x^3 - 3x - 1 = 0$, planteamos

$$F(t, x) = (1 - t)(x^3 - x) + t(1.75x^3 - 3x - 1)$$

vemos que $F(1, x) = 1.75x^3 - 3x - 1$ y $F(0, x) = x^3 - x$. Las soluciones de $F(0, x) = 0$ son $x_0 = 0, \pm 1$. Podemos resolver el problema original integrando la ecuación

$$\begin{cases} \dot{x}(t) = \frac{-0.75x^3 + 2x + 1}{3(1 + 0.75t)x^2 - 2t - 1}, \\ x(0) = -1, 0, 1, \end{cases}$$

obteniendo $x(1) = x_* = -1.09114, -0.360711, 1.45185$ las soluciones de $1.75x^3 - 3x - 1 = 0$. Observemos que si la ecuación cúbica fuese $1.75x^3 - 3x - 2 = 0$, cuyas soluciones son $x_* = 1.5637, -0.781849 \pm i0.345803$, un método similar nos daría una raíz partiendo de $x_0 = 1$, pero no convergería para $x_0 = -1, 0$. Esto se debe a que las raíces son complejas.

3.6.5. Teoría del cable. La evidencia experimental de la importancia de la teoría del cable en el modelado de los axones nerviosos reales comenzó a aparecer en la década de 1930 a partir de los trabajos de Cole, Curtis, Hodgkin, Katz, Rushton, Tasaki y otros. En esta época fueron muy importantes los papeles de Davis y Lorente de No (1947) y los de Hodgkin y Rushton (1946). En la década de 1950 mejoraron en las técnicas para medir la actividad eléctrica de neuronas individuales. Así, la teoría del cable llegó a ser importante para el análisis de los datos recogidos a partir de grabaciones de microelectrodos intracelulares y para el análisis de las propiedades eléctricas de las dendritas neuronales. Científicos como Coombs, Eccles, Fatt, Frank, Fuortes y otros se basaron en gran medida en la teoría de cables para obtener mayor conocimiento sobre el funcionamiento de las neuronas y para orientarse en el diseño de nuevos experimentos.

Más tarde, la teoría del cable con sus derivados matemáticos permitió modelos neuronales cada vez más sofisticados para ser explorados por investigadores como Jack, Christof Koch, Noble, Poggio, Rall, Redman, Rinzel, Idan Segev, Shepherd, Torre y Tsien. Una importante vertiente de investigación se centró en analizar los efectos de las diferentes distribuciones de entrada sináptica en la superficie dendrítica de una neurona.

Vamos a considerar el siguiente problema: un cable coaxial está formado por dos conductores concéntricos, uno central (axoplasma) y uno exterior (fluido externo). Entre ambos se encuentra una capa aislante (membrana). En una punta del cable se conecta un generador de tensión entre ambos conductores, en la otra punta se conectan a través de un resistor (Figura 3.24). El

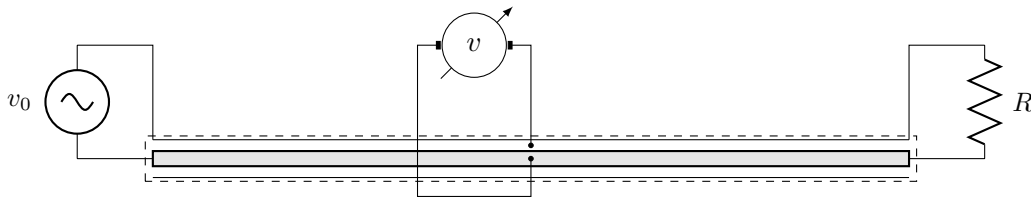


Fig. 3.24: Esquema de las conexiones.

comportamiento cuantitativo del cable está determinado por cuatro constantes eléctricas:

- r_a : resistencia eléctrica por unidad de longitud del axoplasma ($\Omega \text{ cm}^{-1}$),
- r_f : resistencia eléctrica por unidad de longitud del fluido externo ($\Omega \text{ cm}^{-1}$),
- g_m : conductividad eléctrica por unidad de longitud de la membrana ($\Omega^{-1} \text{ cm}^{-1}$),
- c_m : capacidad eléctrica por unidad de longitud de la membrana (F cm^{-1}).

Para estudiar el problema, discretizamos el cable en N secciones y consideramos que cada tramo se modela como se muestra en la Figura 3.25. En cada sección definimos $v_j^{(f)}, v_j^{(a)}$ los potenciales

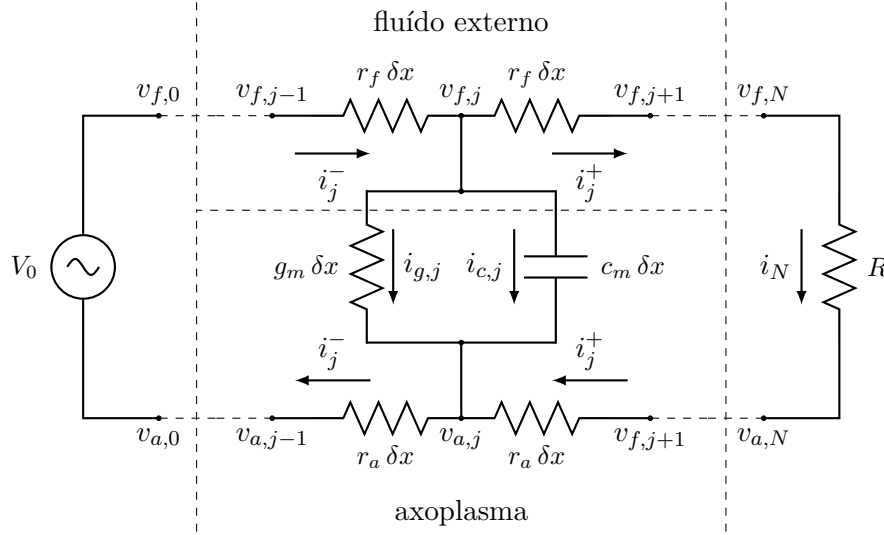


Fig. 3.25: Modelo de transmisión de señales a través de fibras nerviosas.

eléctricos correspondientes al punto j , en el fluido y en el axoplasma, respectivamente. Se puede ver que las corrientes en cada conductor son iguales. En los nodos (f, N) y (a, N) , por la primera ley de Kirchhoff, se verifica la igualdad $i_{N-1}^+ = i_N^- = i_N^+ + i_{g,N} + i_{c,N}$. En forma inductiva podemos ver que en cada sección, la corriente que circula por el fluido externo y por el axoplasma es la misma. Usando la ley de Ohm, vemos que

$$\begin{aligned} v_{f,j-1} - v_{f,j} &= r_f \delta x i_j^-, & v_{a,j} - v_{a,j-1} &= r_a \delta x i_j^-, \\ v_{f,j} - v_{f,j+1} &= r_f \delta x i_j^+, & v_{a,j+1} - v_{a,j} &= r_a \delta x i_j^+, \end{aligned}$$

entonces, si definimos $v_j = v_{f,j} - v_{a,j}$, tenemos

$$v_{j+1} - 2v_j + v_{j-1} = (r_f + r_a) \delta x (i_j^+ - i_j^-).$$

De la primera ley de Kirchhoff se deduce $i_j^- - i_j^+ = i_{g,j} + i_{c,j}$, usando que $i_{g,j} = g_m \delta x v_j$ y $i_{c,j} = c_m \delta x \dot{v}_j$, obtenemos para $j = 1, \dots, N-1$


$$v_{j-1} - 2v_j + v_{j+1} = \delta x^2 ((r_f + r_a) g_m v_j + (r_f + r_a) c_m \dot{v}_j),$$

En los extremos se verifica $v_0 = V_0$ y

$$R v_{N-1} - (R + (r_f + r_a) \delta x) v_N = 0.$$

Si L es la longitud del cable, entonces $\delta x = L/N$. Si llamamos


3.7. Ejercicios.

 **Ejercicio 3.1.** Dada la siguiente ecuación diferencial:

$$\begin{cases} \dot{x}(t) = 2x(t) - 5\sin(t), \\ x(0) = 1, \end{cases}$$

cuya solución exacta es la función $x(t) = 2\sin(t) + \cos(t)$.

- (a) Escribir la iteración del método de Euler para esta ecuación.
- (b) Calcular el error de truncamiento.
- (c) ¿Qué paso h debe elegirse para que el error al estimar $x(\pi/2)$ sea menor que 10^{-2} ?


 **Ejercicio 3.2.** Graficar simultáneamente en la región $[0, 10] \times [0, 10]$ las soluciones que se obtienen del problema de valores iniciales


$$\begin{cases} \dot{x}(t) = (x(t) - 5) \cdot (\cos^2(t) - 0.5), \\ x(0) = k, \end{cases}$$

al utilizar el método de Euler con paso $h = 0.01$ para $k = 0, 1, \dots, 10$.

 **Ejercicio 3.3.** Considerar el problema $\dot{x}(t) = -2tx(t)$, $x(0) = 1$, con $t \geq 0$.

- (a) Determinar una cota, en términos de h , para el error cometido si se usa el método de Euler para calcular $x(1)$.
- (b) ¿Cómo debería tomar h si se desea que el error cometido sea menor que 10^{-2} ?
- (c) Calcular la solución en $t = 1$ usando el valor de h obtenido en el ítem previo, y verificar las estimaciones previstas comparando con la solución exacta.

 **Ejercicio 3.4.** Escribir un programa que implemente el método de Euler explícito descrito en el Algoritmo 3.1 para resolver el problema (3.1).

 **Ejercicio 3.5.** Se quiere verificar numéricamente el orden de convergencia de los métodos de Euler y Taylor de orden 2. Para ello: resolver numéricamente el problema

$$\begin{cases} \dot{x}(t) = x(t), \\ x(t_0) = 1, \end{cases}$$

en el intervalo $[0, 1]$ con ambos métodos, tomando $h = 0.1, 0.0625, 0.05, 0.025, 0.01$. Para cada h , calcular el error que se comete al aproximar $x(1)$: $E_N = |x(1) - x_N|$. Graficar $\log(E_N)$ en función de $\log(h)$. ¿Qué se espera ver? ¿El resultado es consistente con el esperado?

 **Ejercicio 3.6.** Considerar el problema

$$\begin{cases} \dot{x}(t) = \lambda x(t), \\ x(0) = x_0, \end{cases}$$


- (a) Verificar que el método de Euler con paso h genera la sucesión $x_n = (1 + \lambda h)^n x_0$, para $n = 1, \dots, N$.
- (b) Para $\lambda < 0$, determinar para qué valores de h ocurre que $x_n \rightarrow 0$ cuando $n \rightarrow \infty$. Comparar con la solución exacta.

- (c) Resolver usando el programa del Ejercicio 3.4 para distintos valores de $\lambda = 1, 10, 50, 100$ y comparar con la solución exacta. ¿Qué sucede?
- (d) Repetir los items anteriores considerando el método de Euler implícito. ¿Qué se observa?

 **Ejercicio 3.7.** Considerar la ecuación:

$$\begin{cases} \dot{x}(t) = t^{-1} \exp(-x(t)), \\ x(1) = 0, \end{cases}$$


- (a) Probar que $0 \leq x(t) \leq t$ para $t \geq 1$.
- (b) Escribir la iteración dada para esta ecuación por el método de Euler. Probar que la solución numérica resultará creciente.
- (c) Calcular el error de truncado del método de Euler aplicado a la ecuación.
- (d) Dar un valor de paso h que garantice que el error de la estimación numérica de $x(2)$ sea menor que 10^{-3} .

 **Ejercicio 3.8.** Modificar el programa del Ejercicio 3.4 para que acepte ecuaciones vectoriales, la solución \mathbf{x} deberá ser una matriz de $d \times N$, donde d la cantidad de variables del problema y N es el número de pasos temporales. De este modo, $\mathbf{x}_{j,n} = x_j(t_n)$.

 **Ejercicio 3.9.** Considerar el método de Euler modificado:

$$x_n = x_{n-1} + hf(t_{n-1} + h/2, x_{n-1} + h/2f(t_{n-1}, x_{n-1})),$$

probar que el error de truncamiento es $O(h^3)$. ¿Qué ventaja presenta este método respecto del método de Taylor de segundo orden?

 **Ejercicio 3.10.** Probar que si definimos $k_1 = f(t_{n-1}, x_{n-1})$, $k_2 = f(t_{n-1} + \alpha h, x_{n-1} + \alpha h k_1)$ y $x_n = x_{n-1} + h((1 - \beta)k_1 + \beta k_2)$, el error de truncamiento es $O(h^3)$ cuando $\alpha\beta = 1/2$. El método $\alpha = 1$, $\beta = 1/2$ se denomina método de Heun.

 **Ejercicio 3.11. Galileo:** Leer el siguiente párrafo:

– Pero, Simplicio, tengo la esperanza de que no seguirás el ejemplo de muchos otros que desvían la discusión de un punto principal y dicen que algunas de mis afirmaciones se apartan de la verdad por un cabello, y por este cabello esconden las faltas de otras teorías tan gruesas como un cable de navío. Aristóteles dice que ‘una esfera de hierro de 100 libras, cayendo desde una altura de 100 codos, llega a la tierra antes que una bola de 1 libra haya caído un simple codo’. Yo digo que las dos llegan al mismo tiempo. Tú encuentras, al hacer la experiencia, que la más pesada adelanta a la más ligera en 2 ó 3 dedos; ahora, no puedes esconder detrás de estos dos dedos los 99 codos de Aristóteles, ni puedes mencionar mi error y, al mismo tiempo, pasar en silencio el tuyo, mucho mayor.

Salviati, en *Diálogo sobre dos nuevas ciencias* - Galileo Galilei.

Viviani, estudiante de Galileo, afirma que su maestro realizó efectivamente el experimento descrito en el párrafo anterior, arrojando desde lo alto de la torre de Pisa una bala de cañón y una bala de mosquete. El objetivo de este ejercicio es reproducir numéricamente la experiencia de Galileo.

La posición de un objeto en caída libre puede modelarse con la ecuación:

$$(3.4) \quad m\ddot{x}(t) = \gamma\dot{x}^2(t) - mg$$

siendo x la altura, m la masa del cuerpo, $g = 9.81 \text{ m s}^{-2}$ la aceleración gravitatoria y γ una constante que representa el rozamiento con el fluido en que se produce la caída. Deben darse condiciones sobre la altura y la velocidad iniciales.

La Torre de Pisa mide 55.8 m. La masa de una bala de cañón es de 16 kg, y la de una bala de mosquete 8.2 g. Las constantes de rozamiento para cada bala son: $\gamma_c = 5.8 \times 10^{-3}$ y $\gamma_m = 3.74 \times 10^{-5}$, respectivamente (la diferencia se debe a la diferencia de tamaños).

Implementar un programa llamado **galileo** para obtener la dinámica de la caída de ambas balas utilizando el método de Euler modificado, y graficar, en una misma figura, la posición de cada bala en función del tiempo. A partir de los resultados obtenidos, responder:

- ¿Cuánto tiempo tarda cada bala en tocar el suelo?
- Modificar el programa para que se detenga en el momento en que la bala cañón alcanza el suelo. ¿Cuán lejos del piso está la bala de mosquete?

Nota: No debe cometerse el mismo error que Simplicio al juzgar los resultados. La bala de cañón es alrededor de 2000 veces más pesada que la de mosquete. Consecuentemente, Aristóteles hubiese pronosticado que al llegar la bala de cañón al piso, la de mosquete habría descendido apenas 2 cm.

▣ **Ejercicio 3.12.** Implementar un programa que resuelva sistemas de la forma:

$$\begin{cases} \dot{x}(t) = f(t, x(t)), \\ x(t_0) = x_0, \end{cases}$$

utilizando el método de Runge Kutta de orden 4 dado por:

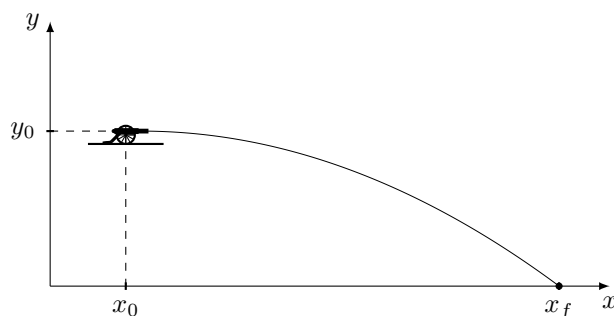
$$x_n = x_{n-1} + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

donde:

$$\begin{aligned} k_1 &= f(t_{n-1}, x_{n-1}), \\ k_2 &= f(t_{n-1} + h/2, x_{n-1} + h/2k_1), \\ k_3 &= f(t_{n-1} + h/2, x_{n-1} + h/2k_2), \\ k_4 &= f(t_{n-1} + h, x_{n-1} + hk_3). \end{aligned}$$

Utilizar este método para resolver nuevamente el Ejercicio 3.17. Comparar la solución con la obtenida con el método de Euler.

🔗 **Ejercicio 3.13. Tiro oblicuo:** Un proyectil de masa m se arroja desde un punto del plano (x_0, y_0) , con una velocidad inicial dada por el vector (v_0^x, v_0^y) .



La trayectoria del proyectil se rige por las ecuaciones dadas por la segunda ley de Newton:

$$\begin{cases} m\ddot{x}(t) = -\gamma\dot{x}(t) \\ m\ddot{y}(t) = -mg - \gamma\dot{y}(t), \end{cases}$$

donde g es la aceleración gravitatoria $g = 9.81 \text{ m s}^{-2}$, y γ es una constante de rozamiento con el medio en que se realiza el lanzamiento. Formular el problema en forma de sistema de orden uno.

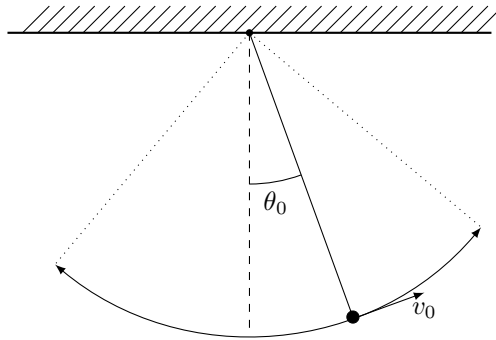
Tomando $m = 10 \text{ kg}$ y $\gamma = 0.2 \text{ kg s}^{-1}$, y suponiendo que el proyectil se lanza desde una altura de 30 m con una velocidad inicial horizontal de 40 m s^{-1} , ¿qué distancia recorre antes de tocar el piso?

Hacer un programa que permita responder esta pregunta, utilizando el método de Euler modificado para resolver el sistema.

 **Ejercicio 3.14. Péndulo:** Se considera el problema del péndulo


$$\begin{cases} \ddot{\theta}(t) = -A \sin(\theta(t)) \\ \theta(0) = \theta_0 \\ \dot{\theta}(0) = v_0 \end{cases}$$

donde θ representa el ángulo que forma la vara del péndulo con la vertical.



- Formular el problema como un sistema de ecuaciones de primer orden.
- Utilizar el método de Euler modificado, con paso $h = 0.05$ para obtener una aproximación de la solución en $[0, T]$ y graficarla.
- Graficar la solución que se obtiene al utilizar método de Runge-Kutta del Ejercicio 3.12.

Pueden considerarse, a modo de ejemplo, los valores $A = 7$, $T = 10$, $\theta_0 = \pi/4$, $v_0 = 0$.

 **Ejercicio 3.15. Oscilador no lineal:** Dada la ecuación $\ddot{x}(t) = -2x^3(t) + x(t)$

- Formular el problema como un sistema de ecuaciones de primer orden.
- Utilizar el método de Runge-Kutta de cuarto orden para obtener las soluciones correspondientes a las condiciones iniciales $x(0) = -2, -1.9, \dots, 1.9, 2$ y $\dot{x}(0) = 0$.
- Graficarlas en el diagrama de fases.
- Graficar la cantidad $\mathcal{H}(t) = \dot{x}^2(t) + x^4(t) - x^2(t)$ para cada solución.
- Obtener el período T en cada caso y graficar T vs. \mathcal{H} .

▣ **Ejercicio 3.16.** Obtener el sistema de ecuaciones del circuito de la Figura 3.26.

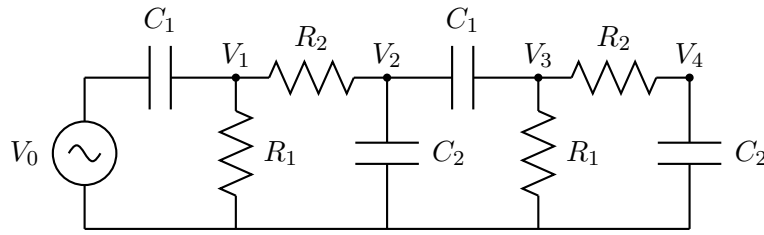


Fig. 3.26

Resolver numéricamente para $R_1 = 2 \text{ k}\Omega$, $R_2 = 1 \text{ k}\Omega$, $C_1 = 2 \text{ }\mu\text{F}$, $C_2 = 1 \text{ }\mu\text{F}$ y $V_0 = \cos(\omega t)$. Mostrar que existen $\hat{V}_4(\omega) \in (0, 1)$ y $\phi(\omega) \in (-\pi, \pi]$ tal que $\lim_{t \rightarrow \infty} V_4(t) - \hat{V}_4(\omega) \cos(\omega t + \phi(\omega)) = 0$.

▣ **Ejercicio 3.17. Sistema predador-presa:** Dado el sistema (3.2):

- Dar condiciones sobre los parámetros y los niveles de x e y que garanticen la estabilidad de las poblaciones. Es decir, que $x(t) = x(t_0)$ e $y(t) = y(t_0)$ para todo $t > t_0$.
- Elegir valores de α , β , γ , δ , x_0 e y_0 que satisfagan las condiciones del ítem anterior y resolver. Realizar dos gráficos: uno de x e y en función de t (simultáneamente) y otro de las trayectorias $(x(t), y(t))$. ¿Se mantiene constante la solución?
- Tomando $\alpha = 0.25$, $\beta = 1$, $\gamma = \delta = 0.01$, $x_0 = 80$ e $y_0 = 30$, resolver y realizar gráficos como los del ítem anterior.
- Modificar $\tilde{\delta} = \lambda\delta$, $\tilde{x}_0 = \lambda^{-1}x_0$ con $\lambda > 0$ (las otras condiciones invariantes) y comparar la solución obtenida con la del ítem anterior.
- Observar los cambios en la solución si se modifica los parámetros en la forma $\tilde{\alpha} = \lambda\alpha$, $\tilde{\beta} = \lambda\beta$, $\tilde{\gamma} = \lambda\gamma$, $\tilde{\delta} = \lambda\delta$, para $\lambda > 0$.

▣ **Ejercicio 3.18. FitzHugh–Nagumo:** Dado el sistema (3.3), integrar las ecuaciones para los valores de parámetros $a = 0.7$, $b = 0.8$ y $\tau = 12.5$ e $I(t) = I_0 H(t - t_{\text{on}})$, donde H es la función Heaviside, $t_{\text{on}} = 100$, $I_0 = 0, 0.1, \dots, 1$ en el intervalo $[0, 500]$. Determinar los valores del estímulo para los cuales el sistema presenta un comportamiento oscilatorio (ciclo límite). Graficar $v(t)$ en función del tiempo y la trayectoria $(v(t), w(t))$ en el espacio de fases.

▣ **Ejercicio 3.19. Atractor de Rössler:** Dado el sistema

$$\begin{cases} \dot{x}(t) = -y(t) - z(t), \\ \dot{y}(t) = x(t) + ay(t), \\ \dot{z}(t) = -b + z(t)(x(t) - c), \end{cases}$$

obtener las soluciones para $x(0) = y(0) = z(0) = 0$, $a = b = 0.1$ y $c = 4, 6, 8, 8.5, 8.7, 12, 14$, en el intervalo $[0, 1000]$. Graficar las soluciones proyectadas en el plano x, y . Para $c = 14$, graficar las soluciones en el espacio x, y, z .

CAPÍTULO 4

Cadenas de Markov

4.1. Introducción. En este capítulo estudiamos la evolución con tiempo discreto de sistemas con un número finito de estados posibles. Vamos a suponer que, contrario a los problemas analizados en el capítulo anterior, la ley que rige las transiciones entre estados es aleatoria, es decir que el estado actual no determina en forma unívoca la transición futura. Si no quisieramos entrar en cuestiones de teoría de probabilidades, podríamos pensar que tenemos un número grande de sistemas iguales e independientes (sin interacción entre ellos) y observamos las evoluciones de estos. De cada estado inicial habrá un porcentaje que evoluciona hacia los distintos estados. Por la Ley de los Grandes Números, esos porcentajes reflejan aproximadamente la probabilidad de transición de un estado a otro.

Ejemplo 4.1 (Movimientos migratorios). Para ilustrar el comentario anterior consideremos un problema de migración humana interterritorial. En el artículo [10] se estudian los movimientos migratorios entre los distintos municipios de España. Cada municipio se clasifica en una de las categorías indicadas en la Tabla 4.1 (Cuadro 1 en [10]). De toda la población migrante (que

Cat.	Tamaño	(en miles de habitantes)
1	$0 < N \leq 10$	
2	$10 < N \leq 20$	
3	$20 < N \leq 50$	
4	$50 < N \leq 100$	
5	$N > 100$	No capital de provincia
6		Capital de provincia

Tabla 4.1: Clasificación de las ciudades según el tamaño.

representa alrededor del 2% de la población total) se la clasifica según la categoría de la ciudad de origen y la de destino. Alguien que migrara de Sevilla a Barcelona¹ tendría como origen y destino la categoría 6, pero aquellos que permanecen residiendo en Sevilla no son considerados. En base a datos del INE², las autoras establecen los siguientes porcentajes de migraciones:

$$(4.1) \quad \begin{pmatrix} 0.278 & 0.244 & 0.218 & 0.203 & 0.204 & 0.288 \\ 0.121 & 0.137 & 0.122 & 0.118 & 0.111 & 0.128 \\ 0.134 & 0.152 & 0.157 & 0.171 & 0.137 & 0.178 \\ 0.083 & 0.100 & 0.110 & 0.121 & 0.093 & 0.113 \\ 0.070 & 0.080 & 0.078 & 0.080 & 0.121 & \textcolor{blue}{0.102} \\ 0.314 & 0.287 & 0.315 & 0.307 & \textcolor{red}{0.334} & 0.191 \end{pmatrix}$$

¹ Al momento de escribir estas notas, Cataluña sigue siendo una provincia de España.

² Instituto Nacional de Estadística, España.

La columna indica la categoría de origen y la fila la categoría de destino. Por ejemplo, el número en rojo establece que el %33.4 (fila 6, columna 5) de los habitantes que migran desde grandes centros urbanos (categoría 5), lo hacen hacia las capitales de provincia (categoría 6). Por otro lado, el número en azul, %10.2 (fila 5, columna 6), muestra el porcentaje del total de la población migrante desde las capitales de provincia que van hacia los grandes centros urbanos.

Obviamente, los porcentajes pueden variar de año a año, de hecho es este punto precisamente lo que se discute en [10]. Pero asumiendo que se mantienen constantes, podríamos predecir a partir de una distribución inicial de la población, como varían cada año las corrientes migratorias. Por ejemplo, si de las zonas rurales (categoría 1) salieran %20.9 del total de migrantes, se distribuirían de la siguiente forma:

Cat.	Porcentaje de migrantes
1	$0.278 \times \%20.9 = \%5.8$
2	$0.121 \times \%20.9 = \%2.5$
3	$0.134 \times \%20.9 = \%2.8$
4	$0.083 \times \%20.9 = \%1.7$
5	$0.070 \times \%20.9 = \%1.5$
6	$0.314 \times \%20.9 = \%6.6$
Total	%20.9

Podríamos calcular de forma similar lo que ocurre con las otras categorías y sumar las contribuciones. Para fijar ideas, consideremos que la población migrante inicial se distribuye como se indica en la columna $t = 0$ de la Tabla 4.2, en las siguientes columnas se muestran sus variaciones en los tres años sucesivos.

Categoría	Población de migrantes (%)			
	$t = 0$	$t = 1$	$t = 2$	$t = 3$
1	20.9	25.7	25.3	25.3
2	19.8	12.6	12.4	12.4
3	12.0	15.7	15.6	15.6
4	5.7	10.3	10.2	10.2
5	8.2	8.8	8.6	8.7
6	33.4	26.9	27.9	27.8
Total	100	100	100	100

Tabla 4.2: Evolución de la población migrante en cada año.

Claramente vemos que hay una tendencia al equilibrio, esto no es algo particular de este ejemplo sino un comportamiento general, como discutiremos más adelante. Observemos que el equilibrio alcanzado es dinámico, con esto queremos decir que las corrientes migratorias continúan pero la diferencia entre los que salen y los que llegan es nula. Por lo tanto, las poblaciones se mantienen constantes.

Una de las condiciones principales que vamos a imponer a la evolución de los sistemas es que el comportamiento futuro dependa solamente del estado inicial (y del azar en el caso aleatorio). En algunos casos, la evolución puede depender de la historia, es decir de los estados en períodos anteriores. Sin embargo, si depende de finitos estados pasados, cambiando la definición de nuestro sistema podemos recuperar esta condición. Ilustramos esto en el siguiente ejemplo.

Ejemplo 4.2. En un semáforo del tránsito, funcionando normalmente, las luces se encienden en la secuencia cíclica: $v \rightarrow a \rightarrow r \rightarrow a \rightarrow v \rightarrow \dots$. Si bien la evolución es determinística, conocer que luz está encendida en un instante no determina completamente el estado futuro, debido a que desde a puede evolucionar hacia v o hacia r , dependiendo del estado inmediatamente anterior al actual. Por lo tanto, el modelo planteado no tiene la propiedad requerida de "falta de memoria". Modificando la noción de estado podemos lograr que la evolución del sistema tenga esta condición. Si consideramos el estado completo del sistema como los colores de luces encendidas en el período anterior y en el presente, es decir $\text{estado} = \langle \text{anterior} \rangle \langle \text{actual} \rangle$, tenemos que las transiciones ocurren en la secuencia: $va \rightarrow ar \rightarrow ra \rightarrow av \rightarrow va \rightarrow \dots$

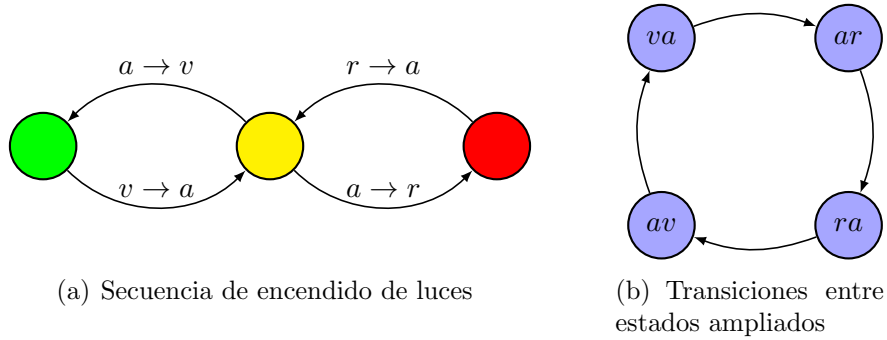


Fig. 4.1: Transiciones del semáforo.

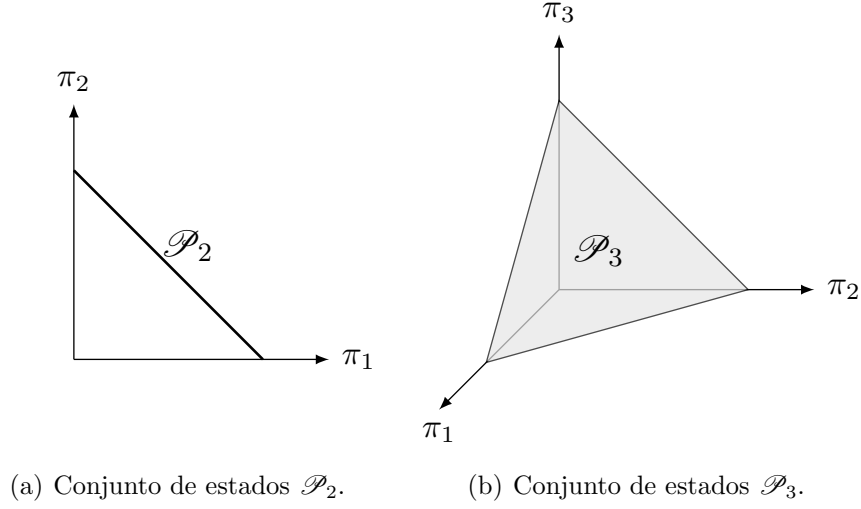
4.2. Estados. En el primer ejemplo, el estado inicial se representa por el vector de proporciones $\mathbf{x}^{(0)} = (0.21 \ 0.20 \ 0.12 \ 0.06 \ 0.08 \ 0.33)^T$ que obviamente es no negativo y sus coordenadas suman 1. Estas dos características se mantienen a lo largo de la evolución, lo que es de esperar tratándose de la distribución de la migración. En el ejemplo del semáforo no es claro como representar los estados en forma vectorial. Podemos asociarlos con los vectores de la base canónica: $va = (1 \ 0 \ 0 \ 0)^T$, $ar = (0 \ 1 \ 0 \ 0)^T$, $ra = (0 \ 0 \ 1 \ 0)^T$, $av = (0 \ 0 \ 0 \ 1)^T$. Vemos que estos vectores también cumplen las condiciones señaladas en el ejemplo anterior.

Vamos a ver que en ambos ejemplos, la evolución se representa en forma sencilla. En el primer caso, la matriz que se muestra en (4.1), que llamaremos P , tiene la información necesaria para calcular el estado siguiente. Concretamente, $\boldsymbol{\pi}^{(1)} = P \cdot \boldsymbol{\pi}^{(0)}$, $\boldsymbol{\pi}^{(2)} = P \cdot \boldsymbol{\pi}^{(1)} = P^{(2)} \cdot \boldsymbol{\pi}^{(0)}$. En general, $\boldsymbol{\pi}^{(n)} = P \cdot \boldsymbol{\pi}^{(n-1)} = P^{(n)} \cdot \boldsymbol{\pi}^{(0)}$. Para el ejemplo del semáforo, podemos considerar la matriz S dada por

$$S = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

tenemos entonces $ar = S \cdot va$, $ra = S \cdot ar$, $av = S \cdot ra$, $va = S \cdot av$.

Consideramos entonces el conjunto de estados $\mathcal{P}_d = \{\boldsymbol{\pi} \in \mathbb{R}_+^d : \pi_1 + \dots + \pi_d = 1\}$ (ver Figura 4.2), que se pueden pensar como una distribución discreta de probabilidades, π_j representa la probabilidad de que el sistema se encuentre en el estado j .

Fig. 4.2: Conjunto de estados \mathcal{P}_d .

La evolución del sistema en general se define por una función $E : \mathcal{P}_d \rightarrow \mathcal{P}_d$. Nosotros nos vamos a restringir a las funciones definidas por una matriz $P \in \mathbb{R}^{d \times d}$, es decir $E(\pi) = P \cdot \pi$. Para que se verifique $\varrho = P \cdot \pi \in \mathcal{P}_d$ para todo $\pi \in \mathcal{P}_d$ es necesario y suficiente que $p_{ij} \geq 0$ y $p_{1j} + \dots + p_{dj} = 1$. En efecto, si $\pi = e_j \in \mathcal{P}_d$ y $\varrho = P \cdot \pi$, entonces $\varrho_i = p_{ij}$ y $\varrho_1 + \dots + \varrho_d = p_{1j} + \dots + p_{dj}$, por lo tanto $p_{ij} \geq 0$ y $p_{1j} + \dots + p_{dj} = 1$. Para ver que es suficiente podemos escribir

$$\begin{aligned} \varrho_1 + \dots + \varrho_d &= (p_{11}\pi_1 + \dots + p_{1d}\pi_d) + \dots + (p_{d1}\pi_1 + \dots + p_{dd}\pi_d) \\ &= (p_{11} + \dots + p_{d1})\pi_1 + \dots + (p_{1d} + \dots + p_{dd})\pi_d = \pi_1 + \dots + \pi_d, \end{aligned}$$

lo que prueba $\varrho_1 + \dots + \varrho_d = 1$. Una matriz que verifica estas dos condiciones se llama matriz de Markov.

4.3. Matrices de Markov. La interpretación probabilística de las propiedades de las matrices de Markov se ilustran en la Figura 4.3.

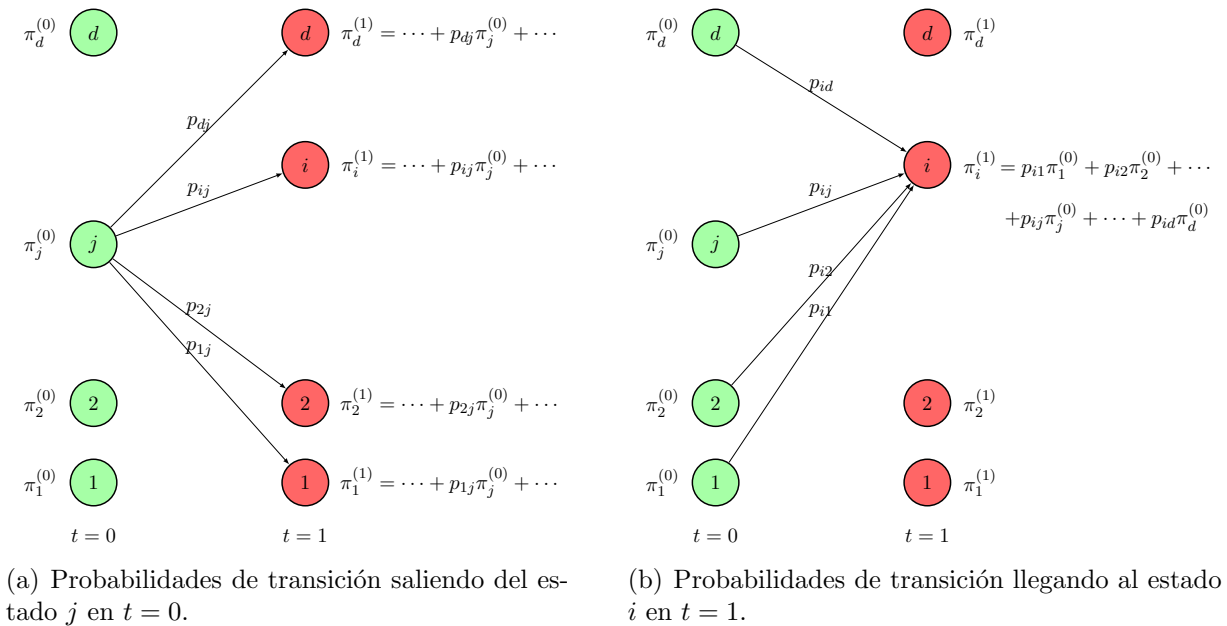


Fig. 4.3: Interpretación probabilística de las matrices de Markov.

El elemento p_{ij} representa la probabilidad de que el sistema pase al estado i en tiempo $t = 1$ sabiendo que estaba en el estado j en tiempo $t = 0$, o en general para tiempos t y $t - 1$. Si pensamos al vector $\boldsymbol{\pi}^{(t)} \in \mathcal{P}_d$ como la función de probabilidades correspondiente al estado del sistema en tiempo t

$$P(\text{el sistema se encuentra en } i \text{ en tiempo } t) = \pi_i^{(t)},$$

entonces por el Teorema de Bayes

$$\pi_i^{(t)} = p_{i1}\pi_1^{(t-1)} + \cdots + p_{ij}\pi_j^{(t-1)} + \cdots + p_{id}\pi_d^{(t-1)},$$

por lo tanto el estado de un sistema evoluciona como $\boldsymbol{\pi}^{(t)} = P \cdot \boldsymbol{\pi}^{(t-1)}$. Inductivamente tenemos que $\boldsymbol{\pi}^{(t)} = P^t \cdot \boldsymbol{\pi}^{(0)}$, o en forma más general $\boldsymbol{\pi}^{(t)} = P^{t-s} \cdot \boldsymbol{\pi}^{(s)}$.

Por lo anterior, vemos que para entender la evolución vamos a tener que estudiar las potencias de la matriz de Markov P . Para eso, necesitamos conocer sus autovalores y autovectores.

4.3.1. Espectro de matrices de Markov. Dada una matriz de Markov P , vamos a estudiar como son sus autovalores y los correspondientes autovectores, lo que se conoce como el análisis espectral de la matriz.

Proposición 4.1. Si P es una matriz de Markov y λ es autovalor de P , entonces $|\lambda| \leq 1$.

Demostración. Si $P \cdot \boldsymbol{\pi} = \lambda \boldsymbol{\pi}$, usando que P es una matriz de Markov

$$|\lambda| |\pi_i| = |\lambda \pi_i| = |p_{i1}\pi_1 + \cdots + p_{id}\pi_d| \leq |p_{i1}\pi_1| + \cdots + |p_{id}\pi_d| = p_{i1}|\pi_1| + \cdots + p_{id}|\pi_d|,$$

sumando i obtenemos

$$\begin{aligned} |\lambda|(|\pi_1| + \cdots + |\pi_d|) &\leq (p_{11}|\pi_1| + \cdots + p_{1d}|\pi_d|) + \cdots + (p_{d1}|\pi_1| + \cdots + p_{dd}|\pi_d|) \\ &\leq (p_{11} + \cdots + p_{d1})|\pi_1| + \cdots + (p_{1d} + \cdots + p_{dd})|\pi_d| \\ &= |\pi_1| + \cdots + |\pi_d|, \end{aligned}$$

como $|\pi_1| + \cdots + |\pi_d| > 0$ vale $|\lambda| \leq 1$. □

Proposición 4.2. Si P es una matriz de Markov, $\lambda \neq 1$ es un autovalor de P y $\boldsymbol{\pi}$ es un autovector asociado, entonces $\pi_1 + \cdots + \pi_n = 0$.

Demostración. Si $\lambda \pi_i = p_{i1}\pi_1 + \cdots + p_{id}\pi_d$, sumando todas las coordenadas π_i obtenemos

$$\begin{aligned} \lambda(\pi_1 + \cdots + \pi_d) &= (p_{11}\pi_1 + \cdots + p_{1d}\pi_d) + \cdots + (p_{d1}\pi_1 + \cdots + p_{dd}\pi_d) \\ &= (p_{11} + \cdots + p_{d1})\pi_1 + \cdots + (p_{1d} + \cdots + p_{dd})\pi_d = \pi_1 + \cdots + \pi_d, \end{aligned}$$

como $\lambda \neq 1$, obtenemos $\pi_1 + \cdots + \pi_n = 0$. □

Recordemos que si $A \in \mathbb{R}^{d \times d}$, $\det(A) = \det(A^T)$, y por lo tanto $f_A(\lambda) = \det(A - \lambda I) = \det(A^T - \lambda I) = f_{A^T}(\lambda)$. Como los polinomios característicos de A y A^T son iguales, los autovalores (incluyendo la multiplicidad) son los mismos.

Proposición 4.3. Si P es una matriz de Markov, $\lambda = 1$ es autovalor.

Demostración. Como $p_{1j} + \cdots + p_{dj} = 1$, tenemos que $P^T \cdot \mathbf{1} = \mathbf{1}$, donde $\mathbf{1} = (1 \ 1 \ \dots \ 1)^T$. Por lo tanto, $\lambda = 1$ es autovalor de P^T con autovector asociado $\mathbf{1}$. Por la observación anterior $\lambda = 1$ es autovalor de P . □

Proposición 4.4. Si P es una matriz de Markov y $\boldsymbol{\pi} \neq \mathbf{0}$ verifica $P \cdot \boldsymbol{\pi} = \boldsymbol{\pi}$, entonces $P \cdot \tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}}$, donde $\tilde{\boldsymbol{\pi}} = (|\pi_1| \ \dots \ |\pi_d|)^T$.

Demostración. Como $\pi_i = p_{i1}\pi_1 + \dots + p_{id}\pi_d$, entonces $|\pi_i| \leq p_{i1}|\pi_1| + \dots + p_{id}|\pi_d|$. Supongamos que alguna de las desigualdades anteriores valiera estrictamente, en ese caso tendríamos

$$\begin{aligned} |\pi_1| + \dots + |\pi_d| &< (p_{11}|\pi_1| + \dots + p_{1d}|\pi_d|) + \dots + (p_{d1}|\pi_1| + \dots + p_{dd}|\pi_d|) \\ &= (p_{11} + \dots + p_{d1})|\pi_1| + \dots + (p_{d1} + \dots + p_{dd})|\pi_d| = |\pi_1| + \dots + |\pi_d|, \end{aligned}$$

lo que es absurdo. Por lo tanto $|\pi_i| = p_{i1}|\pi_1| + \dots + p_{id}|\pi_d|$, lo que prueba $P.\tilde{\pi} = \tilde{\pi}$. \square

Corolario 4.1. Si P es una matriz de Markov, existe $\pi^* \in \mathcal{P}_d$ tal que $P.\pi^* = \pi^*$.

Se puede probar (pero esta fuera del alcance de este apunte) que si λ es autovalor de P con $|\lambda| = 1$, entonces existen tantos autovectores asociados a λ como su multiplicidad. Además, para $\lambda = 1$ se pueden elegir en \mathcal{P}_d .

Supongamos que $\lambda = 1$ es autovalor doble de P , si $\pi^* \in \mathcal{P}_d$ fuera el único autovalor asociado, usando la forma de Jordan, podríamos ver que existe $\varrho \neq 0$ tal que $P.\varrho = \varrho + \pi^*$. Como P es de una matriz de Markov, $\varrho_1 + \dots + \varrho_d = \varrho_1 + \dots + \varrho_d + 1$, lo que resulta absurdo. Por lo tanto, existe otro autovector ϱ asociado a $\lambda = 1$. Queremos ver existe $\varrho^* \in \mathcal{P}_d$, autovector asociado a $\lambda = 1$ independiente de π^* . Consideramos dos casos, primero suponemos que existe $j = 1, \dots, d$ tal que $\varrho_j \neq \pi_j^* = 0$, entonces $\varrho^* \in \mathcal{P}_d$ es independiente de π^* , donde $\varrho^* = (|\varrho_1| + \dots + |\varrho_d|)^{-1}(|\varrho_1| \dots |\varrho_d|)^T \in \mathcal{P}_d$. Si $\pi_j^* \neq 0$ para todo $j = 1, \dots, d$ tal que $\varrho_j \neq 0$, entonces $\hat{\varrho} = \varrho + \beta\pi^*$ verifica $\hat{\varrho}_j = \varrho_j + \beta\pi_j^* \leq 0$ si $\beta > 0$ es suficientemente grande y es independiente de π^* .

4.4. Dinámica.

4.4.1. Estados de equilibrio. Un estado $\pi^* \in \mathcal{P}_d$, es un estado de equilibrio si verifica $P.\pi^* = \pi^*$. En otras palabras, π^* es un autovector asociado al autovalor $\lambda = 1$. Como vimos en las Proposiciones 4.3 y 4.4, existe $\tilde{\pi}$ con $\tilde{\pi}_i \geq 0$ que verifica $P.\tilde{\pi} = \tilde{\pi}$. entonces si definimos $\pi^* = (\tilde{\pi}_1 + \dots + \tilde{\pi}_d)^{-1}\tilde{\pi}$. Tenemos que $\pi^* \in \mathcal{P}_d$ es un estado de equilibrio.

Como el autovalor $\lambda = 1$ puede ser múltiple, el sistema puede tener varios estado de equilibrio como se muestra en el siguiente ejemplo:

Ejemplo 4.3. El sistema mostrado en la Figura 4.4

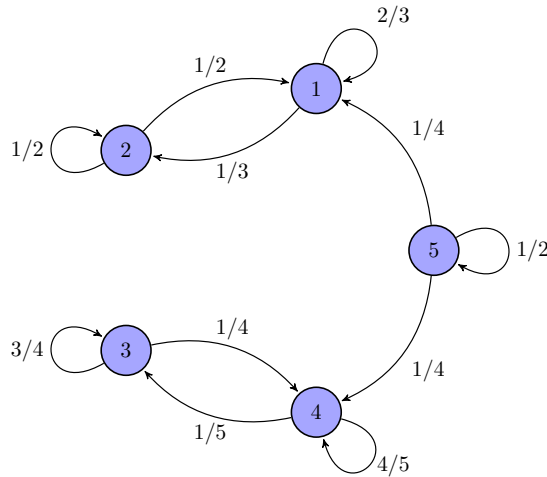


Fig. 4.4

su matriz de transición está dada por

$$P = \begin{pmatrix} 2/3 & 1/2 & 0 & 0 & 1/4 \\ 1/3 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 3/4 & 1/5 & 0 \\ 0 & 0 & 1/4 & 4/5 & 1/4 \\ 0 & 0 & 0 & 0 & 1/2 \end{pmatrix}$$

que tiene $\lambda = 1$ como autovalor doble, los autovectores correspondientes son $\pi^* = (\frac{3}{5} \frac{2}{5} 0 0 0)^T$ y $\pi^{**} = (0 0 \frac{4}{9} \frac{5}{9} 0)^T$, es decir $\pi^*, \pi^{**} \in \mathcal{P}_d$ son estados de equilibrio. Observemos que para $\alpha, \beta \in \mathbb{R}$, $\alpha\pi^* + \beta\pi^{**}$ es también un autovector correspondiente a $\lambda = 1$, pero $\alpha\pi^* + \beta\pi^{**} \in \mathcal{P}_d$ solo si $\alpha, \beta \geq 0$ y $\alpha + \beta = 1$. Los otros autovalores y autovectores son

$$\begin{aligned} \lambda &= \frac{1}{6}, & \pi &= (1 \ -1 \ 0 \ 0 \ 0)^T, \\ \lambda &= \frac{11}{20}, & \pi &= (0 \ 0 \ 1 \ -1 \ 0)^T, \\ \lambda &= \frac{1}{2}, & \pi &= (0 \ -1 \ 4 \ -5 \ 2)^T, \end{aligned}$$

observemos que las coordenadas de los autovectores asociados a $\lambda \neq 1$, suman cero.

4.4.2. Estados límite. Nos preguntamos si dado un estado inicial $\pi^{(0)} \in \mathcal{P}_d$, se verifica $\lim_{t \rightarrow \infty} \pi^{(t)} = \pi^*$. En ese caso, llamamos a π^* estado límite. Como

$$\pi^* = \lim_{t \rightarrow \infty} \pi^{(t+1)} = \lim_{t \rightarrow \infty} P\pi^{(t)} = P\pi^*,$$

todo estado límite es necesariamente un estado de equilibrio. Como ya vimos, los estados de equilibrio no son necesariamente únicos, pero aún en el caso de ser único, no está claro que el sistema evolucione hacia el equilibrio, como se observa en el siguiente ejemplo:

Ejemplo 4.4. La matriz de transición correspondiente al grafo mostrado en la Figura 4.5 es

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

los autovalores son $\lambda = \pm 1$, el autovector correspondiente a $\lambda = 1$ es $\pi = (1/2 \ 1/2)^T$ y para $\lambda = -1$ es $\pi = (1/2 \ -1/2)^T$.

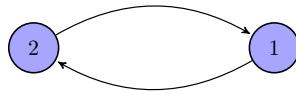


Fig. 4.5

Si tomamos $\pi^{(0)} = (a \ (1-a))^T$, tenemos $\pi^{(t)} = (a \ (1-a))^T$ si t es par y $\pi^{(t)} = ((1-a) \ a)^T$ si t es impar. Vemos que $\pi^{(t)}$ no converge a un estado límite salvo en el caso trivial $\pi^{(0)} = (1/2 \ 1/2)^T$.

4.4.3. Existencia de estados límite. Para estudiar la existencia de estados límite, tenemos que conocer los autovalores y autovectores de la matriz de transición.

4.4.4. Dependencia del estado inicial. El estado límite que se alcanza puede depender del estado inicial.

Ejemplo 4.5. Estudiemos el sistema definido por el grafo de la Figura 4.6(a), su matriz de transición está dada por

$$P = \begin{pmatrix} 1/4 & 11/12 & 0 \\ 3/4 & 1/12 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Los autovalores son $\lambda = 1$ (doble) y $\lambda = -2/3$ y sus autovalores asociados son $(0.55 \ 0.45 \ 0)$, $(0 \ 0 \ 1)$ y $(1 \ -1 \ 0)$.

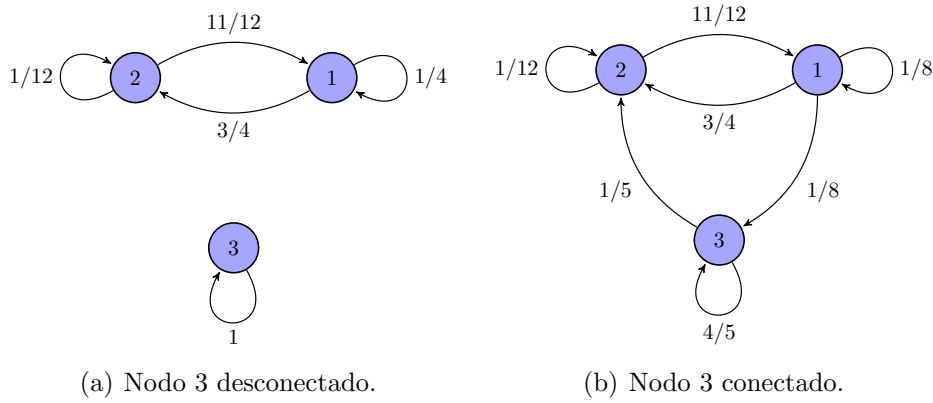


Fig. 4.6: Sistema de tres nodos.

Si partimos del punto $\pi^{(0)} = (0.7 \ 0.1 \ 0.2)^T$, la evolución se muestra en la Tabla 4.3, donde vemos que la solución converge al estado límite $(0.44 \ 0.36 \ 0.2)^T$. En la Figura 4.7(a) se indican los estados $\pi^{(0)}, \dots, \pi^{(4)}$, se puede observar que todos mantienen la tercer coordenada constante. El segmento azul representa a todos los estados de equilibrio y el punto rojo marca el estado límite.

El estado inicial $\pi^{(0)}$ se escribe como combinación de los autovectores de la forma

$$(0.7 \ 0.1 \ 0.2)^T = 0.8 (0.55 \ 0.45 \ 0)^T + 0.2 (0 \ 0 \ 1)^T + 0.26 (1 \ -1 \ 0)^T,$$

lo que nos permite calcular su evolución

$$\pi^{(t)} = 0.8 (0.55 \ 0.45 \ 0)^T + 0.2 (0 \ 0 \ 1)^T + 0.26 (-2/3)^t (1 \ -1 \ 0)^T,$$

de donde concluimos $\lim_{t \rightarrow \infty} \pi^{(t)} = 0.8 (0.55 \ 0.45 \ 0)^T + 0.2 (0 \ 0 \ 1)^T = (0.44 \ 0.36 \ 0.2)^T$. Es claro que el estado límite dependerá de las condiciones iniciales, en general tendremos

$$\begin{aligned} \pi^{(0)} &= \alpha (0.55 \ 0.45 \ 0)^T + (1 - \alpha) (0 \ 0 \ 1)^T + \beta (1 \ -1 \ 0)^T, \\ \pi^{(t)} &= \alpha (0.55 \ 0.45 \ 0)^T + (1 - \alpha) (0 \ 0 \ 1)^T + \beta (-2/3)^t (1 \ -1 \ 0)^T, \end{aligned}$$

de donde se concluye $\lim_{t \rightarrow \infty} \pi^{(t)} = (0.55 \alpha \ 0.45 \alpha \ 1 - \alpha)^T$.

Modificamos el sistema anterior conectando el nodo 3 con los otros, como se muestra en la Figura 4.6(b). La matriz de transición está dada por

$$P = \begin{pmatrix} 1/8 & 11/12 & 0 \\ 3/4 & 1/12 & 1/5 \\ 1/8 & 0 & 4/5 \end{pmatrix}$$

cuyos autovalores son $\lambda = 1, 0.724, -0.716$, con autovectores asociados $\boldsymbol{\pi}^* = (0.388 \ 0.370 \ 0.242)^T$, $\boldsymbol{\varrho}^* = (0.302 \ -0.198 \ -0.500)^T$, $\boldsymbol{\eta}^* = (0.500 \ -0.459 \ -0.041)^T$. Escribiendo el estado inicial

$$\boldsymbol{\pi}^{(0)} = (0.7 \ 0.1 \ 0.2)^T = \boldsymbol{\pi}^* + 0.035 \boldsymbol{\varrho}^* + 0.603 \boldsymbol{\eta}^*,$$

vemos que la evolución está dada por

$$\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}^* + 0.035 \times 0.724^t \boldsymbol{\varrho}^* + 0.603 \times (-0.716)^t \boldsymbol{\eta}^*,$$

por lo tanto $\lim_{t \rightarrow \infty} \boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}^*$, que en este caso es el único estado de equilibrio.

t	$\boldsymbol{\pi}^{(t)}$	t	$\boldsymbol{\pi}^{(t)}$
0	$(0.700 \ 0.100 \ 0.200)^T$	0	$(0.700 \ 0.100 \ 0.200)^T$
1	$(0.267 \ 0.533 \ 0.200)^T$	1	$(0.179 \ 0.573 \ 0.247)^T$
2	$(0.556 \ 0.244 \ 0.200)^T$	2	$(0.548 \ 0.232 \ 0.220)^T$
3	$(0.363 \ 0.437 \ 0.200)^T$	3	$(0.281 \ 0.474 \ 0.245)^T$
4	$(0.491 \ 0.309 \ 0.200)^T$	4	$(0.470 \ 0.299 \ 0.231)^T$
\vdots	\vdots	\vdots	\vdots
50	$(0.440 \ 0.360 \ 0.200)^T$	50	$(0.388 \ 0.370 \ 0.242)^T$

Tabla 4.3: .

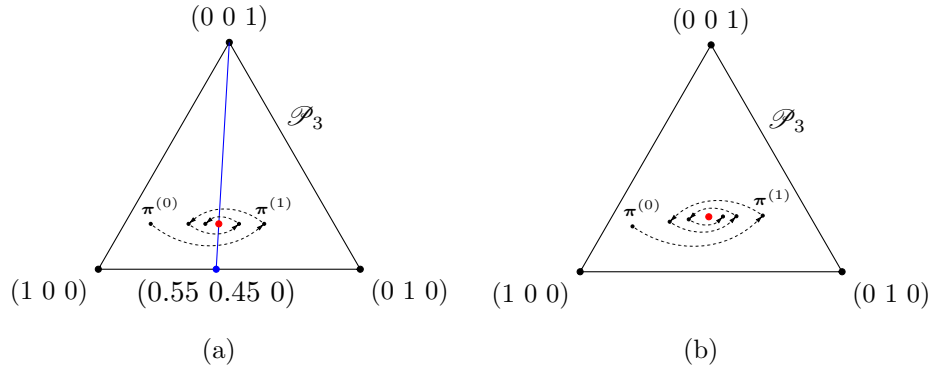


Fig. 4.7

4.4.5. Cadenas de Markov regulares. Una matriz P es positiva si $p_{ij} > 0$ para $1 \leq i, j \leq d$. Una matriz de Markov P es regular si para alguna potencia $t_0 \in \mathbb{N}$ tal que P^{t_0} es positiva. Es posible mostrar que en este caso P^t es positiva para $t \geq t_0$ (ver Ejercicio 4.4). Vamos a estudiar los autovalores de una matriz regular.

Proposición 4.5. Si P es una matriz de Markov positiva, $\lambda = 1$ es autovalor simple y para todo autovalor de P $\lambda \neq 1$, se verifica $|\lambda| < 1$.

Demostración. Consideramos la matriz P^T , ya vimos que $P^T \cdot \mathbf{1} = \mathbf{1}$, donde $\mathbf{1} = (1 \ 1 \dots 1)^T$. Si \mathbf{x} verifica $P^T \mathbf{x} = \mathbf{x}$ y definimos $x_j = \max\{x_1, \dots, x_d\}$, entonces

$$p_{1j}x_1 + \dots + p_{dj}x_d = x_j = p_{1j}x_j + \dots + p_{dj}x_j,$$

por lo tanto $0 = p_{1j}(x_j - x_1) + \dots + p_{dj}(x_j - x_d)$. Como $p_{ij} > 0$ y $x_j - x_i \geq 0$ para $i = 1, \dots, d$, entonces $x_j - x_i = 0$. Por lo tanto $\mathbf{x} = x_j \mathbf{1}$. \square

Corolario 4.2. Si P es una matriz de Markov regular y $\lambda \neq 1$ es un autovalor de P , entonces $|\lambda| < 1$.

Demostración. Si P^{t_0} es positiva, usando que λ^{t_0} es autovalor de P^{t_0} y la proposición anterior, tenemos $|\lambda^{t_0}| < 1$, lo que implica $|\lambda| < 1$. \square

4.4.6. Cadenas de Markov absorbentes. El nodo j es absorbente si $p_{ij} = 0$ para $i \neq j$, por lo tanto $p_{jj} = 1$. En general, un conjunto de nodos $J \subset \{1, \dots, d\}$ es una subcadena absorbente si $p_{ij} = 0$ para $i \notin J$ y $j \in J$. Esto quiere decir que un individuo que entra en algunos de los nodos $j \in J$, no sale de ese subconjunto.

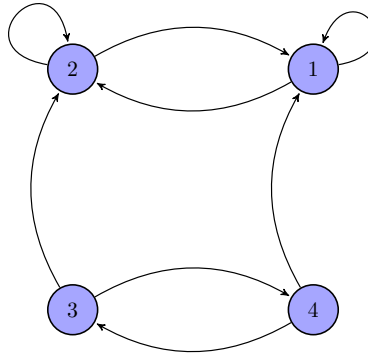


Fig. 4.8: Subcadena absorbente.

En la Figura 4.8 vemos que $J = \{1, 2\}$ es una subcadena absorbente.

4.4.7. Evolución probabilística*. Hasta ahora analizamos el comportamiento del sistema asumiendo que evoluciona estrictamente por lo que indica la matriz de transición. Pero en general, el comportamiento es aleatorio y p_{ij} establece la probabilidad que un individuo pase en un período del nodo j al nodo i . Esto nos dice que no podemos esperar que el sistema converja a un estado constante, cierto *ruido* aparecerá como consecuencia de las fluctuaciones aleatorias alrededor del equilibrio. Para ilustrar este comportamiento consideramos el siguiente ejemplo

Ejemplo 4.6. Consideramos el sistema de tres nodos de la Figura 4.9.

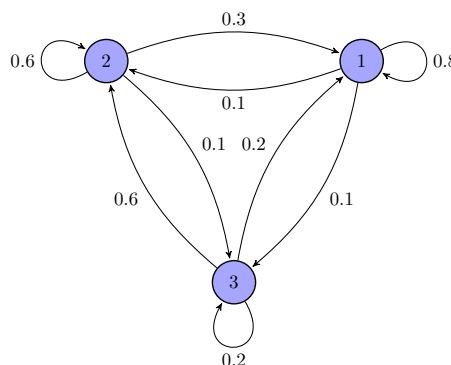


Fig. 4.9

Siendo la matriz de transición P positiva, existe un único estado de equilibrio $\pi^* = (0.578 \ 0.311 \ 0.111)^T$ que es el estado límite de toda punto inicial. Consideramos inicialmente $n = 1000$ individuos distribuidos entre los nodos 1, 2, 3 de la forma $n_1 = 125$, $n_2 = 625$ y $n_3 = 250$ respectivamente. Los individuos se trasladan alguno de los nodos en forma aleatoria, siguiendo las probabilidades

dadas por la matriz P , obteniendo una nueva distribución de la población. Se repite el proceso un número de veces (100). Los resultados se grafican en la Figura 4.10(a), donde podemos observar que rápidamente se acerca al estado límite, pero se mantiene fluctuante alrededor del valor de equilibrio. Podemos ver que después del tiempo transitorio, el valor medio de la proporción n_j/n es π_j^* . De la simulación obtenemos que el desvío estándar es $\sigma_1 \cong 0.016$, $\sigma_2 \cong 0.015$ y $\sigma_3 \cong 0.01$. Repetimos el experimento para $n = 4000$, los resultados se grafican en la Figura 4.10(b). Claramente vemos un nivel menor de ruido, lo que se confirma al calcular el desvío estándar: $\sigma_1 \cong 0.008$, $\sigma_2 \cong 0.007$ y $\sigma_3 \cong 0.005$. Vemos que al multiplicar por 4 la población la intensidad del ruido disminuye a la mitad, lo que es compatible con la regla de la \sqrt{n} (ver [22]).

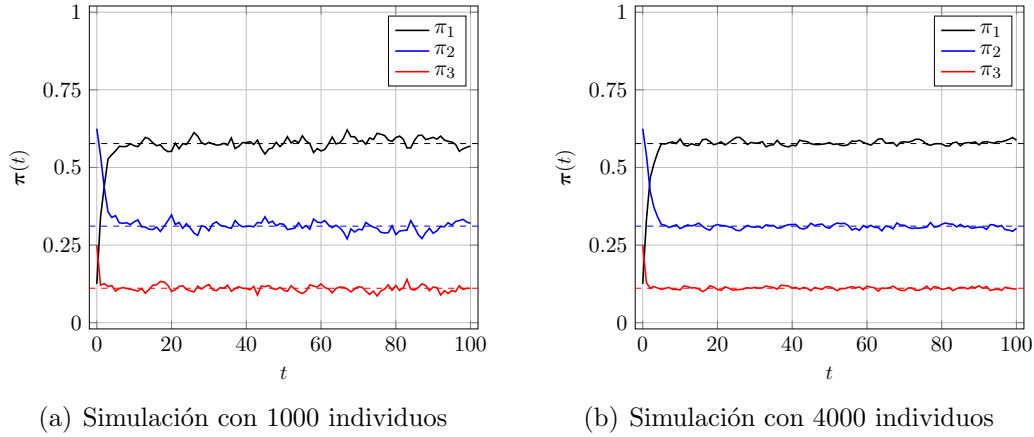


Fig. 4.10: Simulación del sistema.

4.5. Aplicaciones.

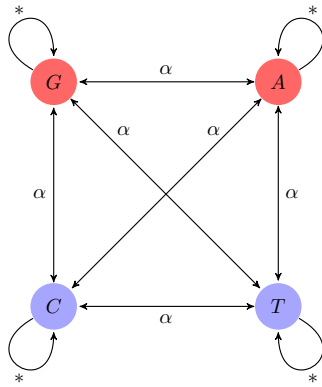
4.5.1. Evolución de poblaciones.

4.5.2. Genética. Se han propuesto modelos de Markov para el estudio de la evolución del ADN. Estos modelos de sustitución describen las mutaciones, donde un nucleótido reemplaza a otro con el paso de las generaciones. Se utilizan frecuentemente en análisis de filogenia molecular y difieren en los parámetros utilizados para describir las tasas de mutación. La aplicación principal es estimar la distancia evolutiva entre secuencias a partir de las diferencias observadas entre estas.

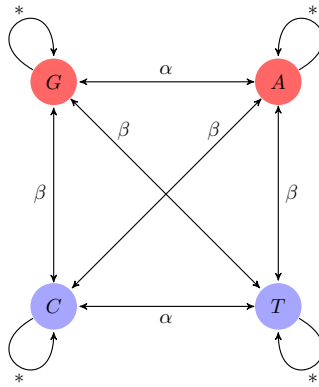
Estos modelos son descripciones fenomenológicas de la evolución del ADN como una cadena de cuatro estados discretos. Estos modelos de Markov no representan explícitamente el mecanismo de mutación ni la acción de la selección natural, sólo describen las tasas relativas de los diferentes cambios. Por ejemplo, sesgos de mutación y selección purificadora favorecen los canse conservativos y probablemente ambos son responsables de la tasa relativamente elevada de transiciones en comparación con las transversiones en las secuencias que evolucionan. Sin embargo, el modelo de Kimura (K80) que se describe más abajo sencillamente intenta capturar el efecto de las dos fuerzas en un parámetro que refleja la tasa relativa de transiciones y transversiones.

Los análisis evolutivos de secuencias se llevan a cabo en una amplia variedad de escalas temporales. Por ello es conveniente expresar estos modelos en términos de tasas instantáneas de cambios entre diferentes estados (las matrices Q de abajo). Si damos un estado inicial (ancestral) en la primera posición, la matriz del modelo Q y la longitud de la rama expresando el número esperado de cambios que se han dado desde el ancestro, entonces podemos derivar la probabilidad de la secuencia descendente teniendo cada uno de los cuatro estadios. Los detalles matemáticos de esta transformación desde tasa-matriz a matriz de probabilidades en

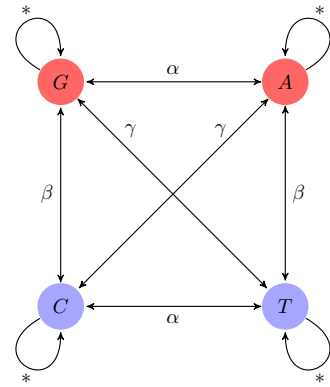
describen en la sección de modelos matemáticos de sustitución de la página de modelos de sustitución. Expresando en términos de tasas instantáneas de cambio podemos evitar estimar grandes números de parámetros para cada rama en un árbol filogenético (o cada comparación si el análisis incluye muchas comparaciones de secuencias por parejas).



(a) Modelo de Jukes-Cantor.



(b) Modelo de Kimura con dos parámetros.



(c) Modelo de Kimura con tres parámetros.

Fig. 4.11: Modelos de evolución de mutaciones.

4.5.3. Modelos epidemiológicos.

4.6. Ejercicios.

Ejercicio 4.1. Determinar cuáles de las siguientes matrices son de Markov.

$$\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1/2 & 2/3 \\ 1/2 & 1/3 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 1/3 \\ 1/2 & 1/2 & 1/2 \\ -1/2 & 1/2 & 1/6 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 & 0.3 \\ 1 & 0 & 0.1\hat{6} \\ 0 & 1 & 0.5 \end{pmatrix}.$$

Ejercicio 4.2. Se tiene un proceso de Markov cuya matriz de transición es

$$P = \begin{pmatrix} 3/5 & 3/10 \\ 2/5 & 7/10 \end{pmatrix},$$

verificar que el vector $\pi = (3/7 \ 4/7)^T$ es un estado de equilibrio del proceso.


Ejercicio 4.3. Sea $P = \begin{pmatrix} 2/3 & 1/4 \\ 1/3 & 3/4 \end{pmatrix}$ la matriz de transición de un proceso de Markov y sea $\pi^{(2)}$ el segundo estado, verificar que P es inversible y calcular $\pi^{(1)}$ y $\pi^{(0)}$.

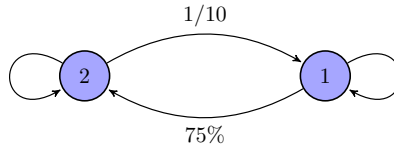
Ejercicio 4.4. Probar que si P y Q son matrices de Markov, entonces

(a) $(1 - \lambda)P + \lambda Q$ es una matriz de Markov para $0 \leq \lambda \leq 1$.


(b) $P \cdot Q$ es una matriz de Markov.

(c) Si P es positiva, $P \cdot Q$ es positiva.

 **Ejercicio 4.5.** La población en estudio está distribuida en un territorio dividido en dos sectores. Esta población es constante y se desplaza. En el momento inicial, exactamente la mitad de la población está en cada sector. Al día siguiente se observa que el 75 % de la población del Sector 1 se ha desplazado al Sector 2, mientras que 1 de cada 10 individuos que estaban en el Sector 2 pasó al Sector 1. Esta pauta de desplazamiento se mantiene.



- (a) Determinar la matriz de transición y el estado inicial.
- (b) Calcular los 5 primeros estados del proceso de Markov.
- (c) Verificar que el vector $\pi = (2/17 \ 15/17)^T$ es estado de equilibrio.
- (d) Simular el comportamiento del sistema con una población total de 100 individuos.

 **Ejercicio 4.6.** En el instante inicial 20 ratones se encuentran en el compartimiento I (ver Figura 4.12). Las puertas que separan los compartimientos permanecen cerradas salvo durante un breve lapso, donde los ratones pueden pasar a un compartimiento adyacente o permanecer en el mismo. Se supone que nada distingue un compartimiento de otro, es decir que es igualmente probable que un ratón pase a cualquiera de los adyacentes o se quede en el compartimiento en el que está. Se realizan observaciones cada hora y se registra el número de ratones en cada compartimiento.

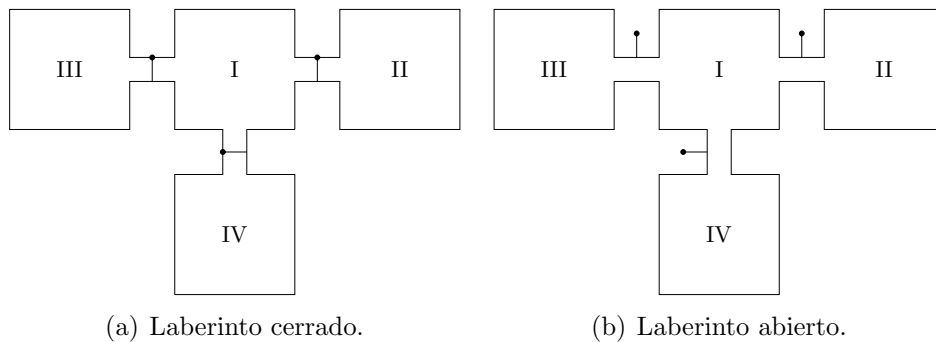



Fig. 4.12: El laberinto se abre unos pocos segundos cada hora.

- (a) Determinar la matriz de transición del proceso.
- (b) Determinar el vector de estado después de 4 horas.
- (c) Decidir si existe o no un estado de equilibrio.

 **Ejercicio 4.7.** Describir las matrices de transición y los estados de equilibrio correspondientes a los sistemas mostrados en la Figura 4.13.

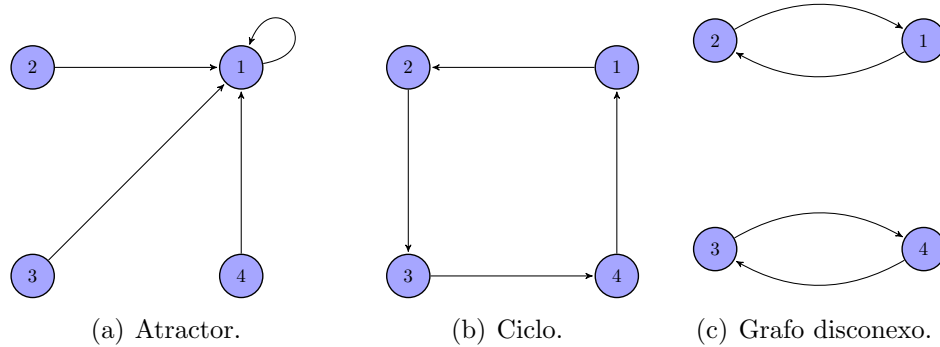


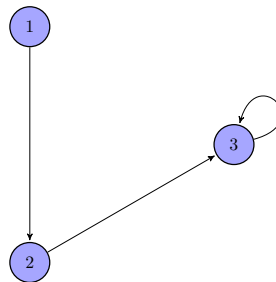
Fig. 4.13: Evolución de los sistemas.

Ejercicio 4.8. Un país, cuya población es constante está dividido en dos regiones. Cada año 1 de cada 10 residentes de la región A se traslada a la región B mientras que 1 de cada 5 habitantes de la zona B se muda a la región A. En el instante inicial (ahora) viven 6 millones en la región A y 30 millones en la B.

- Escribir la matriz de transición para este proceso.
- Determinar si existe un estado de equilibrio.
- Calcular el estado de la población dentro de 10 años y dentro de 30 años.

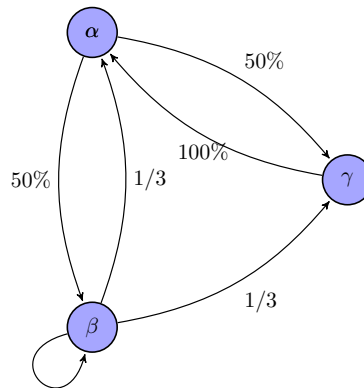
Ejercicio 4.9. La población en estudio es constante y está distribuida en un territorio dividido en tres sectores. Día a día se observan los desplazamientos: el 100 % de la población del Sector 1 se desplaza al Sector 2, el 100 % de la población del Sector 2 se ha desplaza al Sector 1, mientras que la población del Sector 3 permanece (sin desplazarse) en su sector. Esta pauta de desplazamiento se mantiene en el tiempo.

- Determinar la matriz de transición P que describe el proceso.
- Decidir si hay dos estados de equilibrio diferentes.
- Determinar si el proceso tiene un estado límite, y en caso afirmativo hallarlo, con una población inicial de:
 - 200 habitantes en el Sector 1, 200 en el Sector 2 y 300 en el Sector 3.
 - 100 habitantes en el Sector 1, 200 en el Sector 2 y 300 en el Sector 3.
- Determinar si existe P_∞ . Justificar.

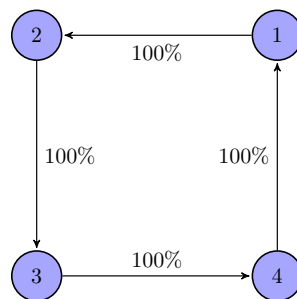


Ejercicio 4.10. Un proceso de Markov admite 3 sectores: α, β, γ . Al cabo de un período el 50 % de los individuos del sector α pasan al β y el otro 50 % al γ . Además, un tercio de los individuos que están en el estado β pasa al α y otro tercio pasa al γ . Finalmente, todos los individuos en el estado γ pasan al α .

- (a) Construir la matriz de transición P que describe el proceso.
- (b) Si el estado actual está dado por $\pi^0 = (0.5 \ 0.25 \ 0.25)^T$, determinar el estado siguiente.
- (c) Analizar el comportamiento de P^n para valores de n grandes.
- (d) Decidir si existe un estado límite.



Ejercicio 4.11. La población en estudio es constante y está distribuida en un territorio dividido en cuatro sectores. Día a día se observan los desplazamientos: el 100 % de la población del Sector 1 se desplaza al Sector 2, el 100 % de la población del Sector 2 se ha desplaza al Sector 3, el 100 % de la población del Sector 3 se ha desplaza al Sector 4 y el 100 % de la población del Sector 4 se ha desplaza al Sector 1. Esta pauta de desplazamiento se mantiene en el tiempo.




- (a) Determinar la matriz de transición P que describe el proceso.
- (b) Decidir si hay dos estados de equilibrio diferentes.
- (c) Determinar si el proceso tiene un estado límite, y en caso afirmativo hallarlo, con una población inicial de:
 - (I) 100 habitantes en cada Sector.
 - (II) 100 habitantes en el Sector 1, 300 en el Sector 2, 100 en el Sector 3 y 300 en el Sector 4.
- (d) Decidir si existe P^∞ (Sugerencia: calcular P^4).

Ejercicio 4.12. Para el proceso mostrado en la Figura 4.11(a).

- (a) Determinar la matriz de transición P que lo describe.

- (b) Establecer el intervalo del parámetro α .
- (c) Determinar, si existe, P^∞ para cada valor de α .

 **Ejercicio 4.13.** Calcular $P^* = \lim_{t \rightarrow \infty} P^t$ para la matriz de transición asociada la cadena de Markov de la Figura 4.4.

CAPÍTULO 5

Análisis de Datos

“Esas ambigüedades, redundancias y deficiencias recuerdan las que el doctor Franz Kuhn atribuye a cierta enciclopedia china que se titula Emporio celestial de conocimientos benévolos. En sus remotas páginas está escrito que los animales se dividen en (a) pertenecientes al Emperador, (b) embalsamados, (c) amaestrados, (d) lechones, (e) sirenas, (f) fabulosos, (g) perros sueltos, (h) incluidos en esta clasificación, (i) que se agitan como locos, (j) innumerables, (k) dibujados con un pincel finísimo de pelo de camello, (l) etcétera, (m) que acaban de romper el jarrón, (n) que de lejos parecen moscas.”

Jorge Luis Borges [4]

5.1. Ajuste por cuadrados mínimos. Es un procedimiento para ajustar a un conjunto de datos (pares ordenados y familia de funciones), la función que mejor se aproxime a los datos (línea de regresión o la línea de mejor ajuste), dentro de una familia de funciones dependiente de uno o varios parámetros. Concretamente, dada una tabla

x	x_1	x_2	\dots	x_N
y	y_1	y_2	\dots	y_N

buscamos una función $f(x; \alpha_1, \dots, \alpha_m)$ que realice el menor error de los residuos r_j definidos por $r_j = y_j - f(x_j; \alpha_1, \dots, \alpha_m)$ en el sentido de mínimos cuadrados. La idea es minimizar la suma S de cuadrados de los residuos:

$$S(\alpha_1, \dots, \alpha_m) = \frac{1}{2}(r_1^2 + \dots + r_N^2) = \frac{1}{2} \sum_{j=1}^N (y_j - f(x_j; \alpha_1, \dots, \alpha_m))^2$$

Este método se utiliza comúnmente para analizar una serie de datos que se obtengan de algún estudio, con el fin de expresar su comportamiento de manera lineal y así minimizar los errores de los datos obtenidos.

5.1.1. Modelo lineal. Vamos a estudiar el siguiente problema, supongamos que dos variables se relacionan de la forma $y = \alpha + \beta x$, pero los coeficientes α y β son desconocidos. Si mediante observaciones tenemos para distintos valores de x , los valores de la variable y correspondiente, podemos hallar los coeficientes desconocidos resolviendo el sistema

$$(5.1) \quad \begin{cases} y_1 = \alpha + \beta x_1, \\ y_2 = \alpha + \beta x_2, \\ \vdots \\ y_N = \alpha + \beta x_N. \end{cases}$$

Si $N = 1$ el sistema admite infinitas soluciones, para $N = 2$ hay una única solución supuesto que $x_1 \neq x_2$. Para $N > 2$, el sistema está sobredeterminado, por lo tanto tendremos solución sólo en el caso de tener los valores exactos de x_j e y_j (y asumiendo que la relación entre las variables es estrictamente lineal). Si dichos valores provienen de observaciones o de cálculos numéricos, esto no sucede y el sistema (5.1) no tiene solución. Vamos a replantear el problema de la siguiente forma: queremos hallar α, β que minimicen la distancia entre las observaciones y la predicción realizada con el modelo. Más concretamente, si definimos $r_j = y_j - \hat{y}_j$, donde $\hat{y}_j = \alpha + \beta x_j$, busquemos que $S = \frac{1}{2}(r_1^2 + \cdots + r_N^2)$ sea mínimo. Podemos escribir S en función de α y β :

$$S(\alpha, \beta) = \frac{1}{2} \sum_{j=1}^N (y_j - \alpha - \beta x_j)^2,$$

sus derivadas valen

$$\begin{aligned} \frac{\partial S}{\partial \alpha}(\alpha, \beta) &= - \sum_{j=1}^N (y_j - \alpha - \beta x_j) = - \sum_{j=1}^N y_j + \alpha N + \beta \sum_{j=1}^N x_j, \\ \frac{\partial S}{\partial \beta}(\alpha, \beta) &= - \sum_{j=1}^N (y_j - \alpha - \beta x_j) x_j = - \sum_{j=1}^N y_j x_j + \alpha \sum_{j=1}^N x_j + \beta \sum_{j=1}^N x_j^2. \end{aligned}$$

Para minimizar buscamos un punto (α, β) crítico, es decir donde ambas derivadas se anulen, esto nos lleva a las ecuaciones normales

$$\begin{cases} \sum_{j=1}^N y_j = \alpha N + \beta \sum_{j=1}^N x_j, \\ \sum_{j=1}^N y_j x_j = \alpha \sum_{j=1}^N x_j + \beta \sum_{j=1}^N x_j^2, \end{cases}$$

por ser la función S convexa, un punto crítico es necesariamente un mínimo (global).

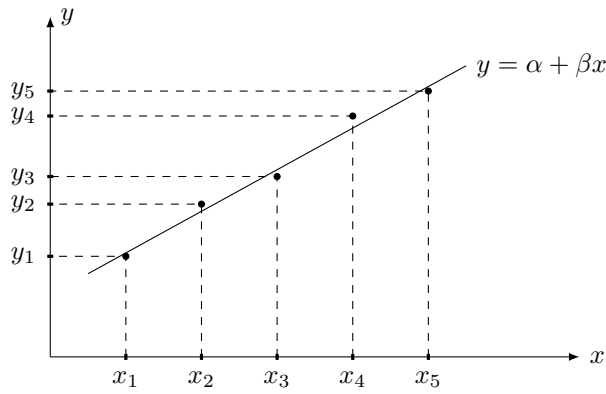
Ejemplo 5.1. Consideremos la Tabla 5.1, las ecuaciones normales se escriben como

x	y	\hat{y}	$y - \hat{y}$
1	2.65707	2.75794	-0.100 87
2	4.04218	3.86579	0.176 39
3	4.76812	4.97365	-0.205 53
4	6.36685	6.08150	0.285 35
5	7.03401	7.18936	-0.155 35

Tabla 5.1

$$\begin{cases} 24.8682 = 5\alpha + 15\beta \\ 85.6833 = 15\alpha + 55\beta \end{cases}$$

cuya solución es $\alpha = 1.65008$, $\beta = 1.10786$, $S \cong 0.095$. En el gráfico 5.1 mostramos los puntos de la tabla y la recta obtenida.

Fig. 5.1: Recta de ajuste $y = \alpha + \beta x$.

5.1.2. Interpretación geométrica. Vamos a darle un sentido geométrico al problema planteado, empezaremos por recordar la noción de producto interno. Dados dos vectores de $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$ se define el producto interno como $\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + \cdots + u_N v_N \in \mathbb{R}$. De la definición se ve claramente que

- $\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$,
- $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$,
- $\mathbf{u} \cdot \mathbf{u} = u_1^2 + \cdots + u_N^2 \geq 0$ y $\mathbf{u} \cdot \mathbf{u} = 0$ si y sólo si $\mathbf{u} = \mathbf{0}$.

Vemos que la norma de un vector se relaciona con el producto interno de la forma: $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$. Se puede probar la desigualdad de Cauchy:

$$-\|\mathbf{u}\| \|\mathbf{v}\| \leq \mathbf{u} \cdot \mathbf{v} \leq \|\mathbf{u}\| \|\mathbf{v}\|,$$

lo que nos permite definir un ángulo $\theta \in [0, \pi]$ como el que verifica $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \cos(\theta)$.

Dada una recta en el plano generado por el vector \mathbf{u} y un punto \mathbf{v} fuera de la misma, sabemos que la mínima distancia entre la recta y el punto, se alcanza en la intersección $\hat{\mathbf{v}}$ con la perpendicular que pasa por el punto exterior (ver Figura 5.2(a)). De forma similar, si consideramos el plano generado por los vectores $\mathbf{u}_1, \mathbf{u}_2$ del espacio tridimensional y un punto \mathbf{v} exterior al mismo, la distancia mínima del punto al plano se obtiene en el punto $\hat{\mathbf{v}}$ proyección ortogonal de \mathbf{v} sobre el plano, como se muestra en la Figura 5.2(b). Si \mathbf{w} es otro punto del plano, por el Teorema de Pitágoras tenemos

$$\|\mathbf{v} - \mathbf{w}\|^2 = \|\mathbf{v} - \hat{\mathbf{v}}\|^2 + \|\hat{\mathbf{v}} - \mathbf{w}\|^2 \geq \|\hat{\mathbf{v}} - \mathbf{w}\|^2$$

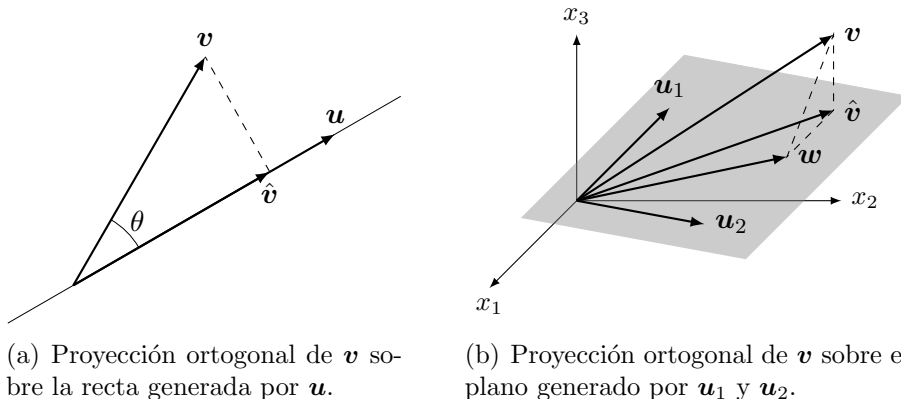


Fig. 5.2: Proyección ortogonal sobre subespacios de dimensión 1 y 2.

Podemos usar estas ideas para interpretar las ecuaciones normales, si $\mathbf{1}, \mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ son los vectores definidos por $\mathbf{1} = (1 \dots 1)^T$, $\mathbf{x} = (x_1 \dots x_N)^T$ e $\mathbf{y} = (y_1 \dots y_N)^T$, buscamos el vector $\hat{\mathbf{y}}$ en el plano generado por los vectores $\mathbf{1}, \mathbf{x}$ que realice la mínima distancia al vector \mathbf{y} . Tenemos que $\hat{\mathbf{y}} = \alpha \mathbf{1} + \beta \mathbf{x}$ debe verificar $\mathbf{y} - \hat{\mathbf{y}} \perp \mathbf{1}$, $\mathbf{y} - \hat{\mathbf{y}} \perp \mathbf{x}$. Se deduce que α, β deben satisfacer el sistema (2×2) :

$$\begin{aligned}\mathbf{y} \cdot \mathbf{1} &= \alpha \mathbf{1} \cdot \mathbf{1} + \beta \mathbf{1} \cdot \mathbf{x}, \\ \mathbf{y} \cdot \mathbf{x} &= \alpha \mathbf{1} \cdot \mathbf{x} + \beta \mathbf{x} \cdot \mathbf{x},\end{aligned}$$

que tiene única solución si y sólo si los vectores $\mathbf{1}, \mathbf{x}$ no son paralelos, es decir no son todos iguales los valores x_j . Se suele denominar residuo a $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$, el valor de S se puede calcular como $\frac{1}{2} \|\mathbf{r}\|^2$. Si \mathbf{z} es otro vector del plano generado por $\mathbf{1}$ y \mathbf{x} , $\mathbf{z} = \gamma \mathbf{1} + \delta \mathbf{x}$, tenemos

$$\|\mathbf{y} - \mathbf{z}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}} + \hat{\mathbf{y}} - \mathbf{z}\|^2 = \|\mathbf{r}\|^2 + \|\hat{\mathbf{y}} - \mathbf{z}\|^2 + 2\mathbf{r} \cdot (\hat{\mathbf{y}} - \mathbf{z}).$$

Como \mathbf{r} es perpendicular a todo elemento generado por $\mathbf{1}$ y \mathbf{x} , el doble producto se anula por lo tanto $\|\mathbf{y} - \mathbf{z}\|^2 = \|\mathbf{r}\|^2 + \|\hat{\mathbf{y}} - \mathbf{z}\|^2$.

Si bien la función S no permite hallar los valores de α y β que hacen óptimo el ajuste del modelo en el sentido de mínimos cuadrados, su valor numérico no tiene significado. Observemos por ejemplo que S se modifica con cambios de escala (de unidades). Queremos un indicador de la bondad del ajuste invariante por transformaciones lineales de la variable y . Si consideramos $\bar{\mathbf{y}} = (\mathbf{y} \cdot \mathbf{1} / \mathbf{1} \cdot \mathbf{1}) \mathbf{1} = (\bar{y} \dots \bar{y})^T$, donde $\bar{y} = (y_1 + \dots + y_N) / N$, podemos escribir

$$\|\mathbf{r}\|^2 = \|\mathbf{r} - \bar{\mathbf{y}}\|^2 - \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 = \|\mathbf{y} - \bar{\mathbf{y}}\|^2 \left(1 - \frac{\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2} \right).$$

Como $\|\mathbf{y} - \bar{\mathbf{y}}\|^2$ es independiente de la elección los parámetros α y β , vemos que el ajuste depende del cociente $R^2 = \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 / \|\mathbf{y} - \bar{\mathbf{y}}\|^2 \in [0, 1]$, mayores valores de R^2 reflejan un mejor ajuste. Al contrario de lo que sucede con S , el coeficiente R^2 no cambia por modificaciones lineales de la variable y , es decir que si consideramos la variable $Y = my + b$, tenemos $\alpha_Y = b + m\alpha_y$, $\beta_Y = m\beta_y$, $\chi_Y^2 = m^2\chi_y^2$ pero $R_Y^2 = R_y^2$.

Ejemplo 5.2. En la Tabla 5.2 (ver <https://www.engineeringtoolbox.com/>) se muestran la temperatura (°C) de fusión del hielo para distinta presiones (MPa). Queremos establecer una relación lineal $T = \alpha + \beta P$.

Presión	Temp	Presión	Temp	Presión	Temp
6.1×10^{-4}	0.01	2.0×10^1	-1.54	9.0×10^1	-7.91
1.0×10^{-1}	0.003	3.0×10^1	-2.36	1.0×10^2	-8.94
1.0	-0.064	4.0×10^1	-3.21	1.2×10^2	-11.09
2.0	-0.14	5.0×10^1	-4.09	1.4×10^2	-13.35
5.0	-0.37	6.0×10^1	-5.00	1.6×10^2	-15.73
1.0×10^1	-0.75	7.0×10^1	-5.94	1.8×10^2	-18.22
1.5×10^1	-1.14	8.0×10^1	-6.91	2.0×10^2	-20.83

Tabla 5.2

Las ecuaciones normales son

$$\begin{cases} -1.2757 \times 10^2 = 21\alpha + 1.3731 \times 10^2 \beta, \\ -1.6498 \times 10^3 = 1.3731 \times 10^2 \alpha + 1.7075 \times 10^3 \beta, \end{cases}$$

de donde obtenemos los parámetros $\alpha = 0.5116$, $\beta = -1.007$ y los indicadores de bondad del ajuste $S = 2.71$, $R^2 = 0.993$. Si expresamos la temperatura en $^{\circ}\text{F}$ en lugar de hacerlo en $^{\circ}\text{C}$, $T(^{\circ}\text{F}) = 32 + 1.8 T(^{\circ}\text{C})$, obtenemos $\alpha = 32.920$, $\beta = -1.813$ y $S = 8.77$, $R^2 = 0.993$. En la Figura 5.3, los puntos presión-temperatura y la recta de ajuste.

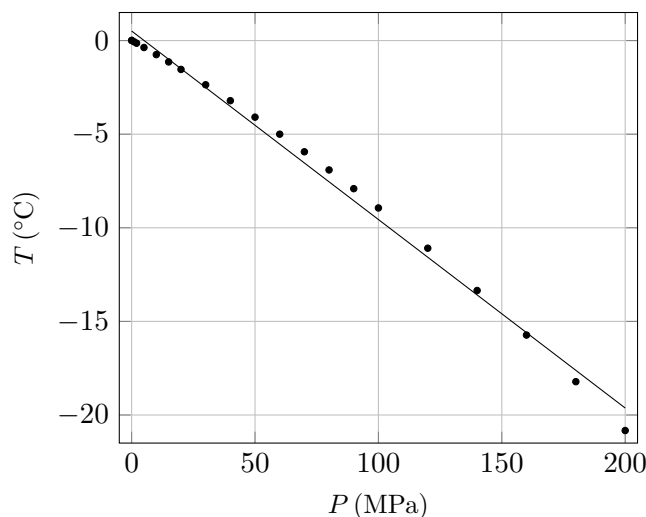


Fig. 5.3

5.1.3. Linealización. En general la ley que relaciona las variables de interés x e y , $y = f(x)$ no es lineal, pero en muchos casos existen transformaciones simples de cada una de ellas que convierten al problema en lineal. Es decir, si consideramos nuevas variables $X = g(x)$, $Y = h(y)$, la relación anterior se escribe como $Y = \alpha + \beta X$. En la Tabla 5.3 mostramos distintos modelos no lineales que aparecen con frecuencia en las aplicaciones y las transformaciones para linealizarlos.

Modelo	Transformación		
	$Y = h(y)$	$X = g(x)$	$Y = \alpha + \beta X$
Exponencial: $y = k e^{bx}$	$Y = \ln(y)$	$X = x$	$Y = \ln(k) + b X$
Potencia: $y = k x^b$	$Y = \ln(y)$	$X = \ln(x)$	$Y = \ln(k) + b X$
Logarítmica: $y = k + b \ln(x)$	$Y = y$	$X = \ln(x)$	$Y = k + b X$
Hiperbólico: $y = k x/(b + x)$	$Y = 1/y$	$X = 1/x$	$Y = 1/k + b/k X$

Tabla 5.3: Modelos no lineales y las linealizaciones.

Ejemplo 5.3 (Modelo potencial). La tercera ley de Kepler establece una relación potencial entre el período orbital (año) de cada planeta y el semieje mayor de su órbita. En la Figura 5.4 mostramos la órbita elíptica del planeta alrededor del sol, el cual ocupa un foco de la elipse (primera ley de Kepler). En la Tabla 5.4 mostramos los valores de las semiejes a medida en metros y el período orbital medido en días.

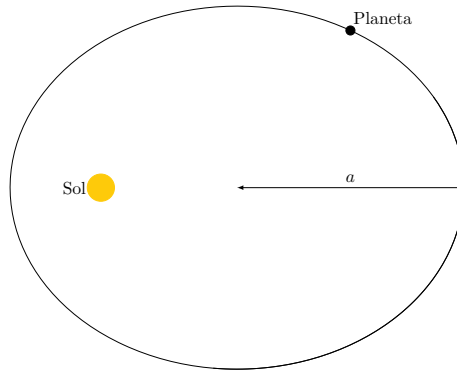


Fig. 5.4: Órbita elíptica y semieje mayor.

Planeta	a (m)	τ (d)
Mercurio (☿)	$5.790\,922\,70 \times 10^{10}$	$8.796\,934 \times 10^1$
Venus (♀)	$1.082\,094\,75 \times 10^{11}$	$2.247\,010 \times 10^2$
Tierra (♁)	$1.495\,982\,62 \times 10^{11}$	$3.652\,570 \times 10^2$
Marte (♂)	$2.279\,388\,24 \times 10^{11}$	$6.869\,601 \times 10^2$
Júpiter (♃)	$7.783\,408\,21 \times 10^{11}$	$4.335\,355 \times 10^3$
Saturno (♄)	$1.426\,666\,42 \times 10^{12}$	$1.075\,774 \times 10^4$
Urano (♅)	$2.870\,658\,19 \times 10^{12}$	$3.079\,910 \times 10^4$
Neptuno (♆)	$4.498\,396\,44 \times 10^{12}$	$6.022\,490 \times 10^4$

Tabla 5.4: Valores del semieje mayor a y el período orbital τ .

La tercera ley de Kepler establece la relación $\tau = k a^b$, que se transforma en $T = \alpha + \beta A$, donde $T = \ln(\tau)$, $\alpha = \ln(k)$ y $b = \beta$. Las ecuaciones normales son

$$\begin{cases} 61.3237 = 8.0 \alpha + 215.260 b, \\ 1677.42 = 215.260 \alpha + 5810.35 b, \end{cases}$$

que tienen como solución $\alpha = -32.7041$ y $b = 1.50031$, es decir

$$\tau = 6.263\,17 \times 10^{-15} a^{1.50031}.$$

Obsevemos que Johannes Kepler no disponía de datos sobre Urano y Neptuno dado que no se conocían estos planetas al momento de enunciar esta ley (1618).¹ De la ley de gravitación de Newton (1685) se deduce que $\tau = 2\pi (GM)^{-1/2} a^{3/2}$, donde G es la constante de gravitación universal y M es la masa del Sol. Usando que $G = 6.674 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$ y $M = 1.989 \times 10^{30} \text{ kg}$, obtenemos el valor $k = 6.311\,83 \times 10^{-15} \text{ d/m}^{3/2}$ ($b = 1.5$).

Notemos que de las observaciones astronómicas, sólo podemos obtener el valor del producto GM , dado que en la fórmula solo interviene el producto de ambas cantidades. No fue hasta 1798 que Henry Cavendish obtuvo implícitamente el valor de G como resultado de los experimentos realizados con balanzas de torsión que permitieron determinar la masa de la Tierra y, como consecuencia, el valor de G .

¹ Los planetas Mercurio, Venus, Marte, Júpiter y Saturno, son los más cercanos a la Tierra y pueden ser vistos sin instrumentos ópticos. Urano fue descubierto en 1781 y Neptuno en 1846.

5.1.4. Mínimos cuadrados generalizados y no lineales*. Debemos observar que al linealizar el problema tomando $X = g(x)$ y $Y = h(y)$ y luego aplicar el método de regresión lineal, estamos minimizando $(Y_1 - \hat{Y}_1)^2 + \cdots + (Y_N - \hat{Y}_N)^2$ que no es equivalente a minimizar $(y_1 - \hat{y}_1)^2 + \cdots + (y_N - \hat{y}_N)^2$. Supongamos que \hat{Y} son los valores del modelo de mínimos cuadrados linealizado $Y = \alpha + \beta X$, si definimos $\tilde{y}_j = h^{-1}(\hat{Y}_j)$, suponiendo $\tilde{y}_j \cong y_j$ tenemos $Y_j - \hat{Y}_j = h(y_j) - h(\tilde{y}_j) \cong h'(y_j)(y_j - \tilde{y}_j)$, de donde se deduce

$$(Y_1 - \hat{Y}_1)^2 + \cdots + (Y_N - \hat{Y}_N)^2 \cong (h'(y_1))^2(y_1 - \tilde{y}_1)^2 + \cdots + (h'(y_N))^2(y_N - \tilde{y}_N)^2.$$

Podemos considerar una función S modificada

$$S_G = \frac{1}{2}(w_1(Y_1 - \hat{Y}_1)^2 + \cdots + w_N(Y_N - \hat{Y}_N)^2),$$

donde $w_j = (h'(y_j))^{-2}$. Al minimizar S_G obtenemos las nuevas ecuaciones normales

$$\begin{cases} \sum_{j=1}^N w_j Y_j = \alpha \sum_{j=1}^N w_j + \beta \sum_{j=1}^N w_j X_j, \\ \sum_{j=1}^N w_j Y_j X_j = \alpha \sum_{j=1}^N w_j X_j + \beta \sum_{j=1}^N w_j X_j^2, \end{cases}$$

asumiendo $(h'(y_j))^{-2}(Y_j - \hat{Y}_j)^2 \cong (y_j - h^{-1}(\hat{Y}_j))^2$.

Ejemplo 5.4. Queremos estudiar el modelo exponencial $y = k e^{bx}$ que mejor ajuste a los datos de la Tabla 5.5.

x	y	$\ln(y)$	x	y	$\ln(y)$
0.068 005	1.816 13	0.596 71	1.110 41	9.0719	2.205 18
0.251 188	1.850 42	0.615 41	1.296 46	12.6341	2.536 40
0.293 132	1.458 38	0.377 33	1.326 62	13.8117	2.625 52
0.559 454	3.399 98	1.223 77	1.524 57	20.1724	3.004 32
0.667 651	3.160 39	1.150 70	1.651 31	27.8109	3.325 43
0.710 296	5.287 71	1.665 39	1.801 58	37.2007	3.616 33
0.955 559	6.373 66	1.852 17	1.974 99	51.1806	3.935 36
0.962 632	6.811 48	1.918 61	1.979 86	52.8378	3.967 23

Tabla 5.5

Si consideramos el mejor ajuste de S , donde

$$S = \frac{1}{2} \sum_{j=1}^{16} (\ln(y_j) - \alpha - \beta x_j)^2$$

obtenemos las ecuaciones normales

$$\begin{cases} 34.6158 = 16.0 \alpha + 17.1337 \beta, \\ 47.9919 = 17.1337 \alpha + 24.0476 \beta, \end{cases}$$

que tienen como solución $\alpha = 0.111283$ y $\beta = 1.91642$, lo que corresponde a $y = 1.11771 e^{1.91642 x}$. Por otro lado, considerando que $(h'(y))^{-2} = y^2$, la función a minimizar es

$$S_G = \frac{1}{2} \sum_{j=1}^{16} y_j^2 (\ln(y_j) - \alpha_G - \beta_G x_j)^2$$

y las ecuaciones normales generalizadas

$$\begin{cases} 31\,512.0 = 8553.6\alpha_G + 15\,761.2\beta_G, \\ 59\,012.9 = 15\,761.2\alpha_G + 29\,515.6\beta_G, \end{cases}$$

que da como resultado $\alpha_G = -4.011\,17 \times 10^{-3}$ y $\beta_G = 2.00152$, es decir $y = 0.995997 e^{2.00152x}$. Para comparar ambos modelos consideramos la suma de cuadrados de la función no lineal:

$$S_N(\alpha, \beta) = \frac{1}{2} ((y_1 - e^{\alpha+\beta x_1})^2 + \cdots + (y_{16} - e^{\alpha+\beta x_{16}})^2),$$

para cada uno de los modelos anteriores obtenemos $S_N(\alpha, \beta) = 11.5404$ y $S_N(\alpha_G, \beta_G) = 2.73312$, donde vemos la mejora del segundo método.

Podemos plantear también directamente la minimización de la función $S_N(\alpha, \beta)$, para esto buscamos un punto crítico de S_N

$$\begin{aligned} 0 &= \frac{\partial S_N}{\partial \alpha}(\alpha, \beta) = \sum_{j=1}^{16} (y_j - e^{\alpha+\beta x_j}) e^{\alpha+\beta x_j}, \\ 0 &= \frac{\partial S_N}{\partial \beta}(\alpha, \beta) = \sum_{j=1}^{16} (y_j - e^{\alpha+\beta x_j}) e^{\alpha+\beta x_j} x_j, \end{aligned}$$

resolviendo estas ecuaciones no lineales obtenemos $\alpha_N = -2.899\,22 \times 10^{-2}$ y $\beta_N = 2.01455$, $S_N = 2.68944$, lo que nos permite concluir que el método de mínimos cuadrados generalizados nos da una buena aproximación.

Resolución de las ecuaciones no lineales Para hallar un punto crítico de S_N planteamos la iteración del método de Newton

$$\begin{pmatrix} \alpha^{(n)} \\ \beta^{(n)} \end{pmatrix} = \begin{pmatrix} \alpha^{(n-1)} \\ \beta^{(n-1)} \end{pmatrix} - \begin{pmatrix} \frac{\partial^2 S_N}{\partial \alpha^2} & \frac{\partial^2 S_N}{\partial \alpha \partial \beta} \\ \frac{\partial^2 S_N}{\partial \alpha \partial \beta} & \frac{\partial^2 S_N}{\partial \beta^2} \end{pmatrix}^{-1} \cdot \begin{pmatrix} \frac{\partial S_N}{\partial \alpha} \\ \frac{\partial S_N}{\partial \beta} \end{pmatrix} \bigg|_{(\alpha^{(n-1)}, \beta^{(n-1)})}$$

Partiendo del punto inicial $(\alpha, \beta) = (0.111283, 1.91642)$ obtenido mediante la linealización obtenemos las iteraciones que mostramos en siguiente tabla:

n	α	β	S_N
0	0.111 283 384	1.916 415 381	11.540 371 940
1	-0.033 322 616	2.018 800 281	2.746 016 039
2	-0.029 027 752	2.014 584 857	2.689 441 908
3	-0.028 992 228	2.014 554 909	2.689 440 046
4	-0.028 992 226	2.014 554 908	2.689 440 046
5	-0.028 992 226	2.014 554 908	2.689 440 046

Ejemplo 5.5 (Modelo hiperbólico). Cinética de Michaelis-Menten

5.1.5. Mínimos cuadrados para suma de funciones. Consideremos el caso general, supongamos que queremos aproximar la relación $y = f(x)$ como combinación de funciones conocidas de la forma $\hat{f}(x) = \alpha_1 \phi_1(x) + \cdots + \alpha_n \phi_n(x)$, donde $\phi_1(x), \dots, \phi_n(x)$ son funciones dadas. Planteamos la elección de los coeficientes basados en el criterio de mínimos cuadrados. Esto implica hallar $\alpha_1, \dots, \alpha_n$ de forma que sea mínimo $S = (y_1 - \hat{y}_1)^2 + \cdots + (y_N - \hat{y}_N)^2$, donde $\hat{y}_j = \hat{f}(x_j)$. Si derivamos respecto a cada coeficiente, en el mínimo se debe verificar

$$0 = \frac{\partial S}{\partial \alpha_i} = 2(y_1 - \hat{y}_1) \frac{\partial \hat{y}_1}{\partial \alpha_i} + \cdots + 2(y_N - \hat{y}_N) \frac{\partial \hat{y}_N}{\partial \alpha_i},$$

para $i = 1, \dots, n$, que se puede reescribir de la forma

$$(5.2) \quad y_1 \frac{\partial \hat{y}_1}{\partial \alpha_i} + \dots + y_N \frac{\partial \hat{y}_N}{\partial \alpha_i} = \hat{y}_1 \frac{\partial \hat{y}_1}{\partial \alpha_i} + \dots + \hat{y}_N \frac{\partial \hat{y}_N}{\partial \alpha_i}.$$

Dado que $\frac{\partial \hat{y}_j}{\partial \alpha_i} = \phi_i(x_j)$, el lado izquierdo vale

$$y_1 \frac{\partial \hat{y}_1}{\partial \alpha_i} + \dots + y_N \frac{\partial \hat{y}_N}{\partial \alpha_i} = y_1 \phi_i(x_1) + \dots + y_N \phi_i(x_N).$$

De la misma forma, el lado derecho se puede plantear

$$\begin{aligned} \hat{y}_1 \frac{\partial \hat{y}_1}{\partial \alpha_i} + \dots + \hat{y}_N \frac{\partial \hat{y}_N}{\partial \alpha_i} &= (\alpha_1 \phi_1(x_1) + \dots + \alpha_n \phi_n(x_1)) \phi_i(x_1) + \dots \\ &\quad + (\alpha_1 \phi_1(x_N) + \dots + \alpha_n \phi_n(x_N)) \phi_i(x_N) \\ &= (\phi_i(x_1) \phi_1(x_1) + \dots + \phi_i(x_N) \phi_1(x_N)) \alpha_1 + \dots \\ &\quad + (\phi_i(x_1) \phi_n(x_1) + \dots + \phi_i(x_N) \phi_n(x_N)) \alpha_n. \end{aligned}$$

Si definimos $\phi_{ij} = \phi_i(x_j)$, las ecuaciones (5.2) se escriben como

$$\sum_{j=1}^N \phi_{ij} y_j = \sum_{k=1}^n \left(\sum_{j=1}^N \phi_{ij} \phi_{kj} \right) \alpha_k$$

las que se conocen como las ecuaciones normales. Si $\Phi \in \mathbb{R}^{N \times n}$ es la matriz cuyos elementos son $(\Phi)_{jk} = \phi_k(x_j)$ y los vectores $\mathbf{y} \in \mathbb{R}^N$, $\boldsymbol{\alpha} \in \mathbb{R}^n$ dados por $\mathbf{y} = (y_1 \dots y_N)^T$, $\boldsymbol{\alpha} = (\alpha_1 \dots \alpha_n)^T$, las ecuaciones normales se escriben matricialmente como $\Phi^T \cdot \mathbf{y} = \Phi^T \cdot \Phi \cdot \boldsymbol{\alpha}$. Si la matriz $\Phi^T \cdot \Phi$ es inversible podemos despejar los coeficientes $\boldsymbol{\alpha} = (\Phi^T \cdot \Phi)^{-1} \Phi^T \mathbf{y}$.

5.1.6. Modelo polinomial. En muchas situaciones, la relación entre dos variables x e y no se puede expresar a través de una función lineal sino que es necesario utilizar otro tipo de funciones, como por ejemplo funciones polinomiales de mayor orden:

$$y = \hat{f}(x) = \alpha_0 + \alpha_1 x + \dots + \alpha_p x^p.$$

Igual que en el caso lineal, buscamos minimizar el error $S = \frac{1}{2}((y_1 - \hat{y}_1)^2 + \dots + (y_N - \hat{y}_N)^2)$, donde $\hat{y}_k = \hat{f}(x_k)$. Si definimos $\boldsymbol{\phi}_0 = (1 \dots 1)^T$, $\boldsymbol{\phi}_1 = (x_1 \dots x_N)^T$, \dots , $\boldsymbol{\phi}_p = (x_1^p \dots x_N^p)^T$, entonces $\hat{\mathbf{y}} = \alpha_0 \boldsymbol{\phi}_0 + \dots + \alpha_p \boldsymbol{\phi}_p$ y $S = \frac{1}{2} \|\mathbf{y} - \hat{\mathbf{y}}\|^2$. Como $\mathbf{y} - \hat{\mathbf{y}} \perp \boldsymbol{\phi}_k$, $k = 0, \dots, n$, obtenemos la ecuaciones normales

$$\begin{aligned} \mathbf{y} \cdot \boldsymbol{\phi}_0 &= \alpha_0 \boldsymbol{\phi}_0 \cdot \boldsymbol{\phi}_0 + \dots + \alpha_p \boldsymbol{\phi}_p \cdot \boldsymbol{\phi}_0, \\ \mathbf{y} \cdot \boldsymbol{\phi}_1 &= \alpha_0 \boldsymbol{\phi}_0 \cdot \boldsymbol{\phi}_1 + \dots + \alpha_p \boldsymbol{\phi}_p \cdot \boldsymbol{\phi}_1, \\ &\vdots \\ \mathbf{y} \cdot \boldsymbol{\phi}_p &= \alpha_0 \boldsymbol{\phi}_0 \cdot \boldsymbol{\phi}_p + \dots + \alpha_p \boldsymbol{\phi}_p \cdot \boldsymbol{\phi}_p. \end{aligned} \quad (5.3)$$

Para que el sistema tenga solución única, los vectores $\boldsymbol{\phi}_0, \dots, \boldsymbol{\phi}_p$ tienen que ser independientes. Si $\alpha_0 \boldsymbol{\phi}_0 + \dots + \alpha_p \boldsymbol{\phi}_p = \mathbf{0}$, tenemos $\hat{f}(x_1) = 0, \dots, \hat{f}(x_N) = 0$. Como $\hat{f}(x)$ es un polinomio de grado p , $\hat{f}(x)$ tiene que ser idénticamente nula, lo que equivale a $\alpha_0 = 0, \dots, \alpha_p = 0$.

Ejemplo 5.6. La ecuación de estado para un sistema compuesto de un mol de un gas ideal es $pV = RT$, donde p es la presión del gas, V el volumen y T la temperatura absoluta. R es una constante universal cuyo valor es $R = 8.314\,472 \text{ J mol}^{-1} \text{ K}^{-1}$. Dicho de otra forma $Z = pV/(RT)$

es una constante. Sin embargo, no existe ningún gas real que cumpla exactamente la ecuación anterior. Las diferencias son más notables a medida que la densidad del gas es mayor, lo que ocurre a temperaturas bajas y altas presiones. En [23] se estudia en forma experimental la ecuación de estado del neón para temperaturas en el rango 80 – 130 K y presiones de hasta 2×10^8 Pa. En la Tabla I de [23] se muestran valores experimentales de p y Z para distintas isothermas (T constante), reproducimos en la Tabla 5.6 los valores correspondientes a $T = 100$ K.

^{2 3} En [23] se ajusta por cuadrados mínimos un polinomio de grado 10. Por simplicidad y

p (Pa)	Z	p (Pa)	Z
$1.431\,78 \times 10^7$	1.023 97	$6.251\,05 \times 10^7$	1.762 59
$1.732\,78 \times 10^7$	1.052 36	$6.595\,53 \times 10^7$	1.820 34
$2.117\,82 \times 10^7$	1.097 05	$6.940\,02 \times 10^7$	1.878 00
$2.471\,87 \times 10^7$	1.144 61	$7.628\,99 \times 10^7$	1.992 51
$2.812\,12 \times 10^7$	1.193 88	$8.317\,94 \times 10^7$	2.106 23
$3.151\,05 \times 10^7$	1.245 90	$9.006\,95 \times 10^7$	2.218 83
$3.495\,48 \times 10^7$	1.301 04	$9.696\,03 \times 10^7$	2.330 30
$3.839\,90 \times 10^7$	1.357 25	$1.038\,51 \times 10^8$	2.440 45
$4.184\,32 \times 10^7$	1.414 25	$1.107\,41 \times 10^8$	2.549 83
$4.528\,76 \times 10^7$	1.471 68	$1.245\,23 \times 10^8$	2.765 54
$4.873\,22 \times 10^7$	1.529 74	$1.383\,07 \times 10^8$	2.978 11
$5.217\,66 \times 10^7$	1.588 02	$1.589\,83 \times 10^8$	3.290 10
$5.562\,12 \times 10^7$	1.646 41	$1.796\,61 \times 10^8$	3.596 27
$5.906\,58 \times 10^7$	1.704 55	$2.072\,35 \times 10^8$	3.997 17

Tabla 5.6: Datos de p y Z correspondientes a $T = 100$ K.

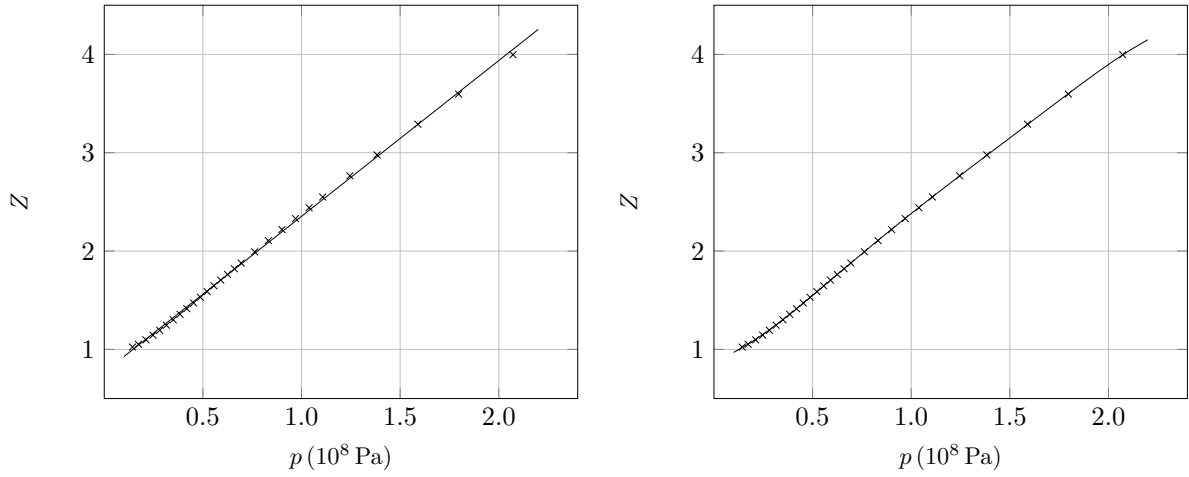
estabilidad numérica, vamos a considerar un ajuste quíntico. Proponemos una aproximación de la forma $Z = \alpha_0 + \alpha_1 x + \dots + \alpha_5 x^5$, donde x es la variable adimensional $x = p/1.0 \times 10^8$ Pa. Resolviendo la ecuaciones normales, obtenemos

$$Z = 0.888 + 0.598\,882 x + 2.376\,95 x^2 - 2.334\,79 x^3 + 1.019\,29 x^4 - 0.166\,425 x^5.$$

El valor del función error S es 2.748×10^{-4} . En la Figura 5.5 se muestran los resultados, comparamos la aproximación quíntica con la aproximación lineal $Z = 0.768\,940 + 1.584\,23x$, cuyo error es $S = 1.127 \times 10^{-2}$.

²La presión está expresada en atmósferas, recordemos $1 \text{ atm} = 101\,325 \text{ Pa}$.

³Observar que en la Tabla I de [23], los datos correspondientes a $P = 209.013 \text{ atm}$ ($2.117\,82 \times 10^7 \text{ Pa}$) y $T = 100 \text{ K}$ aparecen repetidos.

(a) Aproximación lineal: $S = 1.127 \times 10^{-2}$.(b) Aproximación quíntica: $S = 2.748 \times 10^{-4}$.Fig. 5.5: Gráfico de Z en función de p .

5.2. Modelo lineal multivariado. Supongamos que la variable y depende de varias variables x_1, x_2, \dots, x_p . Vamos a limitarnos al caso lineal, es decir nuestro modelo es

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_p x_p.$$

A partir de N mediciones de las variables (Tabla 5.7), definimos $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_p \in \mathbb{R}^N$ como

x_1	x_2	\dots	x_p	y
x_{11}	x_{12}	\dots	x_{1p}	y_1
x_{21}	x_{22}	\dots	x_{2p}	y_2
\vdots	\vdots	\ddots	\vdots	\vdots
x_{N1}	x_{N2}	\dots	x_{Np}	y_N

Tabla 5.7: Mediciones de las variables x_1, x_2, \dots, x_p, y .

$\mathbf{x}_0 = (1 \dots 1)^T, \mathbf{x}_1 = (x_{11} \dots x_{N1})^T, \dots, \mathbf{x}_p = (x_{1p} \dots x_{Np})^T$. Asumiendo que $\{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_p\}$ son independientes, planteamos la mejor aproximación en el sentido de cuadrados mínimos mediante las ecuaciones normales (5.3).

Ejemplo 5.7. Consideremos la Tabla 5.8, planteamos el modelo $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3$. La ecuaciones normales para este problema es

$$\begin{pmatrix} -271.920 \\ 5.115 \\ 446.130 \\ -3955.300 \end{pmatrix} = \begin{pmatrix} 10.000 & -0.172 & -15.594 & 145.460 \\ -0.172 & 0.184 & 2.627 & -2.483 \\ -15.594 & 2.627 & 337.430 & -226.650 \\ 145.460 & -2.483 & -226.650 & 2115.800 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}.$$

La solución del sistema es $\alpha_0 = 2.486, \alpha_1 = 1.880, \alpha_2 = 0.058, \alpha_3 = -2.032$, lo que establece la relación $y = 2.486 + 1.880 x_1 + 0.058 x_2 - 2.032 x_3$.

En muchos casos es importante conocer la influencia de la variable independiente x_i en la variable dependiente y . Un posible indicador de la importancia está dada por el valor del coeficiente α_i . Sin embargo, sin un análisis más profundo, podemos llegar a conclusiones erróneas sobre este punto.

Por ejemplo, si $\mathbf{x}_1 = (100.2 \ 99.3 \ \dots \ 101.0)^T$, es decir $x_{1j} = 100 + \xi_{1j}$ con $\xi_{1j} \cong 1$, podemos escribir el modelo considerando el cambio de variable $x_1 = 100 + \xi_1$, es decir

$$y = \alpha_0 + 100\alpha_1 + \alpha_1\xi_1 + \alpha_2x_2 + \dots + \alpha_px_p.$$

Haremos esto para cada variable $\xi_{ij} = x_{ij} - \mu_i$, donde $\mu_i = (x_{i1} + \dots + x_{iN})/N$ es el promedio de las observaciones. En la variable ξ_i hemos eliminado la componente continua, agregándole el valor medio al coeficiente α_0 .

Por otro lado, la influencia de cada variable depende del coeficiente, pero también del rango de valores de la variable. Por ejemplo, si ξ_i toma valores del orden de 1 y $\alpha_i = 3.5$, su influencia en el valor de y será mucho más significativa que la de la variable ξ_k si se verifica $\alpha_k = 3.5 \times 10^6$ pero ξ_k toma valores del orden de 1.0×10^{-9} . Entonces, para que el coeficiente α_i refleje la influencia de la variable ξ_i debemos normalizar sus valores, para eso consideramos la transformación $x_i^* = \sigma_i^{-1}\xi_i = \sigma_i^{-1}(x_i - \mu_i)$, donde $\sigma_i = N^{-1/2}\|\xi_i\|$. Tenemos entonces

$$y = \beta_0 + \beta_1x_1^* + \dots + \beta_px_p^*,$$

donde $\beta_0 = \alpha_0 + \alpha_1\mu_1 + \dots + \alpha_p\mu_p$ y $\beta_i = \sigma_i\alpha_i$, $i = 1, \dots, n$. Los vectores $\mathbf{x}_0^*, \mathbf{x}_1^*, \dots, \mathbf{x}_p^*$ definidos como $\mathbf{x}_0^* = \mathbf{x}_0$ y $\mathbf{x}_i = \sigma_i^{-1}\xi_i$, $i = 1, \dots, p$, son linealmente independientes y generan el mismo subespacio que los vectores $\mathbf{x}_0^*, \mathbf{x}_1^*, \dots, \mathbf{x}_p^*$. Además, $\mathbf{x}_i^* \cdot \mathbf{x}_i^* = N$ y $\mathbf{x}_0^* \cdot \mathbf{x}_i^* = 0$ si $i = 1, \dots, p$. Podemos normalizar \mathbf{y} , tomando $\psi_j = y_j - \eta$, donde $\eta = (y_1 + \dots + y_N)/N$ y definiendo $\sigma_y = N^{-1/2}\|\psi\|$, si $\mathbf{y}^* = \sigma_y^{-1}\psi$, las ecuaciones normales quedan

$$\begin{aligned} \mathbf{y}^* \cdot \mathbf{x}_0^* &= 0, \\ \mathbf{y}^* \cdot \mathbf{x}_1^* &= \beta_1 \mathbf{x}_1^* \cdot \mathbf{x}_1^* + \dots + \beta_p \mathbf{x}_p^* \cdot \mathbf{x}_1^*, \\ &\vdots \\ \mathbf{y}^* \cdot \mathbf{x}_p^* &= \beta_1 \mathbf{x}_1^* \cdot \mathbf{x}_p^* + \dots + \beta_p \mathbf{x}_p^* \cdot \mathbf{x}_p^*. \end{aligned}$$

Consideremos nuevamente la Tabla 5.8, la normalización de sus valores se muestran en la Tabla 5.9. Si $y = \alpha_1 x_1^* + \alpha_2 x_2^* + \alpha_3 x_3^*$ obtenemos $y^* = 0.539 x_1^* + 0.686 x_2^* - 0.298 x_3^*$. Vemos que, lejos de ser despreciable como podría sugerir el valor de $\alpha_2 = 0.058$ en el modelo sin normalizar, la partición de x_2^* es significativa.

5.3. Regresión de componentes principales. La regresión de componentes principales (PCR) es una técnica de análisis de regresión que se basa en el análisis de componentes principales. Por lo general, considera la regresión de la variable dependiente (también

n	x_1	x_2	x_3	y
1	0.146	0.287	14.458	-26.608
2	-0.108	7.031	14.428	-26.594
3	-0.160	-2.236	14.615	-27.559
4	0.006	4.022	14.639	-27.053
5	0.144	-5.107	14.574	-27.142
6	0.008	-5.486	14.598	-27.481
7	-0.176	-2.852	14.539	-27.627
8	0.191	6.584	14.606	-26.464
9	-0.033	-8.975	14.522	-27.559
10	-0.190	-8.862	14.478	-27.834

Tabla 5.8

conocida como la respuesta) de un conjunto de variables independientes (también conocidas como predictores o explicativas) basadas en un modelo de regresión lineal estándar, pero utilizando técnicas de componentes principales para estimar los coeficientes de regresión desconocidos en el modelo.

En lugar de obtener la variable dependiente directamente como función de las variables originales, se usan variables nuevas (sin significado físico claro en general) relacionadas linealmente con las anteriores. En muchos casos, se reduce la dimensión del problema tomando solo las variables más significativas.

El PCR puede lidiar adecuadamente con modelos que incluyen variables casi dependientes, excluyendo algunas de las componentes principales que tienen bajo peso en la regresión. La reducción del número efectivo de parámetros que caracterizan el modelo subyacente puede ser particularmente útil para problemas de alta dimensión.

5.3.1. Motivación gráfica. Consideremos los datos que se muestran la Tabla 5.10. Planteamos el modelo lineal $y = \alpha_1 x_1 + \alpha_2 x_2$, no sumamos una constante dado que las columnas tienen promedio nulo.

Las ecuaciones normales son

$$\begin{pmatrix} 36.3086 \\ 29.6934 \end{pmatrix} = \begin{pmatrix} 16.9833 & 14.7903 \\ 14.7903 & 14.8291 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix},$$

de donde obtenemos $\alpha_1 = 2.99906$, $\alpha_2 = -0.988843$ con $R^2 = 0.998048$.

Sin embargo, al graficar los puntos (x_{j1}, x_{j2}) para $j = 1, \dots, 18$ en la Figura 5.6(a), vemos que hay una fuerte relación entre las variables x_1, x_2 . En efecto, si planteamos $x_2 = \beta x_1$, obtenemos $x_2 = 0.870874 x_1$ con $R^2 = 0.8686$. Esta relación se grafica en línea punteada.

El ejemplo nos presenta dos interrogantes:

- ¿Es posible considerar una transformación que permita eliminar la dependencia entre las variables?
- Existiendo dependencia, ¿es necesario considerar a ambas variables?

En la Figura 5.6(b) mostramos las variables z_1, z_2 , donde no se observa una correlación importante entre ellas, lo que respondería a la primera pregunta. Podemos ver también la mayor influencia de la coordenada z_1 en la ubicación de los puntos (x_{j1}, x_{j2}) sobre el plano, lo que nos sugiere que podríamos usar solo esta variable para estimar y . Estudiaremos ahora la forma de hacer esto sistemáticamente.

n	x_1^*	x_2^*	x_3^*	y^*
1	1.211	0.330	-1.271	1.243
2	-0.673	1.535	-1.706	1.274
3	-1.062	-0.121	1.009	-0.781
4	0.176	0.997	1.355	0.295
5	1.199	-0.634	0.407	0.107
6	0.184	-0.702	0.758	-0.616
7	-1.182	-0.231	-0.102	-0.926
8	1.547	1.455	0.884	1.552
9	-0.116	-1.325	-0.350	-0.781
10	-1.283	-1.305	-0.984	-1.367

Tabla 5.9

n	x_1	x_2	y	n	x_1	x_2	y
1	-0.732 65	-0.360 50	-1.688 70	10	-0.928 70	-1.244 80	-1.479 70
2	1.251 80	1.472 80	2.218 50	11	-1.379 30	-0.829 41	-3.306 10
3	2.086 70	1.529 20	4.844 50	12	0.723 02	1.212 80	0.935 29
4	-1.398 20	-1.351 90	-3.003 60	13	0.310 71	0.183 90	0.727 82
5	-0.158 03	-0.536 76	-0.042 84	14	0.958 86	0.857 66	1.803 30
6	0.420 51	-0.215 43	1.523 40	15	-0.473 92	-0.613 30	-0.733 00
7	-1.219 20	-0.771 44	-2.888 50	16	-0.444 39	-0.484 58	-0.816 32
8	-0.914 58	-0.716 67	-2.077 60	17	0.745 41	1.076 20	1.279 70
9	0.352 04	-0.028 17	1.064 00	18	0.799 97	0.820 41	1.640 00

Tabla 5.10

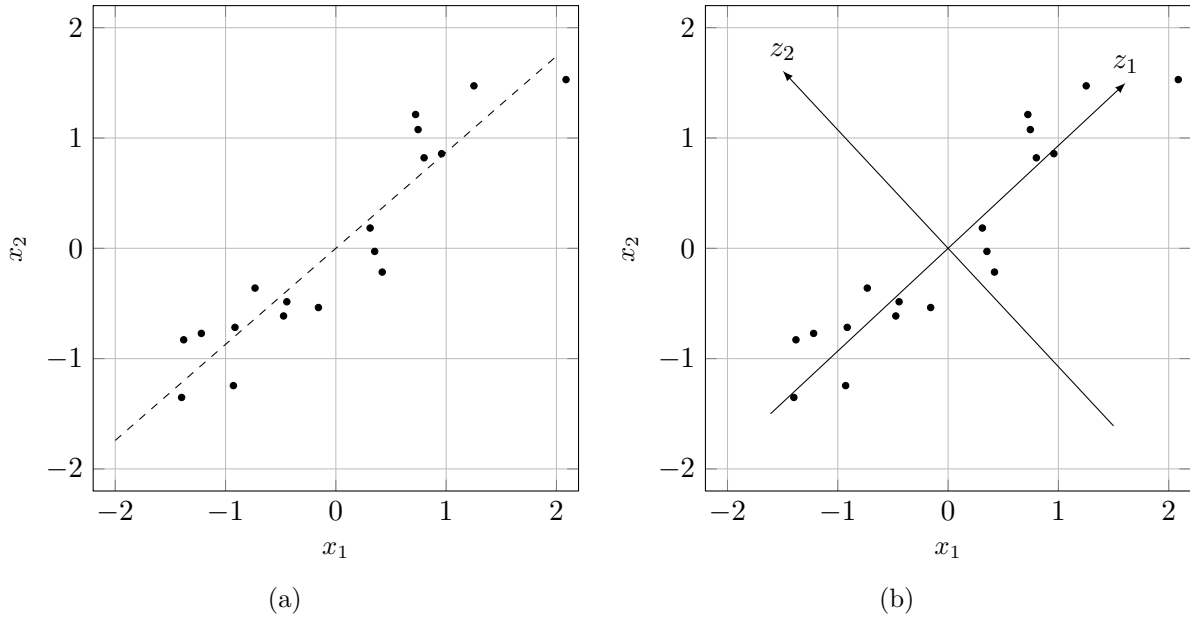


Fig. 5.6

5.3.2. Componentes principales mediante autovalores y autovectores. Si definimos la matriz $X = (x_1 \dots x_p) \in \mathbb{R}^{N \times p}$, como ya vimos los parámetros $\alpha = (\alpha_1 \dots \alpha_p)^T$ del modelo $y = \alpha_1 x_1 + \dots + \alpha_p x_p$ se obtienen resolviendo las ecuaciones normales $X^T y = X^T X \alpha$. La matriz $G = X^T X \in \mathbb{R}^{p \times p}$ es simétrica y semidefinida positiva, entonces sus autovalores son reales y no negativos, $\lambda_1 \geq \dots \geq \lambda_p \geq 0$. Además, tenemos p autovectores independientes u_1, \dots, u_p asociados a los autovalores, es decir $G u_j = \lambda_j u_j$. Recordemos que podemos tomar $u_j \cdot u_k = 0$ si $j \neq k$ y $u_j \cdot u_j = 1$. Si definimos la matriz $U = (u_1 \dots u_p) \in \mathbb{R}^{p \times p}$ y $\Lambda \in \mathbb{R}^{p \times p}$

$$\Lambda = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix},$$

tenemos $GU = U\Lambda$, y como $U^{-1} = U^T$, vale $G = U\Lambda U^T$. Si $Z = XU$, es decir $X = ZU^T$ y $X^T = UZ^T$, las ecuaciones normales se escriben como $UZ^T y = UZ^T ZU^T \alpha$. Simplificando U y usando que $Z^T Z = \Lambda$, las ecuaciones se escriben como $Z^T y = \Lambda \beta$, donde $\beta = U^T \alpha$. Las variables z_1, \dots, z_p se relacionan con x_1, \dots, x_p por $z_j = (x_1 \dots x_p) \cdot u_j$. Si $S = \|y - Z\beta\|^2$, tenemos

$$S = \|y\|^2 - 2 y \cdot (Z\beta) + \|Z\beta\|^2 = \|y\|^2 - 2 (Z^T y) \cdot \beta + (Z^T Z \beta) \cdot \beta = \|y\|^2 - (\Lambda \beta) \cdot \beta,$$

siendo $(\Lambda\beta).\beta = \lambda_1\beta_1^2 + \dots + \lambda_p\beta_p^2$, obtenemos $S = \|\mathbf{y}\|^2 - (\lambda_1\beta_1^2 + \dots + \lambda_p\beta_p^2)$.

En relación al ejemplo de la Tabla 5.10, tenemos $\lambda_1 = 30.7357$, $\lambda_2 = 1.07672$ y

$$U = \begin{pmatrix} -0.732336 & 0.680943 \\ -0.680943 & -0.732336 \end{pmatrix}$$

$$z_1 = -0.732336 x_1 + 0.680943 x_2, \quad z_2 = -0.680943 x_1 - 0.732336 x_2$$

n	z_1	z_2	y	n	z_1	z_2	y
1	0.782 02	-0.234 88	-1.688 70	10	1.527 78	0.279 24	-1.479 68
2	-1.919 61	-0.226 14	2.218 45	11	1.574 90	-0.331 82	-3.306 13
3	-2.569 48	0.301 03	4.844 47	12	-1.355 34	-0.395 84	0.935 29
4	1.944 52	0.037 89	-3.003 59	13	-0.352 77	0.076 90	0.727 82
5	0.481 23	0.285 48	-0.042 84	14	-1.286 22	0.024 83	1.803 26
6	-0.161 25	0.444 11	1.523 43	15	0.764 69	0.126 43	-0.733 00
7	1.418 20	-0.265 28	-2.888 47	16	0.655 41	0.052 27	-0.816 32
8	1.157 79	-0.097 93	-2.077 64	17	-1.278 73	-0.280 57	1.279 67
9	-0.238 63	0.260 35	1.063 98	18	-1.144 50	-0.056 08	1.639 99

Tabla 5.11

El modelo de regresión es en estas variables $y = -1.52297 z_1 + 2.76635 z_2$,

5.3.3. Reducción de la dimensión.

5.4. Métodos de agrupamiento. La agrupación en clúster es una técnica de aprendizaje automático que implica la agrupación de puntos de datos. Dado un conjunto de puntos de datos, podemos usar un algoritmo de agrupamiento para clasificar cada punto de datos en un grupo específico. En teoría, los puntos de datos que están en el mismo grupo deben tener propiedades y / o características similares, mientras que los puntos de datos en diferentes grupos deben tener propiedades y / o características muy diferentes. La agrupación en clústeres es un método de aprendizaje no supervisado y es una técnica común para el análisis estadístico de datos que se utiliza en muchos campos. En Data Science, podemos usar el análisis de agrupamiento para obtener información valiosa de nuestros datos al ver en qué grupos caen los puntos de datos cuando aplicamos un algoritmo de agrupamiento. ¡Hoy, vamos a ver 5 algoritmos de agrupamiento populares que los científicos de datos necesitan saber y sus pros y contras!

Dado un conjunto finito $S \subset \mathbb{R}^d$ de datos, queremos hallar clases $C_1, \dots, C_k \subseteq S$ que verifiquen

- Cada clase C_i contiene por lo menos un elemento, $C_i \neq \emptyset$.
- Dos clases distintas C_i, C_j no tienen puntos en común, $C_i \cap C_j = \emptyset$.
- Todo punto de S pertenece a (exactamente) una clase, $C_1 \cup \dots \cup C_k = S$.
- Los puntos pertenecientes a una misma clase tienen características similares y muy diferentes a los de las otras.

Como vemos, la última condición admite múltiples interpretaciones y debe ser bien definida para poder hacer la clasificación.

5.4.1. Ejemplos gráficos.

5.4.2. Método de k-medias. Este método separa el conjunto $S \subset \mathbb{R}^d$ de datos en un número k fijo de clases minimizando la suma de las distancias al cuadrado entre los puntos de una clase y su baricentro o centroide. Para cada clase C_i definimos el baricentro como

$$\mathbf{b}_i = \frac{1}{\#C_i} \sum_{\mathbf{x} \in C_i} \mathbf{x},$$

donde $\#C_i$ es el número de elementos de la clase C_i , la partición en las clases $C_1, \dots, C_k \subseteq S$ debe minimizar la función \mathcal{W} (within-cluster sum of squares, WCSS) definida por

$$\mathcal{W}(C_1, \dots, C_k) = \sum_{i=1}^k w(C_i) = \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mathbf{b}_i\|^2.$$

La elección del baricentro se relaciona con la definición de w , siendo que

$$w(C_i) = \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mathbf{b}_i\|^2 = \min_{\mathbf{p} \in \mathbb{R}^d} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mathbf{p}\|^2$$

$$\begin{aligned} w(C_i \cup C_j) &= \min_{\mathbf{p} \in \mathbb{R}^d} \sum_{\mathbf{x} \in C_i \cup C_j} \|\mathbf{x} - \mathbf{p}\|^2 \geq \min_{\mathbf{p}_i \in \mathbb{R}^d} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mathbf{p}_i\|^2 + \min_{\mathbf{p}_j \in \mathbb{R}^d} \sum_{\mathbf{x} \in C_j} \|\mathbf{x} - \mathbf{p}_j\|^2 \\ &= w(C_i) + w(C_j), \end{aligned}$$

esto muestra que la minimización no puede tener ninguna clase vacía. También podemos ver que si k aumenta, el valor mínimo disminuye. Por otro lado, para la partición mínima se verifica que si $\mathbf{x} \in C_i$, entonces $\|\mathbf{x} - \mathbf{b}_i\| \leq \|\mathbf{x} - \mathbf{b}_j\|$ para $j \neq i$. En efecto, si $\mathbf{x} \in C_i$ pero $\|\mathbf{x} - \mathbf{b}_i\| > \|\mathbf{x} - \mathbf{b}_j\|$ para algún $j \neq i$, podemos definir una nueva partición: $C'_i = C_i - \{\mathbf{x}\}$, $C'_j = C_j \cup \{\mathbf{x}\}$, $C'_l = C_l$, vamos a mostrar que $\mathcal{W}(C'_1, \dots, C'_k) < \mathcal{W}(C_1, \dots, C_k)$. Si \mathbf{b}'_i es el centroide de C'_i y \mathbf{b}'_j el de C'_j , entonces $w(C'_i) = w(C_i)$ y

$$\begin{aligned} w(C'_i) &= \sum_{\mathbf{y} \in C'_i} \|\mathbf{y} - \mathbf{b}'_i\|^2 \leq \sum_{\mathbf{y} \in C'_i} \|\mathbf{y} - \mathbf{b}_i\|^2 \leq \sum_{\mathbf{y} \in C_i} \|\mathbf{y} - \mathbf{b}_i\|^2 - \|\mathbf{x} - \mathbf{b}_i\|^2, \\ w(C'_j) &= \sum_{\mathbf{y} \in C'_j} \|\mathbf{y} - \mathbf{b}'_j\|^2 \leq \sum_{\mathbf{y} \in C'_j} \|\mathbf{y} - \mathbf{b}_j\|^2 \leq \sum_{\mathbf{y} \in C_j} \|\mathbf{y} - \mathbf{b}_j\|^2 + \|\mathbf{x} - \mathbf{b}_j\|^2, \end{aligned}$$

de donde obtenemos $\mathcal{W}(C'_1, \dots, C'_k) \leq \mathcal{W}(C_1, \dots, C_k) + \|\mathbf{x} - \mathbf{b}_j\|^2 - \|\mathbf{x} - \mathbf{b}_i\|^2$.

El algoritmo consiste en hallar una partición estable para las igualdades

$$\begin{aligned} \mathbf{b}_i &= \frac{1}{\#C_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}, \\ C_i &= \{\mathbf{x} \in S : \|\mathbf{x} - \mathbf{b}_i\| \leq \|\mathbf{x} - \mathbf{b}_j\|\}, \end{aligned}$$

observando que debe definirse un criterio para el caso $\|\mathbf{x} - \mathbf{b}_i\| = \|\mathbf{x} - \mathbf{b}_j\|$.

Una partición estable no es necesariamente óptima como vemos en el siguiente ejemplo

Ejemplo 5.8. Consideramos particiones del conjunto $S = \{-2, -1, 1/5, 1, 2\} \subset \mathbb{R}$ en dos clases, podemos ver que las particiones

(a) $C_1 = \{-2, -1, 1/5\}$, $C_2 = \{1, 2\}$, con centroides $b_1 = -14/15$, $b_2 = 3/2$,

(b) $C_1 = \{-2, -1\}$, $C_2 = \{1/5, 1, 2\}$, con centroides $b_1 = -3/2$, $b_2 = 16/15$,

son ambas estables. Tenemos $\mathcal{W} = 2.71$ para partición (a) y $\mathcal{W} = 2.31$ para (b).

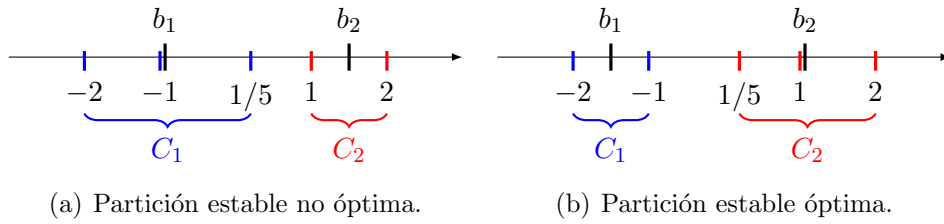


Fig. 5.7

Con un poco de trabajo adicional se puede probar que la partición (b) es óptima. En efecto, si una partición óptima verifica $b_1 < b_2$, entonces todo punto de C_1 tiene que ser menor que cualquier punto de C_2 . En caso contrario, intercambiándolos obtendríamos un menor valor de \mathcal{W} . En la Tabla 5.12 se muestran estas particiones, las estables y la óptima.

C_1	C_2	b_1	b_2	C'_1	C'_2	est.	opt.
$\{-2\}$	$\{-1, 1/5, 1, 2\}$	-2	11/20	$\{-2, -1\}$	$\{1/5, 1, 2\}$	no	—
$\{-2, -1\}$	$\{1/5, 1, 2\}$	-3/2	16/15	$\{-2, -1\}$	$\{1/5, 1, 2\}$	si	si
$\{-2, -1, 1/5\}$	$\{1, 2\}$	-14/15	3/2	$\{-2, -1, 1/5\}$	$\{1, 2\}$	si	no
$\{-2, -1, 1/5, 1\}$	$\{2\}$	-9/20	2	$\{-2, -1, 1/5\}$	$\{1, 2\}$	no	—

Tabla 5.12: Posibles particiones de $S = \{-2, -1, 1/5, 1, 2\}$.

El algoritmo de k -medias se describe en el Algoritmo 5.1 (el cálculo de los centroides y la clasificación en 5.2). La elección de los punto iniciales se hace en forma aleatoria dentro del conjunto S , esto asegura que ninguna clase sea vacía al comenzar. La forma de llevarlo a cabo se discute más adelante. La condición de parada es la estabilidad de la partición. Sin embargo, podría considerarse los cambios de los centroides: dado $\epsilon > 0$ pequeño, si los cambios de los centroide en un paso no supera ϵ , asumimos que alcanzamos una partición estable. Esto puede evitar costos computacionales altos, sin mejoras significativas en el resultado.

Algoritmo 5.1: k -medias.

Data: S, k
Result: $b_1, \dots, b_k, C_1, \dots, C_k$
for $i = 0$ **to** k **do**
 $b_i = \text{rand}(S)$;
 $C'_i = \emptyset$;
end
 $(C_1, \dots, C_k) = \text{clasif}(S, (b_1, \dots, b_k))$;
while $C_1 \neq C'_1 \vee \dots \vee C_k \neq C'_k$ **do**
 for $i = 0$ **to** k **do**
 $b_i = \text{centroide}(C_i)$;
 $C'_i = C_i$;
 end
 $(C_1, \dots, C_k) = \text{clasif}(S, (b_1, \dots, b_k))$;
end

Las funciones `centroide` y `clasif` se muestran en 5.2, la función `rand` se discute más adelante.

Algoritmo 5.2: funciones centroide y clasif.

```

/* cálculo del centroide                                     */
def centroide(C):
    b = 0;
    for x ∈ C do
        b = b + x;
    end
    b = b/length(C);
    return b;
/* ***** */
/* clasificación                                           */
def clasif(S, (b1, ..., bk)):
    for x ∈ S do
        i = arg mín{||x - b1||, ..., ||x - bk||};
        Ci = Ci ∪ {x};
    end
    return (C1, ..., Ck);

```

En las Figuras 5.8, 5.9, 5.10, mostramos al aplicación del algoritmo a una conjunto de $N = 200$ puntos del plano. Se consideran particiones en $k = 3$ subconjuntos, inicialmente se eligen tres puntos arbitrarios (alineados horizontalmente en este caso) y se clasifican los puntos pertenecientes a cada región. Es usual elegir puntos del mismo conjunto S para asegurar que ninguna clase sea vacía. En línea punteada se grafican los segmentos de rectas que separan cada región, formando el diagrama de Voronoi asociado a los puntos b_1, b_2, b_3 .

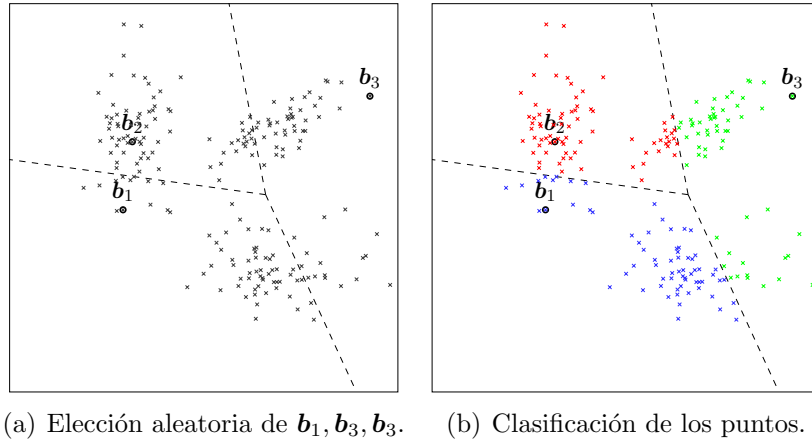


Fig. 5.8: Elección aleatoria de los puntos iniciales y diagrama de Voronoi.

El método de k -medias es fácil de implementar y las iteraciones requieren pocos cálculos. Como desventaja podemos señalar que es necesario indicar el número de clases de la partición. Además, la partición estable puede depender de la elección de los centros iniciales, lo que puede producir diferentes resultados en diferentes ejecuciones del algoritmo. Para determinar k , hallamos \mathcal{W}_k para diferentes valores de k . Sabemos que es decreciente, fijamos k en un valor donde la diferencia entre \mathcal{W}_k y \mathcal{W}_{k+1} no es significativo. De la figura 5.11 se observa un fuerte decaimiento de \mathcal{W}_k al pasar de $k = 2$ a $k = 3$, que se estabiliza para $k \geq 3$. Algunos de los principales inconvenientes de este método son:

- No da buenos resultados si los puntos de un grupo están muy cercanos al centroide de otro.
- Dependen de la forma, tiende a separar en grupos esféricos.

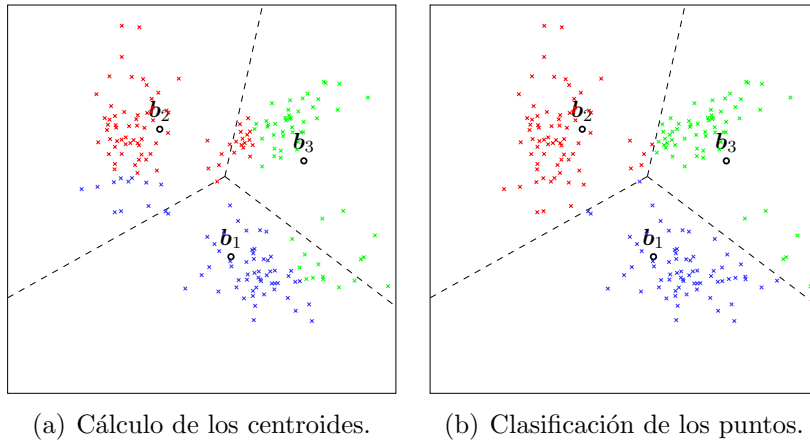


Fig. 5.9: Primer paso.

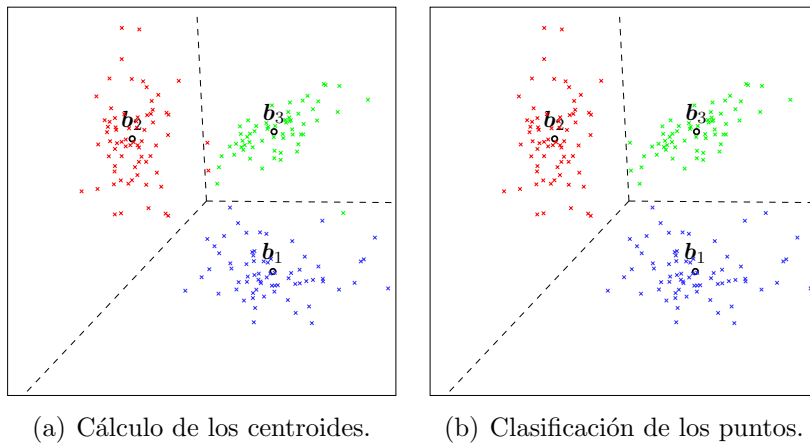


Fig. 5.10: Cuarto paso: partición estable.

- Solo obtiene mínimos locales, los resultados dependen de los puntos iniciales.
- No es robusto: datos atípicos modifican fuertemente las clases.
- Debe darse el número de clases.

Inicialización

Una posible forma de obtener los puntos iniciales es elegir k puntos (sin reposición) de S con igual probabilidad. Correindo varias veces el algoritmo y comparando los valores de \mathcal{W} obtenidos y seleccionando el mejor, podemos esperar que se obtenga el mínimo global. Esto puede dar lugar a punto iniciales muy próximos

Suma de cuadrados totales, intergrupos e internos

Dado un conjunto de datos S y $\{C_1, \dots, C_n\}$ una partición de S , definimos la suma de cuadrados totales (SST) $\mathcal{T} = \sum_{x \in S} \|x - \mathbf{b}\|^2$, y la suma de intergrupos (BSS) $\mathcal{B} = \sum_{i=1}^k \#C_i \|\mathbf{b}_i - \mathbf{b}\|^2$, donde \mathbf{b} es el baricentro de S y \mathbf{b}_i el baricentro de C_i . Podemos ver que $\mathcal{T} = \mathcal{W} + \mathcal{B}$, por lo tanto $0 \leq \mathcal{W}/\mathcal{T} \leq 1$, que se puede considerar como un criterio de bondad de la partición. En el ejemplo estudiado obtenemos: $\mathcal{W} = 780.647$, $\mathcal{B} = 3346.26$, $\mathcal{T} = 4126.9$, por lo tanto $\mathcal{W}/\mathcal{T} = 0.19$.

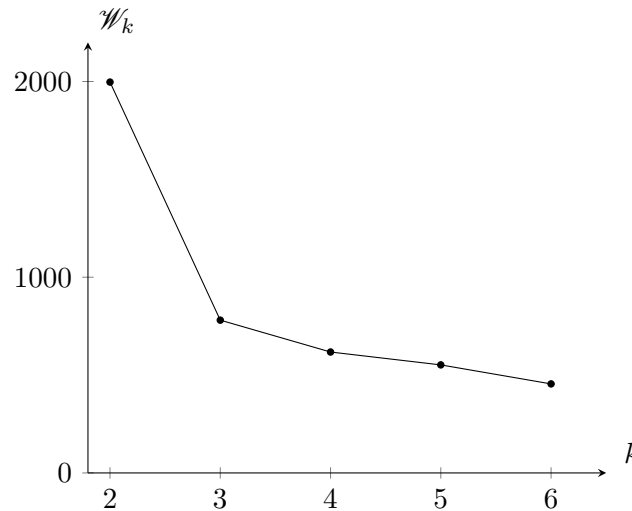


Fig. 5.11: \mathcal{W}_k en función del número de particiones.

5.4.3. Algoritmo de desplazamiento de medias. Este algoritmo se basa en ventanas deslizantes que intentan encontrar áreas densas de puntos de datos. Es un algoritmo basado en centroide que significa que el objetivo es ubicar los puntos centrales de cada grupo, que funciona actualizando los candidatos para que los puntos centrales sean la media de los puntos dentro de la ventana deslizante. Estas ventanas candidatas se filtran en una etapa de postprocesamiento para eliminar casi duplicados, formando el conjunto final de puntos centrales y sus grupos correspondientes. Mira el gráfico a continuación para ver una ilustración.

5.4.4. DBSCAN. DBSCAN es un algoritmo agrupado basado en la densidad similar al desplazamiento medio, pero con un par de ventajas notables. ¡Mira otro gráfico elegante a continuación y comencemos!

1. DBSCAN comienza con un punto de datos de inicio arbitrario que no ha sido visitado. La vecindad de este punto se extrae usando una distancia ϵ (Todos los puntos que están dentro de la distancia ϵ son puntos de vecindad).
2. Si hay un número suficiente de puntos (de acuerdo con minPoints) dentro de esta vecindad, se inicia el proceso de agrupación y el punto de datos actual se convierte en el primer punto en el nuevo clúster. De lo contrario, el punto se etiquetará como ruido (más tarde este punto ruidoso podría convertirse en parte del clúster). En ambos casos, ese punto está marcado como "visitado".
3. Para este primer punto en el nuevo grupo, los puntos dentro de su vecindario de distancia ϵ también se vuelven parte del mismo grupo. Este procedimiento de hacer que todos los puntos de la vecindad ϵ pertenezcan al mismo grupo se repite para todos los puntos nuevos que se acaban de agregar al grupo de grupos.
4. Este proceso de los pasos 2 y 3 se repite hasta que se determinen todos los puntos en el grupo, es decir, se hayan visitado y etiquetado todos los puntos dentro de la vecindad ϵ del grupo.
5. Una vez que hayamos terminado con el clúster actual, se recupera y procesa un nuevo punto no visitado, lo que conduce al descubrimiento de un clúster o ruido adicional. Este proceso se repite hasta que todos los puntos estén marcados como visitados. Dado que al final de esto todos los puntos han sido visitados, cada punto habrá sido marcado como perteneciente a un grupo o como ruido.

DBSCAN presenta algunas grandes ventajas sobre otros algoritmos de agrupamiento. En primer lugar, no requiere un número fijo de clústeres en absoluto. También identifica los valores atípicos como ruidos, a diferencia del cambio medio que simplemente los arroja a un clúster incluso si el punto de datos es muy diferente. Además, puede encontrar grupos de tamaño arbitrario y de forma arbitraria bastante bien. El principal inconveniente de DBSCAN es que no funciona tan bien como los demás cuando los grupos son de densidad variable. Esto se debe a que la configuración del umbral de distancia ϵ y minPoints para identificar los puntos vecinos variará de un grupo a otro cuando la densidad varíe. Este inconveniente también ocurre con datos de muy alta dimensión, ya que nuevamente el umbral de distancia ϵ se vuelve difícil de estimar.

5.4.5. Agrupamiento jerárquico. Las estrategias jerárquicas aglomerativas construyen una jerarquía de agrupamientos, representada gráficamente por un árbol llamado dendograma. Esta árbol parte de las clases formadas por un solo individuo, y en cada paso va uniendo dos clases, hasta conseguir el grupo formado por todos los puntos. La elección de las clases a unir se realiza considerando las dos más cercanas en algún sentido. Existen diferentes nociones de distancia entre subconjuntos, las más utilizadas son la de vecino más cercano (SL), vecino más lejano (CL), distancia promedio (AL) y la distancia entre centroides (C).

5.4.6. Distancia entre clases. Como mencionamos anteriormente, existen varias definiciones diferentes de distancia entre dos clases $C_i, C_j \subset S$, en la Tabla 5.13 mostramos las definiciones de las más usadas

SL	$d_{\text{SL}}(C_i, C_j) = \min_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} \ \mathbf{x} - \mathbf{y}\ $
CL	$d_{\text{CL}}(C_i, C_j) = \max_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} \ \mathbf{x} - \mathbf{y}\ $
AL	$d_{\text{AL}}(C_i, C_j) = \frac{1}{\#C_i \#C_j} \sum_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} \ \mathbf{x} - \mathbf{y}\ $
Centroides	$d_{\text{C}}(C_i, C_j) = \ \mathbf{b}_i - \mathbf{b}_j\ $

Tabla 5.13

Como el algoritmo de agrupamiento jerárquico une las dos clases más cercanas, es necesario calcular las distancias de esta nueva clase a las demás para continuar. Es fácil ver que esto no requiere gran esfuerzo, ya que se pueden obtener las nuevas distancias a partir de las ya conocidas mediante las igualdades

$$\begin{aligned}
 d_{\text{SL}}(C_i \cup C_j, C_l) &= \min\{d_{\text{SL}}(C_i, C_l), d_{\text{SL}}(C_j, C_l)\}, \\
 d_{\text{CL}}(C_i \cup C_j, C_l) &= \max\{d_{\text{CL}}(C_i, C_l), d_{\text{CL}}(C_j, C_l)\}, \\
 d_{\text{AL}}(C_i \cup C_j, C_l) &= \frac{\#C_i}{\#C_i + \#C_j} d_{\text{AL}}(C_i, C_l) + \frac{\#C_j}{\#C_i + \#C_j} d_{\text{AL}}(C_j, C_l).
 \end{aligned}$$

Para la distancia d_{C} también existe una igualdad que permite el cálculo sencillo de las distancias entre las clases de la nueva partición, pero es un poco más difícil de justificar⁴:

$$\begin{aligned}
 d_{\text{C}}(C_i \cup C_j, C_l)^2 &= \frac{\#C_i}{\#C_i + \#C_j} d_{\text{C}}(C_i, C_l)^2 + \frac{\#C_j}{\#C_i + \#C_j} d_{\text{C}}(C_j, C_l)^2 \\
 &\quad - \frac{\#C_i \#C_j}{(\#C_i + \#C_j)^2} d_{\text{C}}(C_i, C_j)^2.
 \end{aligned}$$

⁴Las igualdades anteriores valen para cualquier norma, mientras que para d_{C} solo vale para la norma euclídea.

Observemos que en todos los casos se verifica $d(C_i \cup C_j, C_l) \leq \max\{d(C_i, C_l), d(C_j, C_l)\}$.

Distancia entre baricentros Vamos a mostrar la relación de la distancia entre una unión de dos clases y la distancia de cada una. Si \mathbf{b}_i es el centroide de C_i y \mathbf{b}_j es el centroide de C_j , entonces el centroide \mathbf{b} de $C_i \cup C_j$ es

$$\mathbf{b} = \frac{1}{\#(C_i \cup C_j)} \sum_{\mathbf{x} \in C_i \cup C_j} \mathbf{x} = (1 - \lambda) \frac{1}{\#C_i} \sum_{\mathbf{x} \in C_i} \mathbf{x} + \lambda \frac{1}{\#C_j} \sum_{\mathbf{x} \in C_j} \mathbf{x} = (1 - \lambda)\mathbf{b}_i + \lambda\mathbf{b}_j,$$

donde $\lambda = \#C_j/(\#C_i + \#C_j)$. Si escribimos $\|\mathbf{b} - \mathbf{b}_l\| = \|(1 - \lambda)(\mathbf{b}_i - \mathbf{b}_l) + \lambda(\mathbf{b}_j - \mathbf{b}_l)\|$, usando que

$$\|(1 - \lambda)\mathbf{x} + \lambda\mathbf{y}\|^2 = (1 - \lambda)\|\mathbf{x}\|^2 + \lambda\|\mathbf{y}\|^2 - \lambda(1 - \lambda)\|\mathbf{x} - \mathbf{y}\|^2,$$

obtenemos la igualdad tomando $\mathbf{x} = \mathbf{b}_i - \mathbf{b}_l$ y $\mathbf{y} = \mathbf{b}_j - \mathbf{b}_l$.

Ejemplo 5.9. Consideremos el conjunto $S = \{0, 2, 5.5, 6.5, 8\} \subset \mathbb{R}$, la matriz de distancias se muestra en la Tabla 5.14 (izquierda). Queda claro que el primer agrupamiento es $\{\mathbf{x}_3, \mathbf{x}_4\}$.

d_{SL}	\mathbf{x}_1	\mathbf{x}_2	\mathbf{x}_3	\mathbf{x}_4	\mathbf{x}_5	d_{SL}	\mathbf{x}_1	\mathbf{x}_2	$\{\mathbf{x}_3, \mathbf{x}_4\}$	\mathbf{x}_5
\mathbf{x}_1	0	2	5.5	6.5	8	\mathbf{x}_1	0	2	5.5	8
\mathbf{x}_2		0	3.5	4.5	6	\mathbf{x}_2		0	3.5	6
\mathbf{x}_3			0	1	2.5	$\{\mathbf{x}_3, \mathbf{x}_4\}$			0	1.5
\mathbf{x}_4				0	1.5	\mathbf{x}_5				0
\mathbf{x}_5					0					

Tabla 5.14: Distancia entre elementos (izquierda) y distancia entre clases (derecha).

Para continuar debemos definir la distancia entre clases, para la distancia del vecino más cercano (d_{SL}) obtenemos las distancias de la Tabla 5.14 (derecha). Vemos que el nuevo agrupamiento es $\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$. Continuando con el algoritmo obtenemos las distancias siguientes:

d_{SL}	\mathbf{x}_1	\mathbf{x}_2	$\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$	d_{SL}	$\{\mathbf{x}_1, \mathbf{x}_2\}$	$\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$
\mathbf{x}_1	0	2	5.5	$\{\mathbf{x}_1, \mathbf{x}_2\}$	0	3.5
\mathbf{x}_2		0	3.5	$\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$		0
$\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$			0			

En la Figura 5.12 mostramos gráficamente las clases obtenidas en cada paso. Este gráfico se denomina dendograma, cortando el dendograma a una altura, obtenemos una partición de S , en la Figura se muestra un corte a altura $d_{SL} = 3$ que da como resultado $\{\mathbf{x}_1, \mathbf{x}_2\}$ y $\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$.

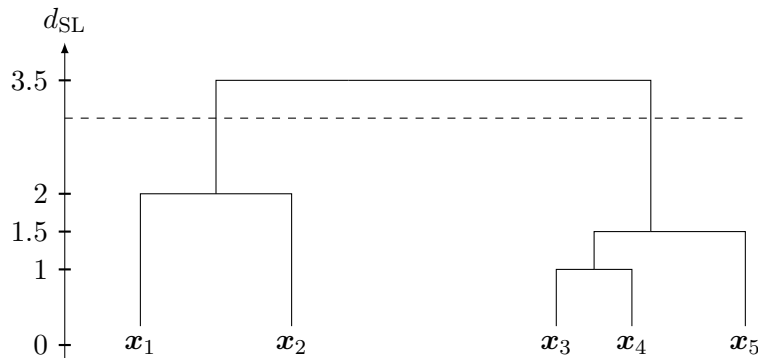


Fig. 5.12: Dendograma correspondiente a $S = \{0, 2, 5.5, 6.5, 8\}$.

Distancia de Hamming

Solo para la distancia d_C es importante que el conjunto sea \mathbb{R}^d y la distancia esté dada por la norma euclídea. Podemos estudiar los problemas de agrupamiento en un conjunto arbitrario

donde se defina una distancia. Como ejemplo, vamos a considerar la distancia de Hamming entre palabras. Si $\mathcal{A} = \{\alpha_1, \dots, \alpha_q\}$ es un alfabeto, la distancia entre las secuencias de longitud p , $a = a_1 \dots a_p \in \mathcal{A}^p$ se define de la siguiente forma dadas dos secuencias $a, b \in \mathcal{A}^p$ la distancia es la cantidad de elementos distintos en la misma posición, es decir $d_H(a, b) = r$ si existen exactamente r posiciones $1 \leq j_1 < \dots < j_r \leq p$ tales que $a_{j_1} \neq b_{j_1}, \dots, a_{j_r} \neq b_{j_r}$ y $a_j = b_j$ si $j \neq j_1, \dots, j_r$. Por ejemplo, con el alfabeto latino $\mathcal{A} = \{a, \dots, z\}$, las palabras “perro” y “perno” pertenecientes a \mathcal{A}^5 están a distancia 1, pero “perro” y “peras” se encuentran a distancia 2. Esta definición no permite definir la distancia entre palabras de longitudes diferentes. Para hacer esto, redefinamos la distancia de Hamming, si consideramos la operación sobre una palabra dada por la modificación de una letra por otra letra del alfabeto, la distancia es igual a la cantidad mínima de operaciones necesarias para modificar una palabra en la otra:

$$\begin{aligned} \text{“perro”} & \xrightarrow{+1} \text{“perao”} \\ \text{“perao”} & \xrightarrow{+1} \text{“peras”}. \end{aligned}$$

De esta forma, podemos analizar la distancia entre palabras de distinta longitud agregando como operaciones elementales agregar y eliminar una letra. Por ejemplo, la distancia entre “pera” y “perros”

$$\begin{aligned} \text{“pera”} & \xrightarrow{+1} \text{“pero”} \\ \text{“pero”} & \xrightarrow{+1} \text{“perro”} \\ \text{“perro”} & \xrightarrow{+1} \text{“perros”} \end{aligned}$$

es entonces 3, dado que no podemos hacer la transformación con una o dos operaciones elementales). Esto define una distancia d_L (distancia de Levenshtein) entre palabras del conjunto $\mathcal{A}^* = \bigcup_{p \geq 0} \mathcal{A}^p$, donde \mathcal{A}^0 es la palabra de longitud nula, es decir \emptyset . Se define en \mathcal{A}^* la operación concatenación \diamond : si $a = a_1 a_2 \dots a_m \in \mathcal{A}^m$ y $b = b_1 b_2 \dots b_n \in \mathcal{A}^n$, $a \diamond b = a_1 a_2 \dots a_m b_1 b_2 \dots b_n \in \mathcal{A}^{m+n}$ y $a \diamond \emptyset = \emptyset \diamond a = a$.

CAPÍTULO 6

Análisis de Fourier y Filtros

“Haffner gime dolorosamente. El infierno se ha dado cita a la orilla de su cama. Un rectángulo negro gira ante sus ojos, y alguien escribe con una tiza: $\cos \alpha + i \sin \alpha$ ”

Roberto Arlt [2]

6.1. Señales periódicas. Como ya vimos, muchos sistemas, físicos, químicos, biológicos, etc., presentan un comportamiento regular, es decir el estado del sistema se repite cada un tiempo fijo llamado período. Esto significa que las variables que describen al sistema, como por ejemplo posición, velocidad, presión, concentración, etc., son funciones periódicas del tiempo. Generalmente, esto no deja de ser una aproximación, el movimiento de un sistema comienza en algún momento y termina alguna vez. Pero dependerá de la escala temporal de las variaciones del sistema respecto a la duración del fenómeno, para que podamos considerar o no, que el sistema tiene un comportamiento periódico. Por ejemplo, si observamos la órbita del planeta Júpiter durante unos pocos siglos, no cometeremos un error significativo si suponemos que su movimiento es periódico con período de 11 años, 315 días, 1 hora y 6 minutos. En otros casos, la dinámica del sistema está muy lejos de ser periódica, en ese caso decimos que la señal observada es aperiódica, postergamos la discusión de tales señales a otra sección de este capítulo. Para tratar estos dos casos usamos herramientas diferentes, aunque relacionadas. En un sentido que habrá que justificar, podemos pensar que una señal aperiódica es una señal periódica de período infinito.

6.1.1. Señales armónicas. El más simple del comportamiento periódico es el armónico simple, $x(t) = \hat{x} \sin(2\pi t/T + \varphi)$, donde \hat{x} es la amplitud máxima, T es el período y φ la fase inicial. Definimos la frecuencia $\nu = 1/T$ y la frecuencia angular $\omega = 2\pi/T = 2\pi\nu$, podemos escribir $x(t) = \hat{x} \sin(2\pi\nu t + \varphi) = \hat{x} \sin(\omega t + \varphi)$. Las unidades de $x(t)$ depende de la magnitud considerada (milímetros, voltios, amperios, etc.), pero en general la unidad para t es el segundo (o alguna fracción: milisegundos, microsegundos, etc.), que se denota por s. Definimos la unidad de frecuencia $[\nu] = \text{s}^{-1} = \text{Hz}$, denominada Hercio¹ (Hertz en inglés), la unidad de la frecuencia angular es $[\omega] = \text{rad s}^{-1}$.

Observemos que $x(t) = \hat{x} \sin(\omega t + \varphi) = a \cos(\omega t) + b \sin(\omega t)$, con $a = \hat{x} \sin(\varphi)$ y $b = \hat{x} \cos(\varphi)$. Vemos que a partir de a y b podemos obtener $\hat{x} = (a^2 + b^2)^{1/2}$ y $\tan(\varphi) = a/b$. Geométricamente podemos interpretar las relaciones anteriores en el triángulo mostrado de la Figura 6.2(b)

Ejemplo 6.1. La tensión eléctrica domiciliaria $E(t)$ es de 220V (voltios) y 50Hz. Hay que

¹En honor a Heinrich Rudolf Hertz.

aclarar que 220 no es el valor máximo \hat{E} , sino lo que se conoce como valor eficaz. Para una señal senoidal, vale $\hat{E} = \sqrt{2} E_{\text{ef}}$ (ver subsección 6.1.7).

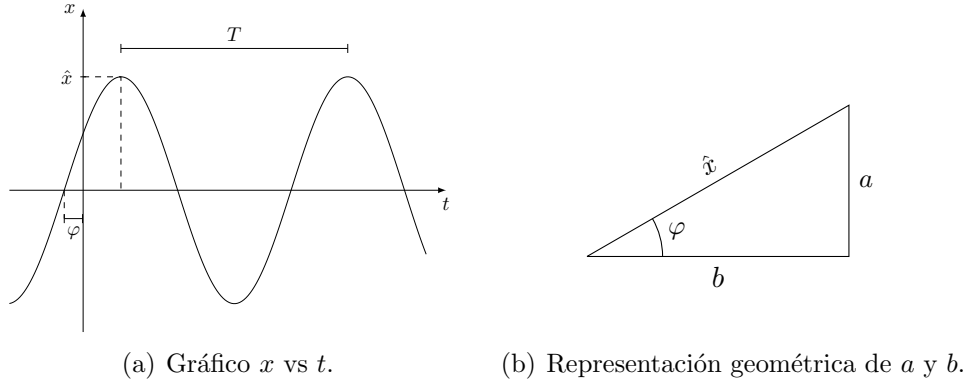


Fig. 6.2: Señal armónica

En muchas aplicaciones la señal de interés es vectorial, es decir $\mathbf{x}(t) \in \mathbb{R}^d$. Claramente $\mathbf{x}(t)$ es periódica de período T si cada componente $x_j(t)$ lo es. Para $d = 2, 3$, se puede interpretar como una curva cerrada en el plano o en el espacio. Todas las soluciones periódicas de sistemas de ecuaciones diferenciales y su representación gráfica en el espacio de fases son ejemplos de esto (ver subsecciones 3.6.1 y 3.6.2).

Ejemplo 6.2. Consideremos la señal en $\mathbf{x}(t) \in \mathbb{R}^2$ dada por $\mathbf{x}(t) = (\sin(\omega t), \sin(\omega t + \varphi))$. Claramente es periódica con período $T = 2\pi/\omega$, la curva que dibuja en el plano depende de φ . En la Figura 6.3 vemos las curvas cerradas correspondientes a $\varphi = 0, \pi/4, \pi/2$. En general vemos que se obtiene una elipse (Figura 6.3(b)), para algunos valores particulares de φ sus semiejes son iguales quedando determinada una circunferencia (Figura 6.3(c)). También existen valores para los cuales uno de sus semiejes se anula, degenerando la elipse en un segmento como se observa en la Figura 6.3(a).

En óptica, el campo eléctrico \mathbf{E} de una onda electromagnética es un vector perpendicular a la dirección de propagación. En el caso de tratarse de una onda monocromática, las componentes de \mathbf{E} varían en forma senoidal con la misma frecuencia, pero con distintas fases. Esto da lugar a diferentes polarizaciones según los casos ilustrados en la Figura 6.3.

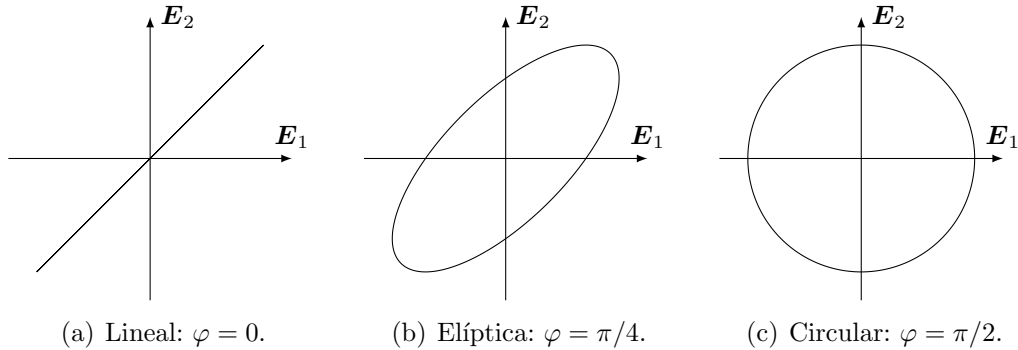


Fig. 6.3: Diferentes polarizaciones.

Curvas de Lissajous En el caso de tener dos componentes que varían en forma senoidal con distinta frecuencia angular $\mathbf{x}(t) = (x_1(t), x_2(t)) = (\sin(\omega_1 t), \cos(\omega_2 t))$. Las curvas del plano que resultan se conocen como curvas de Lissajous. La función vectorial $\mathbf{x}(t)$ es periódica si las frecuencias angulares son conmensurables, es decir, si

su cociente ω_1/ω_2 es racional. En el caso en el que el cociente de frecuencias no sea una fracción racional, la trayectoria será abierta y la curva nunca pasará dos veces por un mismo punto del plano con la misma dirección, dado que es la solución del sistema de segundo orden

$$\begin{aligned}\ddot{x}_1(t) &= \omega_1^2 x_1(t), \\ \ddot{x}_2(t) &= \omega_2^2 x_2(t),\end{aligned}$$

con condiciones iniciales $\mathbf{x}(0) = (0 \ 1)$, $\dot{\mathbf{x}}(0) = (\omega_1 \ 0)$. Supongamos que $\omega_1/\omega_2 = p/q$, si m el mínimo común múltiplo entre p y q , entonces $\mathbf{x}(t)$ es T -periódica, con $T = \frac{m}{q}T_1 = \frac{m}{p}T_2$.

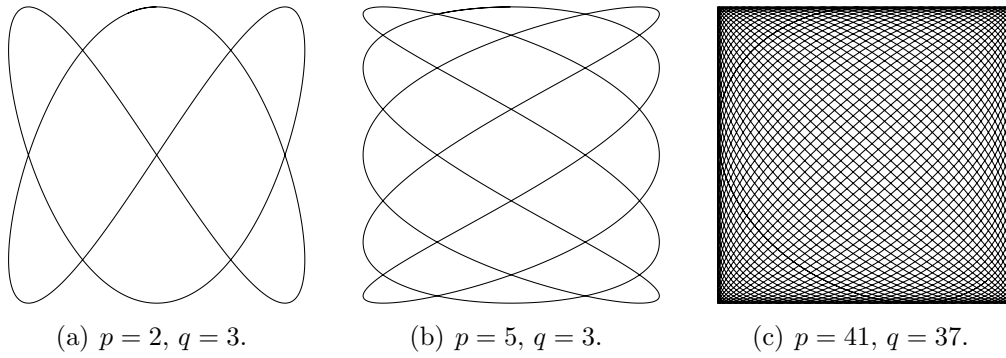


Fig. 6.4: Figura de Lissajous para distintos valores de p y q .

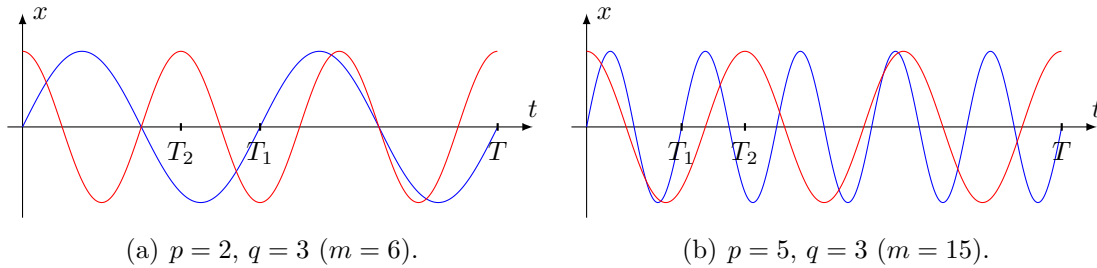


Fig. 6.5: Períodos para distintos valores de p y q .

6.1.2. Períodos y período mínimo. Como ya vimos, una señal $x(t)$ es periódica si se repite regularmente cada cierto tiempo T , es decir

$$(6.1) \quad x(t + T) = x(t), \quad t \in \mathbb{R},$$

en ese caso decimos que $x(t)$ es T -periódica. Obviamente, $x(t + 2T) = x(t + T) = x(t)$, y en general $x(t + kT) = x(t)$ para k entero (ver ejercicio 6.1). Es decir que hay más de un tiempo T para el cual se verifica (6.1). En el caso particular que $x(t)$ sea constante, es decir $x(t) = \hat{x}$, esto sucede todo $T \in \mathbb{R}$. Pero fuera de este caso trivial², se puede probar que existe $T > 0$ mínimo, es decir si $x(t + \tau) = x(t)$, $t \in \mathbb{R}$, entonces $\tau = kT$, con k número entero, en particular $|\tau| \geq T$. Decimos que T es el período de $x(t)$.

✎ **Ejercicio 6.1.** Probar que si $x(t)$ es T -periódica, entonces es kT -periódica para k entero.

✎ **Ejercicio 6.2.** Probar que si T_1 y T_2 son períodos de $x(t)$, entonces $k_1T_1 + k_2T_2$, con k_1, k_2 enteros, es un período de $x(t)$.

²Una señal constante casi no contiene información.

Existencia de período mínimo Vamos a discutir la afirmación anterior sobre la existencia de período mínimo. Esto requiere de algunas herramientas matemáticas más avanzadas. Si $x(t)$ es una señal periódica no constante, definimos

$$T = \inf \{ \tau > 0 : x(t + \tau) = x(t), t \in \mathbb{R} \}$$

Por definición $T \geq 0$, probaremos que $T > 0$, en otro caso $x(t)$ es constante. En efecto, si $T = 0$, existe una sucesión de tiempos $\tau_n \rightarrow 0$ para los cuales se verifica $x(t + \tau_n) = x(t)$, entonces

$$x'(t) = \lim_{n \rightarrow \infty} \frac{x(t + \tau_n) - x(t)}{\tau_n} = 0,$$

lo que muestra que $x(t)$ es constante. El resultado vale aunque $x(t)$ no sea derivable, pero su demostración es más complicada y excede a este apunte. La generalización no es solamente un problema académico, dado que nos interesan señales no derivables en algunos puntos, e inclusive no continuas (ver Ejemplos 6.3 y 6.4). Por otro lado, alguna *regularidad* de la señal se tiene que verificar para que el resultado sea cierto. Por ejemplo, para la señal³ definida como

$$x(t) = \begin{cases} 1 & \text{si } t \text{ racional,} \\ 0 & \text{si } t \text{ irracional,} \end{cases}$$

se puede ver que para todo τ racional, se verifica $x(t + \tau) = x(t)$ y claramente $x(t)$ no es constante. Pero en un sentido que no discutiremos aquí, $x(t)$ es equivalente a la señal nula. Podemos asumir entonces que para toda $x(t)$ periódica y no constante, el período T es positivo.

Queremos ver si $\tau > 0$ verifica $x(t + \tau) = x(t)$, entonces $\tau = kT$ con k entero. Como T es el menor de los tiempos que verifican $x(t + \tau) = x(t)$, vemos que $T \leq \tau$. Si k es el mayor entero menor que $\tau/T \geq 1$, es decir $k = \lfloor \tau/T \rfloor$, tenemos $0 \leq \tau - kT < T$. Como τ y T son períodos de $x(t)$, vale $x(t + \tau - kT) = x(t - kT) = x(t)$. Pero $\tau - kT < T$, por lo tanto $\tau - kT = 0$.

6.1.3. Operaciones con señales periódicas. En otros capítulos vimos que las magnitudes físicas que describen un sistema se relacionan entre sí mediante operaciones algebraicas, suma y producto, o por operaciones del cálculo, derivación e integración. Vamos a ver que si las señales son T -periódicas, el resultado de estas operaciones también lo es. En efecto, $x_1(t), x_2(t)$ son dos señales periódicas T -periódicas, entonces $x_1(t + T) = x_1(t)$, $x_2(t + T) = x_2(t)$, por lo tanto $x_1(t + T) + x_2(t + T) = x_1(t) + x_2(t)$; lo mismo vale para el producto. Sin embargo el período mínimo puede cambiar, como vemos en el siguiente ejemplo. Si $x_1(t) = \sin(2\pi t/T) + \sin(4\pi t/T)$ y $x_2(t) = -\sin(2\pi t/T)$, ambas tienen período mínimo T . Pero $x_1(t) + x_2(t) = \sin(4\pi t/T)$ tiene período mínimo es $T/2$. Para la multiplicación los ejemplos son más sencillos aún, si $x_1(t) = x_2(t) = \sin(2\pi t/T)$, entonces

$$x_1(t) \cdot x_2(t) = \sin^2(2\pi t/T) = \frac{1}{2}(1 - \cos(4\pi t/T)),$$

nuevamente el período es $T/2$.

De manera similar, podemos ver que si $x(t)$ es una señal T -periódica, entonces

- $x_1(t) = x(-t)$ es T -periódica,
- $x_2(t) = x(t - \tau)$ es T -periódica para todo $\tau \in \mathbb{R}$,
- $x_3(t) = x(t/\lambda)$ es λT -periódica para todo $\lambda > 0$.⁴

Como se verifica

$$\frac{x(t + T + h) - x(t + T)}{h} = \frac{x(t + h) - x(t)}{h},$$

³Esta función se conoce como función de Dirichlet.

⁴Observemos que tomando $\lambda = T^{-1}$, la nueva señal tiene período 1.

tomando límite $h \rightarrow 0$ obtenemos $x'(t+T) = x'(t)$, lo que prueba que la derivada $x'(t)$ es T -periódica. No vale el resultado equivalente para primitivas de $x(t)$, por ejemplo la primitiva de una señal constante no nula, $x(t) = a \neq 0$ es $X(t) = X_0 + at$, que no es periódica. Sin embargo, la señal $x(t) = \hat{x} \sin(\omega t + \varphi)$ tiene primitiva, $X(t) = X_0 - \hat{x}/\omega \cos(\omega t + \varphi)$, que si es periódica. La condición necesaria y suficiente para que la primitiva de una señal periódica sea también periódica es que su valor medio sea nulo, es decir

$$\int_0^T x(t) dt = 0.$$

En efecto, si $X(t)$ es una primitiva T -periódica, por la regla de Barrow obtenemos

$$\int_0^T x(t) dt = X(T) - X(0) = 0.$$

Recíprocamente, si $x(t)$ tiene media nula, definimos

$$X(t) = \int_0^t x(t') dt',$$

entonces

$$X(t+T) = \int_0^{t+T} x(t') dt' = \int_0^T x(t') dt' + \int_T^{t+T} x(t') dt'.$$

La primera integral del lado derecho es la media de $x(t)$, y por lo tanto es nula. La segunda integral, verifica

$$\int_T^{t+T} x(t') dt' = \int_0^t x(t') dt'$$

como se ve en la Figura 6.6, lo que implica $X(t+T) = X(t)$.

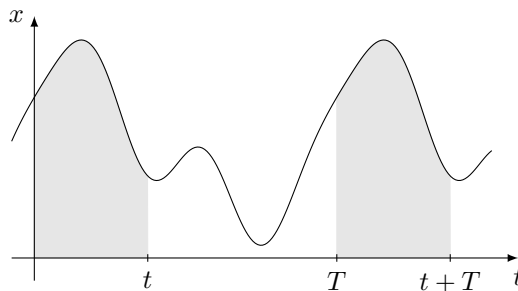



Fig. 6.6

 **Ejercicio 6.3.** Probar que si $x(t)$ es T -periódica, entonces para todo $\tau \in \mathbb{R}$

$$\int_0^T x(t) dt = \int_\tau^{\tau+T} x(t) dt,$$

como lo muestra la Figura 6.7.

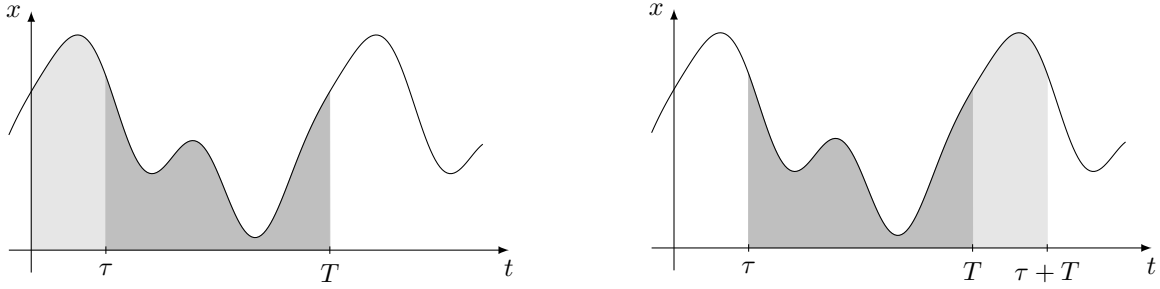


Fig. 6.7: Cálculo de la media en los intervalos $[0, T]$ y $[\tau, \tau + T]$.

6.1.4. Polinomios trigonométricos. Vamos a considerar de ahora en más, señales de período 1. Como un ejemplo importante de señales periódicas tenemos los polinomios trigonométricos, cuya expresión general está dada por

$$(6.2) \quad x(t) = a_0 + \sum_{k=1}^n (a_k \cos(2\pi kt/T) + b_k \sin(2\pi kt/T)).$$

Tiene sentido llamarlos polinomios dado que si $x_1(t)$ y $x_2(t)$ son polinomios trigonométricos, entonces $x_1(t) + x_2(t)$ y $x_1(t) \cdot x_2(t)$ también lo son. Es fácil ver que la suma es un polinomio trigonométrico. Para el producto usamos las identidades del apéndice A.1.1:


$$\begin{aligned} \cos(2\pi jt/T) \cos(2\pi kt/T) &= \frac{1}{2} (\cos(2\pi(k-j)t/T) + \cos(2\pi(k+j)t/T)) \\ \sin(2\pi jt/T) \sin(2\pi kt/T) &= \frac{1}{2} (\cos(2\pi(k-j)t/T) - \cos(2\pi(k+j)t/T)) \\ \cos(2\pi jt/T) \sin(2\pi kt/T) &= \frac{1}{2} (\sin(2\pi(k+j)t/T) - \sin(2\pi(k-j)t/T)) \end{aligned}$$

Usando que $e^{i2k\pi t/T} = \cos(2k\pi t/T) + i \sin(2k\pi t/T)$, un polinomio trigonométrico $x(t)$ se puede reescribir de la forma

$$(6.3) \quad x(t) = \sum_{k=-n}^n c_k e^{i2k\pi t/T},$$

donde $a_0 = c_0$, $a_k = c_k + c_{-k}$ y $b_k = ic_k - ic_{-k}$, o equivalentemente

$$c_k = \begin{cases} \frac{a_k - ib_k}{2} & , \text{ si } k > 0, \\ \frac{a_{-k} + ib_{-k}}{2} & , \text{ si } k < 0. \end{cases}$$

 **Ejercicio 6.4.** Probar que $a_0, \dots, a_n, b_1, \dots, b_n$ son reales si y sólo si $c_{-k} = \bar{c}_k$.

6.1.5. Cálculo de los coeficientes. Supongamos que sabemos que la señal $x(t)$ periódica se representa por un polinomio trigonométrico, pero sus coeficientes son desconocidos. Nos planteamos como hallar los coeficientes a partir de los valores de $x(t)$. Vamos a proponer dos posibles respuestas. La primera forma consiste en evaluar $x(t)$ en número suficientemente grande de tiempos t_1, \dots, t_N y resolver un sistema de ecuaciones lineales en los coeficientes c_{-n}, \dots, c_n :

$$\begin{pmatrix} e^{-i2n\pi t_1/T} & \dots & e^{-i2\pi t_1/T} & 1 & e^{i2\pi t_1/T} & \dots & e^{i2n\pi t_1/T} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ e^{-i2n\pi t_N/T} & \dots & e^{-i2\pi t_N/T} & 1 & e^{i2\pi t_N/T} & \dots & e^{i2n\pi t_N/T} \end{pmatrix} \begin{pmatrix} c_{-n} \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} x(t_1) \\ \vdots \\ x(t_N) \end{pmatrix}$$

Para asegurarnos que el sistema tenga solución única tomamos $N = 2n + 1$, la matriz resultante debe ser inversible. Este problema será estudiado con detalle en la Sección 6.2.

La segunda forma de obtener los coeficientes se basa en expresiones integrales. Estas relaciones nos permiten extender estas ideas a señales arbitrarias. Observando que se verifican las igualdades

$$\begin{aligned}\frac{1}{T} \int_0^T \cos(2k\pi t/T) dt &= \frac{1}{T} \int_0^T \sin(2k\pi t/T) dt = 0, \\ \frac{1}{T} \int_0^T \cos(2k\pi t/T) \sin(2m\pi t/T) dt &= 0, \\ \frac{1}{T} \int_0^T \cos(2k\pi t/T) \cos(2m\pi t/T) dt &= \frac{1}{2} \delta_{k,m}, \\ \frac{1}{T} \int_0^T \sin(2k\pi t/T) \sin(2m\pi t/T) dt &= \frac{1}{2} \delta_{k,m},\end{aligned}$$

vemos que

$$(6.4a) \quad a_0 = \frac{1}{T} \int_0^T x(t) dt,$$

$$(6.4b) \quad a_k = \frac{2}{T} \int_0^T x(t) \cos(2k\pi t/T) dt,$$

$$(6.4c) \quad b_k = \frac{2}{T} \int_0^T x(t) \sin(2k\pi t/T) dt.$$

Podemos obtener fórmulas equivalentes para los coeficientes en la forma exponencial de $x(t)$, si $k \geq 1$ tenemos

$$(6.5) \quad \begin{aligned}c_k &= \frac{a_k - ib_k}{2} = \frac{1}{T} \int_0^T x(t) (\cos(2k\pi t/T) - i \sin(2k\pi t/T)) dt \\ &= \frac{1}{T} \int_0^T x(t) e^{-i2k\pi t/T} dt.\end{aligned}$$

Es posible ver que esta última expresión, vale también para $k \leq 0$. Podemos observar que si $x(t)$ toma valores reales si y solo si los coeficientes a_0, a_k, b_k son reales, lo que es equivalente a $c_{-k} = \bar{c}_k$. Si $x(t)$ es par, entonces $b_k = 0$ y

$$a_k = \frac{2}{T} \int_0^T x(t) \cos(2k\pi t/T) dt = \frac{4}{T} \int_0^{T/2} x(t) \cos(2k\pi t/T) dt.$$

De la misma forma, si $x(t)$ es impar, entonces $a_k = 0$ y

$$b_k = \frac{2}{T} \int_0^T x(t) \sin(2k\pi t/T) dt = \frac{4}{T} \int_0^{T/2} x(t) \sin(2k\pi t/T) dt.$$

6.1.6. Aproximación de Fourier. Hasta ahora hemos considerado polinomios trigonométricos, pero podemos extender estas ideas al caso general. Para una señal arbitraria $x(t)$, definimos los coeficientes de Fourier por (6.4) (o por (6.5)) Claramente siguen valiendo las relaciones $c_k = (a_k - ib_k)/2$, $c_{-k} = (a_k + ib_k)/2$ para $k \geq 1$. Si $x(t)$ toma valores reales, entonces los coeficientes a_0, a_k, b_k son reales y $c_{-k} = \bar{c}_k$.

Para todo $n \geq 1$, definimos la aproximación de Fourier

$$(6.6) \quad x_n(t) = \sum_{k=-n}^n c_k e^{i2k\pi t/T} = a_0 + \sum_{k=1}^n a_k \cos(2k\pi t/T) + b_k \sin(2k\pi t/T).$$

La teoría de Fourier afirma que los polinomios trigonométricos $x_n(t)$ aproximan a cualquier señal $x(t)$ con un grado de exactitud arbitrario, es decir si $x(t)$ es una señal T -periódica, entonces los polinomios trigonométricos $x_n(t)$ definidos por (6.6) verifican $x_n(t) \rightarrow x(t)$ cuando $n \rightarrow \infty$. En que sentido se da esta convergencia no es un problema sencillo y debe ser analizado con cuidado.

La discusión sobre que funciones se pueden aproximar por polinomios trigonométricos comenzó a mediados del siglo XVIII y se extendió por mucho más de un siglo, participaron entre otros Leonhard Euler, Jean le Rond D'Alembert, Daniel Bernoulli, Joseph-Louis Lagrange, Jean-Baptiste Joseph Fourier (ver [12]). Este problema puso en crisis la noción misma de función y la de integración, llevando a una revisión profunda de estos conceptos durante la primera mitad del siglo XX.⁵ El argumento de quienes pensaban que no toda función se podía expresar como límite de polinomios trigonométricos (Euler, D'Alembert, Lagrange), se basaba en la imposibilidad de representar una función partida⁶ por una única suma de senos y cosenos. El siguiente ejemplo refuta en forma gráfica esta objeción.

Ejemplo 6.3. Consideremos $x(t)$ la señal triangular de la Figura 6.8, correspondiente a la función periódica, con período 1, dada por $x(t) = 1 - 4|t|$ si $t \in [-1/2, 1/2]$. Siendo $x(t)$ par, los coeficientes b_k , correspondientes a $\sin(2k\pi t)$, son nulos. Los coeficientes a_k valen

$$a_k = 2 \int_{-1/2}^{1/2} (1 - 4|t|) \cos(2k\pi t) dt = 4 \int_0^{1/2} (1 - 4t) \cos(2k\pi t) dt = \frac{4(1 - (-1)^k)}{\pi^2 k^2}.$$

Si definimos los polinomios trigonométricos

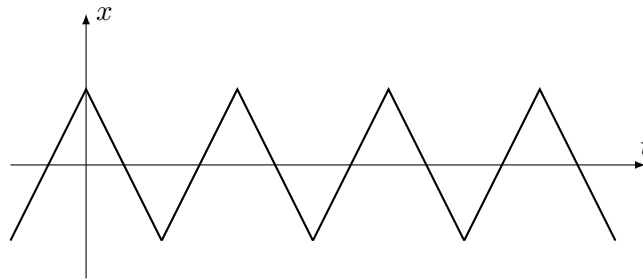


Fig. 6.8: Señal triangular.

$$x_{2n+1}(t) = \frac{8}{\pi^2} \sum_{k=0}^n \frac{1}{(2k+1)^2} \cos(2(2k+1)\pi t),$$

vemos en la Figura 6.9 que a medida que aumenta el número de términos, la aproximación es cada vez mejor. Consideremos por ejemplo $t = 0$, que es un punto donde cambia la definición de $x(t)$ (un punto de discontinuidad según Euler), en la Tabla 6.1 mostramos el valor de $x_{2n+1}(0)$ para $n = 10, 100, 1000, 10000$. Claramente vemos que $x_{2n+1}(0) \rightarrow x(0) = 1$. En la Tabla 6.1 mostramos la suma para distinto valores de n .

n	10	100	1000	10000
$x_{2n+1}(0)$	0.981 591	0.997 994	0.999 798	0.999 980

Tabla 6.1

⁵Casi todos los conceptos del análisis matemático tuvieron que ser reformulados.

⁶Esto era lo que Euler entendía por una función discontinua.

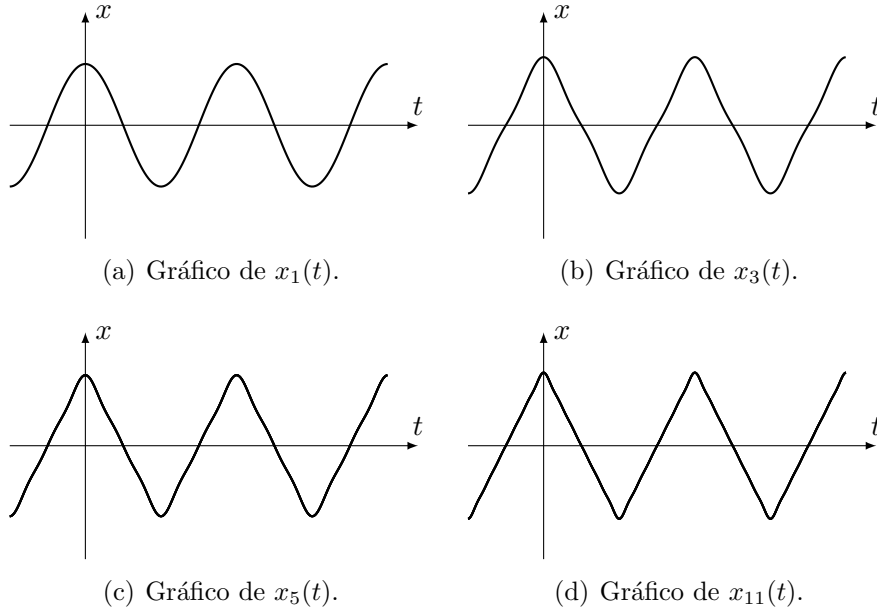


Fig. 6.9: Aproximaciones por polinomios trigonométricos de la señal triangular.

En particular, vemos que

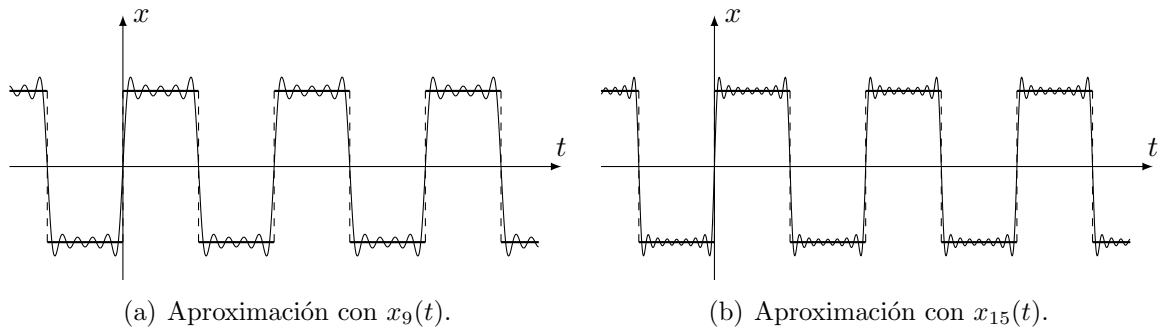
$$\lim_{n \rightarrow \infty} \sum_{k=0}^n \frac{1}{(2k+1)^2} = \frac{\pi^2}{8} \lim_{n \rightarrow \infty} x_{2n+1}(0) = \frac{\pi^2}{8}.$$

Observamos en el ejemplo anterior, que los polinomios trigonométricos aproximan en forma uniforme a $x(t)$, es decir

$$\max_{t \in \mathbb{R}} |x(t) - x_n(t)| = \max_{-1 \leq t \leq 1} |x(t) - x_n(t)| \rightarrow 0, \text{ si } n \rightarrow \infty.$$

Dado que los polinomios trigonométricos son continuos, no podrían converger uniformemente a una señal discontinua. En el siguiente ejemplo vemos en que sentido podemos plantear la convergencia.

Ejemplo 6.4. Estudiamos ahora $x(t)$ la señal cuadrada de la Figura 6.10, 1–periódica, definida por $x(t) = 1$ si $t \in [0, 1/2)$ y $x(t) = -1$ si $t \in [-1/2, 0)$. Siendo impar, solo calcularemos los

Fig. 6.10: Señal cuadrada $x(t)$ y la aproximación $x_n(t)$.

coeficientes b_k , integrando obtenemos

$$b_k = 2 \int_{-1/2}^{1/2} x(t) \sin(2k\pi t) dt = 4 \int_0^{1/2} \sin(2k\pi t) dt = \frac{2(1 - (-1)^k)}{\pi k},$$

vemos que para k par, $b_k = 0$ y $b_k = 4(\pi k)^{-1}$ si k es impar. En este caso, los polinomios trigonométricos que aproximan a $x(t)$ son

$$x_{2n+1}(t) = \frac{4}{\pi} \sum_{k=0}^n \frac{1}{2k+1} \sin(2(2k+1)\pi t),$$

En la Figura 6.10(a), superpusimos sobre la señal cuadrada el polinomio trigonométrico $x_9(t)$, y en la Figura 6.10(b), $x_{15}(t)$. Vemos que en las discontinuidades, $t = 0, \pm 1/2, \pm 1, \dots$, la aproximación toma el valor 0, que es el punto medio del salto. Esto no es particular del ejemplo, puede probarse que esto sucede para cualquier señal con saltos. Observamos que cerca de la discontinuidad aparecen oscilaciones, que al aumentar n no disminuyen en amplitud, solo se acercan al punto de discontinuidad. A este comportamiento se lo conoce como fenómeno de Gibbs. En la Figura 6.11 vemos una ampliación de las oscilaciones en el primer milisegundo,

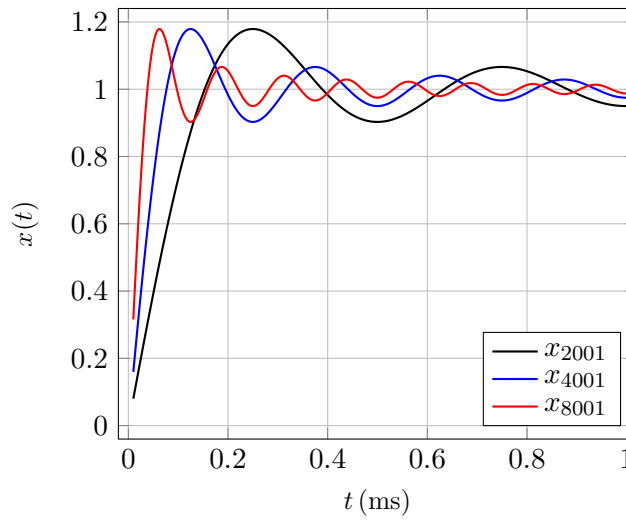


Fig. 6.11: Fenómeno de Gibbs.

$0 \leq t \leq 10^{-3}$, para $n = 1000, 2000, 4000$. La altura del máximo no se modifica, de hecho se puede probar que corresponde aproximadamente al 9% del valor del salto ($2 = 1 - (-1)$ en este caso) para cualquier discontinuidad. Sin embargo, las oscilaciones se van concentrando cada vez más cerca del tiempo del salto, a medida que n crece. Esto nos dice que $x_n(t)$ converge a $x(t)$ para cualquier tiempo fuera de las discontinuidades.

6.1.7. Potencia de una señal y convergencia en media cuadrática. En un circuito eléctrico, la energía disipada en un resistor es proporcional al cuadrado de $v(t)$, donde $v(t)$ es la diferencia de potencial entre los bornes del resistor. Concretamente, por la ley de Joule, la potencia es $P(t) = v(t)i(t) = v^2(t)/R$, por lo tanto la energía disipada por período es

$$E = \frac{1}{R} \int_0^T |v(t)|^2 dt,$$

la potencia media está dada por $\bar{P} = E/T$. Definimos la diferencia de potencial efectiva como $V_{\text{ef}} = (\bar{P}R)^{1/2}$, es decir $\bar{P} = V_{\text{ef}}^2/R$ que corresponde a la potencia disipada en el resistor para una diferencia de potencial continua $V = V_{\text{ef}}$. Despejando obtenemos

$$V_{\text{ef}} = \left(\frac{1}{T} \int_0^T |v(t)|^2 dt \right)^{1/2}.$$

Para el caso armónico, $v(t) = \hat{v} \sin(\omega t + \varphi)$, se verifica

$$V_{\text{ef}} = \hat{v} \left(\frac{1}{T} \int_0^T |\sin(\omega t + \varphi)|^2 dt \right)^{1/2} = \frac{\hat{v}}{\sqrt{2}}.$$

En general, si $v(t)$ es un polinomio trigonométrico T -periódico,

$$v(t) = a_0 + \sum_{k=1}^n a_k \cos(2k\pi t/T) + b_k \sin(2k\pi t/T),$$

entonces, usando las igualdades A.1.1 obtenemos

$$\bar{P}R = \frac{1}{T} \int_0^T v^2(t) dt = a_0^2 + \frac{1}{2} \sum_{k=1}^n (a_k^2 + b_k^2) = \sum_{k=-n}^n |c_k|^2.$$

Estudiemos el problema general, si $x(t)$ es una señal T -periódica y $x_n(t)$ el polinomio trigonométrico aproximante

$$x_n(t) = a_0 + \sum_{k=1}^n a_k \cos(2k\pi t/T) + b_k \sin(2k\pi t/T),$$

Por simplicidad en el análisis vamos a suponer que $x(t)$ toma valores reales, fácilmente se puede extender al caso general. Vamos a medir la bondad de la aproximación como la diferencia en media cuadrática:

$$\frac{1}{T} \int_0^T (x(t) - x_n(t))^2 dt = \frac{1}{T} \int_0^T x^2(t) dt - \frac{2}{T} \int_0^T x(t) x_n(t) dt + \frac{1}{T} \int_0^T x_n^2(t) dt.$$

Ya vimos como calcular la última integral a partir de los coeficientes, la segunda integral vale

$$\begin{aligned} \frac{1}{T} \int_0^T x(t) x_n(t) dt &= \frac{a_0}{T} \int_0^T x(t) dt + \sum_{k=1}^n \frac{a_k}{T} \int_0^T x(t) \cos(2k\pi t/T) dt \\ &+ \frac{b_k}{T} \int_0^T x(t) \sin(2k\pi t/T) dt = a_0^2 + \frac{1}{2} \sum_{k=1}^n a_k^2 + b_k^2 = \frac{1}{T} \int_0^T x_n^2(t) dt, \end{aligned}$$

de donde obtenemos

$$0 \leq \frac{1}{T} \int_0^T (x(t) - x_n(t))^2 dt = \frac{1}{T} \int_0^T x^2(t) dt - \frac{1}{T} \int_0^T x_n^2(t) dt,$$

que se puede escribir como la denominada desigualdad de Bessel:

$$\frac{1}{T} \int_0^T x_n^2(t) dt = a_0^2 + \frac{1}{2} \sum_{k=1}^n a_k^2 + b_k^2 \leq \frac{1}{T} \int_0^T x^2(t) dt.$$

Además, $x_n(t)$ converge en media cuadrática a $x(t)$ si y sólo si

$$\lim_{n \rightarrow \infty} \frac{1}{T} \int_0^T x_n^2(t) dt = \frac{1}{T} \int_0^T x^2(t) dt,$$

en ese caso, obtenemos la identidad de Parseval

$$\lim_{n \rightarrow \infty} a_0^2 + \frac{1}{2} \sum_{k=1}^n a_k^2 + b_k^2 = \frac{1}{T} \int_0^T x^2(t) dt.$$

Por ejemplo, para la señal cuadrada tenemos

$$\lim_{n \rightarrow \infty} \frac{1}{T} \int_0^T x_n^2(t) dt = \lim_{n \rightarrow \infty} \frac{1}{2} \sum_{k=1}^n b_k^2 = \lim_{n \rightarrow \infty} \frac{8}{\pi^2} \sum_{k=0}^n \frac{1}{(2k+1)^2} = 1,$$

por lo visto en el Ejemplo 6.3, lo que muestra la convergencia en $x_n(t)$ en media cuadrática a $x(t)$. Se puede probar que bajo hipótesis muy generales sobre la señal, los polinomios trigonométricos obtenidos mediante (6.4) (o (6.5)), convergen en media cuadrática a $x(t)$.⁷

Si $x_T(t)$ es la señal triangular considerada en el Ejemplo 6.3 y $x_C(t)$ la señal cuadrada del Ejemplo 6.4, vemos que $\dot{x}_T(t) = -4x_C(t)$. Analicemos que sucede con las aproximaciones de Fourier

$$\dot{x}_{T,2n+1}(t) = -\frac{16}{\pi} \sum_{k=0}^n \frac{1}{2k+1} \sin(2(2k+1)\pi t) = -4x_{C,2n+1}(t).$$

Esto es general para cualquier señal $x(t)$. Por ejemplo, la primitiva de $x_T(t)$ con media nula, está dada por la función 1-periódica que verifica

$$X_T(t) = \begin{cases} t + 2t^2 & \text{si } -1/2 \leq t < 0, \\ t - 2t^2 & \text{si } 0 \leq t < 1/2, \end{cases}$$

Por otro lado, la primitiva de $x_{T,2n+1}(t)$ vale

$$X_{T,2n+1}(t) = \frac{4}{\pi^3} \sum_{k=0}^n \frac{1}{(2k+1)^3} \sin(2(2k+1)\pi t),$$

que corresponde a la aproximación de Fourier de $X_T(t)$.

6.1.8. Soluciones periódicas de ecuaciones diferenciales. Como vimos en el Capítulo 3, existen problema de ecuaciones diferenciales que admiten soluciones periódicas. En el caso de sistemas lineales con coeficientes constantes, podemos hallar estas soluciones en forma analítica simplemente mediante el método de coeficientes indeterminados que se propone a continuación. Supongamos que queremos estudiar el sistema $\dot{\mathbf{x}}(t) = \mathbf{C} \mathbf{x}(t) + \mathbf{f}(t)$, donde \mathbf{f} es una función vectorial periódica. La solución general se puede descomponer como $\mathbf{x}(t) = \mathbf{x}_h(t) + \mathbf{x}_p(t)$, donde \mathbf{x}_h es cualquier solución del sistema homogéneo y \mathbf{x}_p es una solución particular. Vamos a buscar una solución particular periódica para el caso donde no homogeneidad es armónica:

$$\mathbf{f}(t) = \hat{\mathbf{f}} \cos(\omega t + \varphi) = \mathbf{f}_a \cos(\omega t) + \mathbf{f}_b \sin(\omega t),$$

podemos proponer $\mathbf{x}(t) = \mathbf{x}_a \cos(\omega t) + \mathbf{x}_b \sin(\omega t)$ obteniendo

$$-\omega \mathbf{x}_a \sin(\omega t) + \omega \mathbf{x}_b \cos(\omega t) = \mathbf{C} \mathbf{x}_a \cos(\omega t) + \mathbf{C} \mathbf{x}_b \sin(\omega t) + \mathbf{f}_a \cos(\omega t) + \mathbf{f}_b \sin(\omega t),$$

separando los términos que corresponden a $\cos(\omega t)$ y a $\sin(\omega t)$

$$\begin{aligned} \omega \mathbf{x}_b - \mathbf{C} \mathbf{x}_a &= \mathbf{f}_a, \\ -\omega \mathbf{x}_a - \mathbf{C} \mathbf{x}_b &= \mathbf{f}_b, \end{aligned}$$

despejando \mathbf{x}_b de la primera ecuación y reemplazando en la segunda obtenemos

$$\begin{aligned} \mathbf{x}_a &= -(\mathbf{C}^2 + \omega^2 \mathbf{I})^{-1} (\mathbf{C} \mathbf{f}_a + \omega \mathbf{f}_b), \\ \mathbf{x}_b &= -(\mathbf{C}^2 + \omega^2 \mathbf{I})^{-1} (\mathbf{C} \mathbf{f}_b - \omega \mathbf{f}_a). \end{aligned}$$

⁷Sin embargo la convergencia puntual no vale, aún para señales $x(t)$ continuas.

Vemos que el problema tiene solución si $\pm i\omega$ no es autovalor de C . En otro caso, decimos que $\nu = \omega/(2\pi)$ es una frecuencia de resonancia. Si definimos la matriz $H(\omega)$

$$H(\omega) = \begin{pmatrix} -(C^2 + \omega^2 I)^{-1} C & -\omega(C^2 + \omega^2 I)^{-1} I \\ \omega(C^2 + \omega^2 I)^{-1} I & -(C^2 + \omega^2 I)^{-1} C \end{pmatrix},$$

entonces

$$(6.7) \quad \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} = H(\omega) \begin{pmatrix} \mathbf{f}_a \\ \mathbf{f}_b \end{pmatrix}.$$

Matriz de transferencia Si asumimos $\mathbf{x}, \mathbf{f}, C$ reales, en lugar de considerar el sistema (6.7) podemos tomar los vectores complejos $\mathcal{X} = \mathbf{x}_a - i\mathbf{x}_b$, $\mathcal{F} = \mathbf{f}_a - i\mathbf{f}_b$ y la matriz $\mathcal{H}(\omega) = -(C^2 + \omega^2 I)^{-1} (C + i\omega I)$, para plantear el sistema complejo equivalente $\mathcal{X} = \mathcal{H}(\omega) \mathcal{F}$.

En el caso general, si \mathbf{f} es una función vectorial periódica más general

$$\mathbf{f}(t) = \sum_{k=1}^N \mathbf{f}_{a,k} \cos(k\omega t) + \mathbf{f}_{b,k} \sin(k\omega t),$$

como el sistema es lineal, por el principio de superposición la solución periódica se escribe de la forma $\mathbf{x}_{\text{per}}(t) = \sum_{k=1}^N \mathbf{x}_{a,k} \cos(k\omega t) + \mathbf{x}_{b,k} \sin(k\omega t)$, donde

$$(6.8a) \quad \mathbf{x}_{a,k} = -(C^2 + k^2\omega^2 I)^{-1} (C\mathbf{f}_{a,k} + k\omega \mathbf{f}_{b,k}),$$

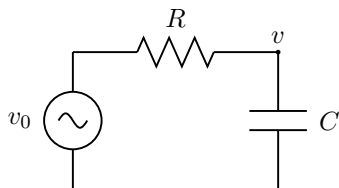
$$(6.8b) \quad \mathbf{x}_{b,k} = -(C^2 + k^2\omega^2 I)^{-1} (C\mathbf{f}_{b,k} - k\omega \mathbf{f}_{a,k}).$$

Los sistemas reales suelen disipar energía generalmente en forma de calor (rozamiento, efecto joule, etc.), que se refleja en el comportamiento del sistema libre, es decir $\mathbf{f} = 0$, donde se observa que las soluciones deben converger a 0. Esto se relaciona con los autovalores de la matriz C , los cuales tienen parte real negativa. Vemos entonces que no existen frecuencias de resonancia, lo que permite construir la solución particular periódica $\mathbf{x}_{\text{per}}(t)$ dada por los coeficientes (6.8). La solución general se escribe entonces como $\mathbf{x}(t) = \exp(tC) \mathbf{x}_0 + \mathbf{x}_{\text{per}}(t)$, pero por lo que asumimos sobre el sistema C , $\exp(tC) \mathbf{x}_0 \rightarrow \mathbf{0}$ cuando $t \rightarrow \infty$ en forma exponencial. Esto muestra que asintóticamente $\mathbf{x}(t)$ se comporta como $\mathbf{x}_{\text{per}}(t)$, más precisamente existe $\alpha > 0$ que verifica

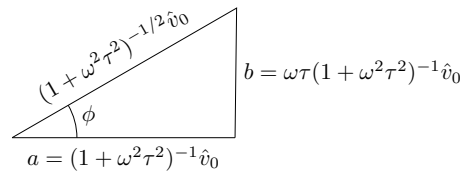
$$\mathbf{x}(t) = \sum_{k=1}^N (\mathbf{x}_{a,k} \cos(k\omega t) + \mathbf{x}_{b,k} \sin(k\omega t)) + o(e^{-\alpha t}),$$

para $t \rightarrow \infty$. Esto vale para $\alpha > 0$ que verifique $\text{Re}(\lambda) < -\alpha < 0$ para todo λ autovalor de C .

Ejemplo 6.5. Consideremos el circuito de la Figura 6.12(a), por lo visto en el Capítulo 3, el potencial $v(t)$ satisface al ecuación diferencial $\tau \dot{v}(t) + v(t) = v_0(t)$, donde $\tau = RC$.



(a) Circuito R-C.



(b) Representación geométrica.

Fig. 6.12

Para $v_0(t) = \hat{v}_0 \cos(\omega t)$ ($\hat{v}_0 > 0$), buscamos una solución $v(t) = a \cos(\omega t) + b \sin(\omega t)$. Reemplazando en la ecuación, obtenemos

$$\hat{v}_0 \cos(\omega t) = (a + \omega \tau b) \cos(\omega t) + (-\omega \tau a + b) \sin(\omega t),$$

por lo tanto $a + \omega \tau b = \hat{v}_0$, $-\omega \tau a + b = 0$. Resulta $a = (1 + \omega^2 \tau^2)^{-1} \hat{v}_0$ y $b = \omega \tau (1 + \omega^2 \tau^2)^{-1} \hat{v}_0$. Si definimos

$$\begin{aligned}\hat{v} &= (a^2 + b^2)^{1/2} \hat{v}_0 = (1 + \omega^2 \tau^2)^{-1/2} \hat{v}_0, \\ \phi &= -\arctan(\omega \tau),\end{aligned}$$

(ver Figura 6.12(b)), podemos escribir $a = \hat{v} \cos(\phi)$, $b = -\hat{v} \sin(\phi)$ y $v(t) = \hat{v} \cos(\omega t + \phi)$. Si $v_0(t) = \hat{v}_{0,1} \cos(\omega_1 t) + \cdots + \hat{v}_{0,k} \cos(\omega_k t)$, superponiendo las soluciones tenemos

$$v(t) = \hat{v}_{0,1} (1 + \omega_1^2 \tau^2)^{-1/2} \cos(\omega_1 t + \phi_1) + \cdots + \hat{v}_{0,k} (1 + \omega_k^2 \tau^2)^{-1/2} \cos(\omega_k t + \phi_k),$$

donde $\phi_j = \phi(\omega_j) = -\arctan(\omega_j \tau)$. Si definimos $H(\omega) = (1 + \omega^2 \tau^2)^{-1/2}$, tenemos

$$v(t) = \hat{v}_{0,1} H(\omega_1) \cos(\omega_1 t + \phi(\omega_1)) + \cdots + \hat{v}_{0,k} H(\omega_k) \cos(\omega_k t + \phi(\omega_k)).$$

Si escribimos $v_0(t) = \frac{1}{2} \hat{v}_0 e^{i\omega t} + \frac{1}{2} \hat{v}_0 e^{-i\omega t}$, la solución $v(t)$ se puede expresar como

$$v(t) = \frac{1}{2} \hat{v}_0 H(\omega) e^{i\phi(\omega)} e^{i\omega t} + \frac{1}{2} \hat{v}_0 H(-\omega) e^{-i\phi(\omega)} e^{-i\omega t} = \frac{1}{2} \hat{v}_0 H(\omega) e^{i\omega t} + \frac{1}{2} \hat{v}_0 H(-\omega) e^{-i\omega t},$$

donde $\mathcal{H}(\omega)$ es la función (con valores complejos) dada por $\mathcal{H}(\omega) = H(\omega) e^{i\phi(\omega)}$ ⁸.

En particular si $v_0(t)$ está dada por $v_0(t) = \sum_{k=-n}^n \hat{v}_{0,k} e^{i2k\pi t/T}$, la solución $v(t)$ es el polinomio trigonométrico dado por

$$(6.9) \quad v(t) = \sum_{k=-n}^n \hat{v}_{0,k} \mathcal{H}(2k\pi/T) e^{i2k\pi t/T}.$$

6.1.9. Convolución. Motivados por la igualdad (6.9) de la sección anterior, vamos a definir una nueva operación sobre las señales periódicas, llamado producto de convolución: dadas dos señales, $x_1(t), x_2(t)$, definidas por los polinomios trigonométricos

$$x_1(t) = \sum_{j=-n}^n c_{1,k} e^{i2j\pi t/T}, \quad x_2(t) = \sum_{k=-n}^n c_{2,k} e^{i2k\pi t/T},$$

definimos el producto de convolución, $x = x_1 * x_2$, como el polinomio trigonométrico

$$x(t) = \sum_{k=-n}^n c_{1,k} c_{2,k} e^{i2k\pi t/T}.$$

Volviendo a (6.9), si $v_0(t) = \sum_{k=-n}^n \hat{v}_{0,k} e^{i2k\pi t/T}$ y definimos $h(t) = \sum_{k=-n}^n H(2k\pi/T) e^{i2k\pi t/T}$

$$v(t) = (h * v_0)(t) = \sum_{k=-n}^n H(2k\pi/T) \hat{v}_{0,k} e^{i2k\pi t/T}.$$

De la definición de producto de convolución, claramente valen:

$$\blacksquare \quad x_1 * x_2 = x_2 * x_1,$$

⁸Se puede ver que $\mathcal{H}(\omega) = (1 + i\omega\tau)^{-1}$.

- $x_1 * (x_2 + x_3) = x_1 * x_2 + x_1 * x_3,$
- $(x_1 * x_2) * x_3 = x_1 * (x_2 * x_3).$

Vamos a probar que se puede obtener $x(t) = (x_1 * x_2)(t)$ mediante la fórmula

$$(6.10) \quad x(t) = \frac{1}{T} \int_0^T x_1(t') x_2(t - t') dt'.$$

Usando las expresiones de $x_1(t)$ y $x_2(t)$, obtenemos

$$x(t) = \frac{1}{T} \int_0^T \sum_{j=-n}^n \sum_{k=-n}^n c_{1,j} c_{2,k} e^{i2j\pi t'/T} e^{i2k\pi(t-t')/T} dt',$$

intercambiando la integral con las sumatorias, vemos que

$$x(t) = \sum_{j=-n}^n \sum_{k=-n}^n c_{1,j} c_{2,k} e^{i2k\pi t/T} \frac{1}{T} \int_0^T e^{i2(j-k)\pi t'/T} dt'$$

Como

$$\frac{1}{T} \int_0^T e^{i2(j-k)\pi t'/T} dt' = \delta_{j,k} = \begin{cases} 1 & \text{si } j = k, \\ 0 & \text{si } j \neq k, \end{cases}$$

concluimos que

$$x(t) = \sum_{j=-n}^n \sum_{k=-n}^n c_{1,j} c_{2,k} e^{i2k\pi t/T} \delta_{j,k} = \sum_{k=-n}^n c_{1,k} c_{2,k} e^{i2k\pi t/T}.$$

📌 **Ejercicio 6.5.** Probar que si los coeficientes de $x(t) = (x_1 * x_2)(t)$ en forma trigonométrica verifican $a_k = (a_{1,k}a_{2,k} - b_{1,k}b_{2,k})/2$ y $b_k = (a_{1,k}b_{2,k} + b_{1,k}a_{2,k})/2$.

Es posible mostrar que si $x(t) = (x_1 * x_2)(t)$, definida por (6.10), entonces $c_k = c_{1,k} c_{2,k}$ para señales en general (lo habíamos probado para polinomios trigonométricos).

Nota: En efecto,

$$\begin{aligned} c_k &= \frac{1}{T} \int_0^T x(t) e^{-i2k\pi t} dt = \frac{1}{T} \int_0^T \frac{1}{T} \int_0^T x_1(t') x_2(t - t') dt' e^{-i2k\pi t} dt \\ &= \frac{1}{T^2} \int_0^T \int_0^T x_1(t') x_2(t - t') e^{-i2k\pi t'} e^{-i2k\pi(t-t')} dt' dt \\ &= \frac{1}{T^2} \int_0^T \int_{-t'}^{-t'+T} x_1(t') x_2(t) e^{-i2k\pi t'} e^{-i2k\pi t} dt dt' \\ &= \frac{1}{T^2} \int_0^T \int_0^T x_1(t') x_2(t) e^{-i2k\pi t'} e^{-i2k\pi t} dt dt' = c_{1,k} c_{2,k}. \end{aligned}$$

En particular, las aproximaciones de Fourier verifican $x_n(t) = (x_{1,n} * x_{2,n})(t)$.

Ejemplo 6.6. Consideremos $h(t)$ la función T -periódica que vale $h(t) = Ae^{-t/\tau}$, sus coeficientes de Fourier son

$$c_k = \frac{1}{T} \int_0^T Ae^{-t/\tau} e^{-i2k\pi t/T} dt = \frac{\tau}{T} A (1 - e^{-T/\tau}) \frac{1}{1 + i\tau 2k\pi/T}$$

Si tomamos $A = T/\tau (1 - e^{-T/\tau})^{-1}$, obtenemos $c_k = (1 + i\tau 2k\pi/T)^{-1}$. Por lo tanto la solución periódica del circuito de la Figura 6.12(a) dada en (6.9) se escribe como

$$v(t) = (h * v_0)(t) = \frac{1}{T} \int_0^T h(t - t') v_0(t') dt'.$$

Función de Green: El resultado anterior se puede obtener directamente, pero es un poco laborioso. Si dividimos en el intervalo $[0, T]$ en $[0, t]$ y $[t, T]$, $v(t)$ se escribe en la forma

$$v(t) = (h * v_0)(t) = \frac{1}{T} \int_0^t h(t - t') v_0(t') dt' + \frac{1}{T} \int_t^T h(t - t') v_0(t') dt',$$

como $h(t)$ es T -periódica, tenemos $h(t - t') = h(T + t - t')$, de donde se deduce

$$v(t) = \frac{1}{T} \int_0^t h(t - t') v_0(t') dt' + \frac{1}{T} \int_t^T h(T + t - t') v_0(t') dt',$$

usando que $h(t) = Ae^{-t/\tau}$ para $t \in [0, T]$ (observemos que $T + t - t' \in [0, T]$ si $t' \in [t, T]$) se obtiene

$$v(t) = \frac{A}{T} \int_0^t e^{-(t-t')/\tau} v_0(t') dt' + \frac{A}{T} \int_t^T e^{-(T+t-t')/\tau} v_0(t') dt',$$

que se reescribe como

$$v(t) = \frac{Ae^{-t/\tau}}{T} \int_0^t e^{t'/\tau} v_0(t') dt' + \frac{Ae^{-(T+t)/\tau}}{T} \int_t^T e^{t'/\tau} v_0(t') dt'.$$

Si derivamos esta expresión, usando el teorema fundamental del cálculo, obtenemos

$$\begin{aligned} \dot{v}(t) &= -\frac{Ae^{-t/\tau}}{T\tau} \int_0^t e^{t'/\tau} v_0(t') dt' - \frac{Ae^{-(T+t)/\tau}}{T\tau} \int_t^T e^{t'/\tau} v_0(t') dt' \\ &\quad + \frac{Ae^{-t/\tau}}{T} e^{t/\tau} v_0(t) - \frac{Ae^{-(T+t)/\tau}}{T} e^{t/\tau} v_0(t), \end{aligned}$$

agrupando los dos primeros términos y los dos últimos, vemos que

$$\dot{v}(t) = -\frac{1}{\tau} v(t) + \frac{A(1 - e^{-T/\tau})}{T} v_0(t)$$

y por lo tanto $\tau \dot{v}(t) + v(t) = A\tau(1 - e^{-T/\tau})/T v_0(t)$. Por la elección que hicimos de la constante A , resulta $\tau \dot{v}(t) + v(t) = v_0(t)$. La función $h(t)$ se conoce como la función de Green asociada a la ecuación diferencial.

6.2. Transformada discreta de Fourier. Para calcular los coeficientes de una señal $x(t)$ debemos calcular las integrales indicadas en (6.5). El cálculo analítico de estas integrales puede ser difícil o directamente imposible si, por ejemplo, la señal considerada es el resultado de mediciones. Planteamos entonces alternativas numéricas.

6.2.1. Método de los trapecios. Entre los algoritmos más usados para calcular integrales en forma numérica se encuentra el método de trapecios. Para integrar $x(t)$ en un intervalo $[0, T]$, tomamos $0 = t_0 < t_1 < \dots < t_{N-1} < t_N = T$ y aproximamos la integral mediante el área (con signo) de la región trapezoidal encerrada por el segmento secante que pasa por dos puntos consecutivos del gráfico, es decir

$$\begin{aligned} \int_0^T x(t) dt &\cong \frac{x(t_0) + x(t_1)}{2} (t_1 - t_0) + \dots + \frac{x(t_{N-1}) + x(t_N)}{2} (t_N - t_{N-1}) \\ &= \frac{1}{2} (x(t_0) + 2x(t_1) + \dots + 2x(t_{N-1}) + x(t_N)). \end{aligned}$$

Si los intervalos son de la misma longitud: $(t_j - t_{j-1}) = \Delta t = T/N$ ($t_j = jT/N$), entonces

$$\int_0^T x(t)dt \cong \frac{T}{2N}(x(t_0) + 2x(t_1) + \cdots + 2x(t_{N-1}) + x(t_N)).$$

Si $x(t)$ es T -periódica, tenemos $x(t_0) = x(t_N)$, por lo tanto

$$\frac{1}{T} \int_0^T x(t)dt \cong \frac{1}{N}(x(t_0) + x(t_1) + \cdots + x(t_{N-1})).$$

6.2.2. Muestreo de señales. Para poder el cálculo de los coeficientes de una señal mediante intergación por trapecios necesitamos los valores de la misma en los tiempos $t_j = jT/N$ con $j = 0, \dots, N-1$. En la Figura 6.13(a) mostramos esquemáticamente el dispositivo físico para obtener el vector de muestras $\mathbf{x} = (x(t_0) x(t_1) \dots, x(t_{N-1}))$. Inicialmente, la señal física $x(t)$ se convierte en una señal eléctrica mediante un sistema sensor (micrófono, pHmetro, termopar, sensor fotoeléctrico, etc.) La señal eléctrica obtenida se procesa, generalmente en forma analógica, y luego se conecta a un interruptor activado por la señal de reloj $s(t)$ (Figura 6.13(b)), el circuito se cierra en los tiempos t_j , dando a la salida valores x_0, x_1, \dots, x_{N-1} proporcionales a los valores $x(t_j)$. Varios problemas prácticos deben ser tenidos en cuenta. Generalmente los sensores tienen respuesta no lineal, es decir la señal eléctrica obtenida por el sensor no es proporcional a la señal $x(t)$. Además, los sensores suelen tener baja sensibilidad, esto significa que las señales eléctricas obtenidas son muy débiles, lo que dificulta su medición. Ambos problemas pueden ser solucionados mediante circuitos electrónicos, que linealizan y amplifican la salida del sensor. Por otro lado, el muestreo teórico debe ser instantáneo, lo que es físicamente imposible. Por lo que, como resulta del muestreo, obtenemos un valor promedio en el intervalo de medición en lugar de $x(t_j)$. En la mayoría de las aplicaciones, las variaciones de la señal son lentas comparadas con la velocidad del interruptor, y por lo tanto se obtienen resultados similares a los teóricos del proceso de muestreo.

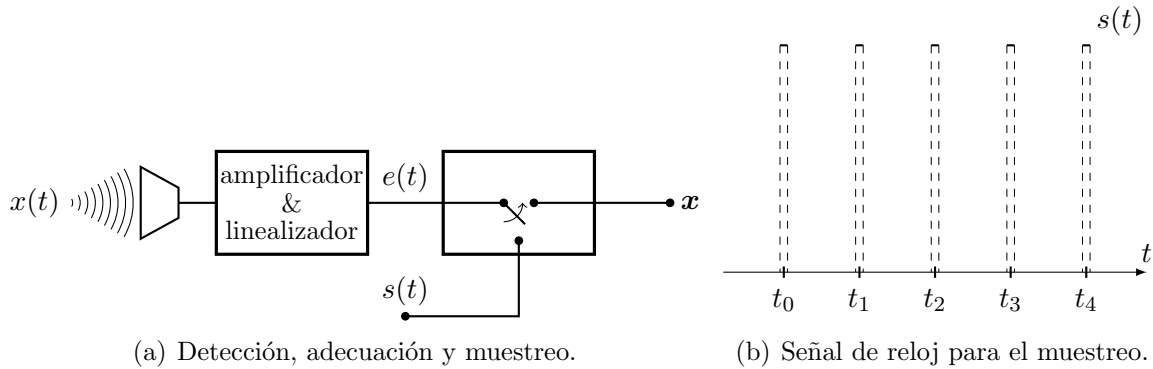


Fig. 6.13: Conversión de señales continuas en discretas.

Otro punto a tener en cuenta es la discretización de los resultados en un soporte digital. En principio $x_j = x(t_j)$ es un número real, pero para almacenar los valores muestreados se toma un conjunto finito de valores posibles. Por ejemplo, si se guardan en una computadora, los valores se representarían en punto flotante. Si se guardan en un soporte físico, como por ejemplo en un CD, se consideran niveles equiespaciados que se pueden expresar por un número entero. En los CD de audio, la frecuencia de muestreo es de 44.1 kHz, es decir $\Delta t = t_j - t_{j-1} = 22.6757 \mu s$, los valores adquiridos se representan en forma digital con 16 bits, lo que equivale a $2^{16} = 65536$ niveles.

6.2.3. Definición de DFT. Como resultado de la discretización de una señal $x(t)$, obtenemos un vector $\mathbf{x} \in \mathbb{C}^N$ dado por $\mathbf{x} = (x(t_0) x(t_1) \dots x(t_{N-1}))$, queremos calcular los coeficientes de Fourier de la señal definidos por (6.5) mediante la integración por trapecios:

$$(6.11) \quad \hat{x}_k = \frac{1}{N} \sum_{j=0}^{N-1} x(jT/N) e^{-i2\pi jk/N}.$$

definido para todo entero k . Observemos que $\hat{x}_{k+N} = \hat{x}_k$, dado que $e^{-i2\pi j} = 1$ y por lo tanto

$$e^{-i2\pi j(k+N)/N} = e^{-i2\pi jk/N - i2\pi j} = e^{-i2\pi jk/N} e^{-i2\pi j} = e^{-i2\pi jk/N}.$$

A partir de esta observación, definimos la transformada discreta de Fourier del vector $\mathbf{x} \in \mathbb{C}^N$ como el vector $\hat{\mathbf{x}} \in \mathbb{C}^N$ dado por $\hat{\mathbf{x}} = (\hat{x}_0 \hat{x}_1 \dots \hat{x}_{N-1})$. El Algoritmo 6.1 permite calcular el vector $\hat{\mathbf{x}}$ (*xf_list*) teniendo como dato \mathbf{x} (*x_list*) y N . El número de multiplicaciones necesarias para calcular la transformada discreta de Fourier con este algoritmo es N^2 . En la subsección 6.2.6 presentamos el algoritmo denominado FFT (Fast Fourier Transform) que realiza el cálculo con muchas menos operaciones.

Algoritmo 6.1: Transformada discreta de Fourier.

Data: N, x_list
Result: xf_list
for $k = 0$ **to** $N - 1$ **do**
 $xf_list(k) = 0;$
 for $j = 0$ **to** $N - 1$ **do**
 $xf_list(k) = xf_list(k) + x_list(j) \exp(-i2\pi jk/N);$
 end
 $xf_list(k) = xf_list(k)/N;$
end
return $xf_list;$

Ejemplo 6.7. Consideremos al señal armónica $x(t) = \sin(2\pi t)$ de período $T = 1$, para $N = 7$ ($\Delta t \cong 0.142857$) obtenemos la señal discreta

$$\mathbf{x} = (0, 0.781831, 0.974928, 0.433884, -0.433884, -0.974928, -0.781831)$$

La transformada discreta de Fourier del vector \mathbf{x} vale

$$\hat{\mathbf{x}} = (0, -i 0.5, 0, 0, 0, 0, i 0.5)$$

Si modificamos la frecuencia de la señal y tomamos $x(t) = \sin(6\pi t)$, también como una señal de período $T = 1$, obtenemos

$$\begin{aligned} \mathbf{x} &= (0, 0.433884, -0.781831, 0.974928, -0.974928, 0.781831, -0.433884), \\ \hat{\mathbf{x}} &= (0, 0, 0, -i 0.5, i 0.5, 0, 0). \end{aligned}$$

Recordando que $\sin(2k\pi t) = -0.5 i e^{i2k\pi t} + 0.5 i e^{-i2k\pi t}$, vemos que el vector $\hat{\mathbf{x}}$ contiene información sobre los coeficientes de Fourier de estas señales. Tomemos el caso $x(t) = \cos(4\pi t)$, los cálculos dan como resultado

$$\begin{aligned} \mathbf{x} &= (1., -0.222521, -0.900969, 0.62349, 0.62349, -0.900969, -0.222521), \\ \hat{\mathbf{x}} &= (0, 0, 0.5, 0, 0, 0.5, 0). \end{aligned}$$

Observemos que $\cos(2k\pi t) = 0.5 e^{i2k\pi t} + 0.5 e^{-i2k\pi t}$.

Si N es impar, $N = 2n + 1$, y \mathbf{x} es la señal discreta que se obtiene por muestreo de la señal $x(t)$, dada por (6.3), se puede ver analíticamente que el vector transformado es

$$\hat{\mathbf{x}} = (c_0 \ c_1 \ \dots \ c_n \ c_{-n} \ \dots \ c_{-1}).$$

En el caso N par, $N = 2n$, vale

$$\hat{\mathbf{x}} = (c_0 \ c_1 \ \dots \ c_{n-1} \ a_n \ c_{-n+1} \ \dots \ c_{-1}),$$

recordemos que $a_n = c_n + c_{-n}$.

6.2.4. Propiedades de DFT. Desde un punto de vista algebraico, la transformada discreta de Fourier es una aplicación lineal de \mathbb{C}^N , es decir si $\mathbf{z} = \alpha \mathbf{x} + \beta \mathbf{y}$, donde $\alpha, \beta \in \mathbb{C}$ y $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$, entonces $\hat{\mathbf{z}} = \alpha \hat{\mathbf{x}} + \beta \hat{\mathbf{y}}$. Existe una versión de la identidad de Parseval para la transformada discreta de Fourier:

$$\|\mathbf{x}\|^2 = \sum_{j=0}^{N-1} |x_j|^2 = N \sum_{k=0}^{N-1} |\hat{x}_k|^2 = N \|\hat{\mathbf{x}}\|^2.$$

Podemos escribir la identidad anterior de la siguiente forma: si $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$, entonces $\mathbf{x} \cdot \mathbf{y} = N \hat{\mathbf{x}} \cdot \hat{\mathbf{y}}$, donde el producto interno en \mathbb{C}^N se define como

$$\mathbf{x} \cdot \mathbf{y} = \sum_{j=0}^{N-1} \bar{x}_j y_j$$

Como consecuencia de la identidad de Parseval, vemos que la transformación es inyectiva y por lo tanto inversible. Su inversa es una transformación similar la transformada discreta de Fourier, su expresión es

$$x_j = \sum_{k=0}^{N-1} \hat{x}_k e^{i2\pi jk/N}.$$

Nota: Existen otras parametrizaciones de la transformada discreta de Fourier y su inversa, si p es un número real y q un entero coprimo⁹ con N , entonces

$$\hat{x}_k = \frac{1}{N^p} \sum_{j=0}^{N-1} x_j e^{-i2\pi qjk/N}, \quad x_j = \frac{1}{N^{1-p}} \sum_{k=0}^{N-1} \hat{x}_k e^{i2\pi qjk/N}.$$

forman un par de transformadas discretas de Fourier, directa e inversa. Salvo aviso en contrario, vamos a considerar $p = 1$ y $q = 1$. Si $p = 1/2$, tanto la transformada directa como la inversa son transformaciones unitarias¹⁰.

Varias operaciones sobre vectores de \mathbb{C}^N se pueden ver en términos de su transformada. En algunos casos es más simple pensar al espacio \mathbb{C}^N como las sucesiones N -periódicas

$$(\dots x_{N-1} \underbrace{x_0 \ x_1 \ \dots \ x_{N-1}}_{\mathbf{x}} \ x_0 \ x_1 \ \dots),$$

por ejemplo, la rotación hacia la izquierda $\mathbf{y} = L \cdot \mathbf{x} = (x_1 \ x_2 \ \dots \ x_0)$ se expresa como $y_j = x_{j+1}$:

$$(\dots x_0 \underbrace{x_1 \ x_2 \ \dots \ x_0}_{\mathbf{y}} \ x_1 \ x_2 \ \dots).$$

⁹El máximo común divisor entre q y N es 1.

¹⁰Una transformación lineal es unitaria si la norma de un vector y el de su transformado valen lo mismo.

La rotación hacia la derecha $\mathbf{z} = \mathbf{R}\mathbf{x} = (x_{N-1} x_0 \dots x_{N-2})$, equivale a $z_j = x_{j-1}$:

$$(\dots x_{N-2} \underbrace{x_{N-1} x_0 \dots x_{N-2}}_{\mathbf{z}} x_{N-1} x_0 \dots).$$

Claramente, $\sum_{j=0}^{N-1} x_j = \sum_{j=m}^{N+m-1} x_j$ para cualquier m entero. Es fácil ver que $\hat{y}_k = e^{i2\pi k/N} \hat{x}_k$ y $\hat{z}_k = e^{-i2\pi k/N} \hat{x}_k$. Efectivamente,

$$\hat{y}_k = \frac{1}{N} \sum_{j=0}^{N-1} y_j e^{-i2\pi jk/N} = \frac{1}{N} \sum_{j=0}^{N-1} x_{j+1} e^{-i2\pi jk/N},$$

cambiando el índice $j \rightarrow j-1$

$$\hat{y}_k = \frac{1}{N} \sum_{j=0}^{N-1} x_j e^{-i2\pi(j-1)k/N} = e^{i2\pi k/N} \frac{1}{N} \sum_{j=0}^{N-1} x_j e^{-i2\pi jk/N} = e^{i2\pi k/N} \hat{x}_k.$$

Se puede definir la convolución de vectores de \mathbb{C}^N , si $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$, definimos $\mathbf{z} = \mathbf{x} * \mathbf{y}$

$$z_j = \frac{1}{N} \sum_{l=0}^{N-1} x_{j-l} y_l = \frac{1}{N} \sum_{l=0}^j x_{j-l} y_l + \sum_{l=j+1}^{N-1} x_{N+j-l} y_l,$$

la transformada discreta de Fourier de la convolución está dada por

$$\hat{z}_k = \frac{1}{N} \sum_{j=0}^{N-1} z_j e^{-i2\pi jk/N} = \frac{1}{N^2} \sum_{j=0}^{N-1} \sum_{l=0}^{N-1} x_{j-l} y_l e^{-i2\pi jk/N},$$

usando $e^{-i2\pi jk/N} = e^{-i2\pi(j-l)k/N} e^{-i2\pi lk/N}$

$$\hat{z}_k = \frac{1}{N} \sum_{j=0}^{N-1} \frac{1}{N} \sum_{l=0}^{N-1} x_{j-l} y_l e^{-i2\pi(j-l)k/N} e^{-i2\pi lk/N} = \hat{x}_k \hat{y}_k.$$

Ejemplo 6.8. Tomemos los vectores $\mathbf{x}, \mathbf{y} \in \mathbb{C}^5$ $\mathbf{x} = (-5, 9, -7, 2, 7)$, $\mathbf{y} = (8, -3, -9, 8, -5)$, el vector convolución $\mathbf{z} = \mathbf{x} * \mathbf{y}$ vale $\mathbf{z} = (-36, 15, 1.6, -23.8, 42)$. Sus transformadas son

$$\begin{aligned} \hat{\mathbf{x}} &= (1.200, 0.798 + i0.678, -3.900 - i1.950, -3.900 + i1.950, 0.798 - i0.678), \\ \hat{\mathbf{y}} &= (-0.200, 1.270 + i1.620, 2.830 - i3.470, 2.830 + i3.470, 1.270 - i1.620), \\ \hat{\mathbf{z}} &= (-0.240, -0.085 + i2.150, -17.800 + i8.010, -17.800 - i8.010, -0.085 - i2.150), \end{aligned}$$

vemos que se verifica $\hat{z}_k = \hat{x}_k \hat{y}_k$. Se puede verificar también $\mathbf{x} \cdot \mathbf{y} = 5 \hat{\mathbf{x}} \cdot \hat{\mathbf{y}} = -23$.

6.2.5. Reconstrucción exacta y solapamiento. Es posible mostrar que si una señal $x(t)$ está dada por (6.3) y $N > 2n$, entonces \hat{x}_k coincide exactamente con c_k . Esto permite la reconstrucción de la señal sin pérdida de información.

Ejemplo 6.9. Consideremos una señal de período 1 dada por (6.3) con $n = 3$, por ejemplo:

$$x(t) = -e^{-6i\pi t} + 2e^{-4i\pi t} + 4e^{-2i\pi t} + 5 - e^{2i\pi t} + 3e^{4i\pi t} - 2e^{6i\pi t},$$

si tomamos $N = 8 > 2n$ y definimos $\mathbf{x} = (x(t_0) x(t_1) \dots x(t_7))$ con $t_j = j\Delta t$ para $j = 0, \dots, 7$ y $\Delta t = 1/8$, obtenemos

$$\mathbf{x} = (10, 9.243 - i3.243, -i4, 0.757 - i5.243, 10, 0.757 + i5.243, i4, 9.243 + i3.243),$$

la transformada de Fourier de \mathbf{x} vale

$$\hat{\mathbf{x}} = (5, -1, 3, -2, 0, -1, 2, 4) = (c_0 c_1 c_2 c_3 0 c_{-3} c_{-2} c_{-1})$$

que recupera los coeficientes de $x(t)$. Como $N > 2n + 1$, aparecen $N - 2n - 1$ ceros.

Vamos a estudiar que sucede si $N \leq 2n$, la muestra $\mathbf{x} = (x(t_0) x(t_1) \dots x(t_{N-1}))$ no es suficiente para recuperar los coeficientes de la señal. En la Figura 6.14 mostramos las señales $x(t) = 1.5 + \cos(2\pi t)$ y $y(t) = 1.5 + \cos(14\pi t)$, si tomamos $N = 8$ vemos que $\mathbf{x} = \mathbf{y}$, por lo tanto $\hat{\mathbf{x}} = \hat{\mathbf{y}}$. Esto muestra que no podemos distinguir entre estas dos señales.

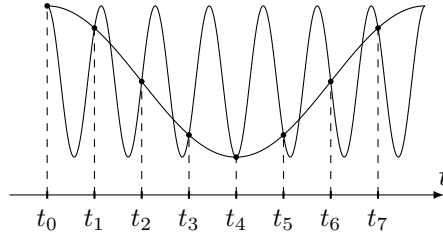


Fig. 6.14: Muestras de $x(t)$, $y(t)$ para $N = 8$.

Este fenómeno se conoce como solapamiento (aliasing) y produce distorsiones cuando se digitalizan señales (audio, imágenes, etc.) con baja definición.

Ejemplo 6.10. Si $y(t) = x(t) + e^{-10i\pi t}$, $x(t)$ es la señal del Ejemplo 6.9, es decir

$$y(t) = -e^{-6i\pi t} + 2e^{-4i\pi t} + 4e^{-2i\pi t} + 5 - e^{2i\pi t} + 3e^{4i\pi t} - 2e^{6i\pi t} + 4e^{-10i\pi t},$$

si $N = 8$ y $\mathbf{y} = (y(t_0) y(t_1) \dots y(t_7))$ con $t_j = j\Delta t$ para $j = 0, \dots, 7$ y $\Delta t = 1/8$, obtenemos

$$\mathbf{y} = (14, 6.414 - i0.414, -i8, 3.586 - i2.414, 6, 3.586 + i2.414, i8, 6.414 + i0.414),$$

la transformada de Fourier de \mathbf{y} vale

$$\hat{\mathbf{y}} = (5, -1, 3, 2, 0, -1, 2, 4) = (c_0, c_1, c_2, c_3 + c_{-5}, 0, c_{-3}, c_{-2}, c_{-1}),$$

vemos que en \hat{x}_3 se solapan c_3 y c_{-5} .

Consideremos ahora una señal periódica arbitraria $x(t)$, por lo visto arriba si discretizamos con N puntos por período $\mathbf{x} = (x(t_0) x(t_1) \dots x(t_{N-1}))$ la transformada discreta de Fourier $\hat{\mathbf{x}}$ no coincide con los primeros coeficientes de Fourier de $x(t)$, de hecho se puede probar que

$$\hat{x}_k = \dots + c_{k-N} + c_k + c_{k+N} + \dots = \sum_{l=-\infty}^{\infty} c_{k+lN},$$

de donde obtenemos para $0 \leq k < N/2$

$$\begin{aligned} \hat{x}_k - c_k &= \dots + c_{k-N} + c_{k+N} + c_{k+2N} + \dots = \sum_{\substack{l=-\infty \\ l \neq 0}}^{\infty} c_{k+lN}, \\ \hat{x}_{N-k} - c_{-k} &= \dots + c_{-k-N} + c_{-k+N} + c_{-k+2N} + \dots = \sum_{\substack{l=-\infty \\ l \neq 0}}^{\infty} c_{-k+lN}. \end{aligned}$$

Podríamos pensar que no hay información sobre c_k en el vector $\hat{\mathbf{x}}$. Pero si suponemos que los coeficientes de Fourier tienden rápidamente a cero, $|c_k| = O(|k|^{-q})$ con $q > 1$, entonces se puede probar que $|\hat{x}_k - c_k|, |\hat{x}_{N-k} - c_{-k}| = O(N^{-q})$, si $0 \leq k < N/2$.

Ejemplo 6.11. Volviendo a la señal triangular del Ejemplo 6.3, vemos que $c_0 = a_0 = 0$ y siendo $b_k = 0$, vale

$$c_k = a_{|k|} = \frac{2(1 - (-1)^k)}{\pi^2 k^2}$$

por lo tanto $|c_k| = O(|k|^{-2})$. En la Tabla 6.2 exhibimos las aproximaciones de c_1 , c_7 y c_{-3} mediante la transformada discreta de Fourier para $N = 16, 1024, 16384$. La última columna muestra el comportamiento asintótico de la aproximación.

N	c_1	\hat{x}_1	$ \hat{x}_1 - c_1 $	$N^2 \hat{x}_1 - c_1 $
2^4	0.405 28	0.410 53	5.2487×10^{-3}	1.343 70
2^{10}		0.405 29	1.2716×10^{-6}	1.333 30
2^{14}		0.405 28	4.9671×10^{-9}	1.333 30
N	c_7	\hat{x}_7	$ \hat{x}_7 - c_7 $	$N^2 \hat{x}_7 - c_7 $
2^4	8.2711×10^{-3}	1.6243×10^{-2}	7.9721×10^{-3}	2.040 90
2^{10}		8.2724×10^{-3}	1.2717×10^{-6}	1.333 50
2^{14}		8.2711×10^{-3}	4.9671×10^{-9}	1.333 30
N	c_{-3}	\hat{x}_{N-3}	$ \hat{x}_{N-3} - c_{-3} $	$N^2 \hat{x}_{N-3} - c_{-3} $
2^4	4.5032×10^{-2}	5.0622×10^{-2}	5.5907×10^{-3}	1.431 20
2^{10}		4.5033×10^{-2}	1.2716×10^{-6}	1.333 40
2^{14}		4.5032×10^{-2}	4.9671×10^{-9}	1.333 30

Tabla 6.2: Aproximación de los coeficientes de Fourier por DFT.

6.2.6. Transformada rápida de Fourier. Vamos a estudiar ahora un algoritmo eficiente para calcular la transformada discreta de Fourier. Podemos ver que para calcular cada coeficiente \hat{x}_k mediante la fórmula (6.11) requiere N operaciones de multiplicación, es decir que se necesitan N^2 operaciones (sin contar sumas) para calcular $\hat{\mathbf{x}}$. Esto representa el costo computacional de la multiplicación de una matriz $N \times N$ por un vector. Pero por la estructura particular de esta matriz, permite calcular el vector producto con muchas menos operaciones. El algoritmo que se conoce como transformada rápida de Fourier (FFT) fue propuesto por J. Cooley y W. Tukey en 1965 [?]. Se pueden encontrar algunos datos históricos sobre este algoritmo en la página <http://ocw.nctu.edu.tw/course/fourier/supplement/heideman-johnson-et-al1985.pdf>. El algoritmo se basa en la técnica *divide et impera*: suponiendo que N es par, podemos escribir $\mathbf{x} = (x_0, 0, x_2, \dots, x_{N-2}, 0) + (0, x_1, 0, \dots, 0, x_{N-1})$, entonces

$$\begin{aligned} \hat{x}_k &= \frac{1}{N} \sum_{j=0}^{N/2-1} x_{2j} e^{-i4\pi jk/N} + \frac{1}{N} \sum_{j=0}^{N/2-1} x_{2j+1} e^{-i2\pi(2j+1)k/N} \\ &= \frac{1}{N} \sum_{j=0}^{N/2-1} x_{2j} e^{-i4\pi jk/N} + e^{-i2\pi k/N} \frac{1}{N} \sum_{j=0}^{N/2-1} x_{2j+1} e^{-i4\pi jk/N}. \end{aligned}$$

Si definimos $\mathbf{y}, \mathbf{z} \in \mathbb{C}^{N/2}$ como $\mathbf{y} = (x_0 \ x_2 \ \dots \ x_{N-2})$, $\mathbf{z} = (x_1 \ x_3 \ \dots \ x_{N-1})$, tenemos

$$\hat{y}_k = \frac{2}{N} \sum_{j=0}^{N/2-1} x_{2j} e^{-i4\pi jk/N}, \quad \hat{z}_k = \frac{2}{N} \sum_{j=0}^{N/2-1} x_{2j+1} e^{-i4\pi jk/N}$$

y por lo tanto $\hat{x}_k = (\hat{y}_k + \hat{\zeta}_k)/2$ para $k = 0, \dots, N/2 - 1$, con $\hat{\zeta}_k = e^{-i2\pi k/N} \hat{z}_k$. Para el rango $k = N/2, \dots, N - 1$, usando que $e^{-i4\pi jk/N} = e^{-i4\pi j(k-N/2)/N}$ y $e^{-i2\pi k/N} = -e^{-i2\pi(k-N/2)/N}$, tenemos $\hat{x}_k = (\hat{y}_{k-N/2} - \hat{\zeta}_{k-N/2})/2$. En resumen

$$\hat{x}_k = \frac{1}{2} \begin{cases} \hat{y}_k + \hat{\zeta}_k & k = 0, \dots, N/2 - 1, \\ \hat{y}_{k-N/2} - \hat{\zeta}_{k-N/2} & k = N/2, \dots, N - 1, \end{cases}$$

Costo computacional de FFT

Para calcular el vector $\hat{\mathbf{x}} \in \mathbb{C}^N$ mediante el algoritmo descrito anteriormente, necesitamos calcular $\hat{\mathbf{y}}, \hat{\mathbf{z}} \in \mathbb{C}^{N/2}$. Además, deben realizarse un número de operaciones proporcional a N . Por lo tanto, si C_N es el costo computacional (medido en operaciones) de obtener $\hat{\mathbf{x}} \in \mathbb{C}^N$, vemos que $C_N = 2C_{N/2} + rN$. Si $N/2$ es par, podemos repetir el procedimiento obteniendo

$$C_N = 2(2C_{N/4} + rN/2) + rN = 4C_{N/4} + 2rN.$$

En general, si $N = 2^k p$, inductivamente obtenemos $C_N = 2^k C_p + krN$. Para $N = 2^k$, vemos que $C_N = 2^k C_1 + krN$, como $C_1 = 0$ resulta $C_N = rN \log_2(N)$. El número r se relaciona con la forma de ponderar las operaciones: para calcular $\hat{\zeta}_k$, $k = 0, \dots, N/2 - 1$ se necesitan $N/2$ productos de números complejos y para calcular $\hat{y}_k \pm \hat{\zeta}_k$, N sumas complejas. Una suma compleja se realiza mediante dos sumas reales y un producto entre números complejos requiere cuatro productos y dos sumas. Por lo tanto, se necesitan $2N$ productos y $3N$ sumas reales.

Producto de números complejos Si $z = a + ib$ y $w = c + id$, el producto zw se calcula como

$$\operatorname{Re}(zw) = ac - bd, \quad \operatorname{Im}(zw) = (ad + bc),$$

lo que requiere cuatro productos y dos sumas (parte real y parte imaginaria se almacenan en registros separados). Definiendo los productos $p_1 = (a + b)c$, $p_2 = b(c + d)$, $p_3 = a(c - d)$, tenemos

$$\operatorname{Re}(zw) = p_1 - p_2, \quad \operatorname{Im}(zw) = p_1 - p_3,$$

de esta forma se necesitan solo tres productos y cinco sumas. En las antiguas implementaciones de las operaciones aritméticas, los productos solían ser significativamente más costosos que las sumas, por lo que esta última forma de calcular el producto era más eficiente.

Diezmado en frecuencia

El algoritmo visto anteriormente se conoce como transformada rápida de Fourier con diezmado en tiempo (DIT-FFT). Existe una forma equivalente de calcular $\hat{\mathbf{x}}$ denominada diezmado en frecuencia (DIF-FFT), la cual consiste en considerar

$$\begin{aligned} \hat{x}_k &= \frac{1}{N} \sum_{j=0}^{N/2-1} x_j e^{-i2\pi jk/N} + \frac{1}{N} \sum_{j=N/2}^{N-1} x_j e^{-i2\pi jk/N} \\ &= \frac{1}{N} \sum_{j=0}^{N/2-1} x_j e^{-i2\pi jk/N} + \frac{1}{N} \sum_{j=0}^{N/2-1} x_{j+N/2} e^{-i2\pi(j+N/2)k/N}. \end{aligned}$$

Usando que $e^{-i2\pi(j+N/2)k/N} = (-1)^k e^{-i2\pi jk/N}$ obtenemos

$$\hat{x}_k = \frac{1}{N} \sum_{j=0}^{N/2-1} x_j e^{-i2\pi jk/N} + \frac{1}{N} \sum_{j=0}^{N/2-1} x_{j+N/2} (-1)^k e^{-i2\pi jk/N}.$$

Si $k = 2m$, entonces

$$\hat{x}_{2m} = \frac{1}{N} \sum_{j=0}^{N/2-1} (x_j + x_{j+N/2}) e^{-i4\pi jm/N},$$

y para $k = 2m + 1$ tenemos

$$\hat{x}_{2m+1} = \frac{1}{N} \sum_{j=0}^{N/2-1} (x_j - x_{j+N/2}) e^{-i2\pi j/N} e^{-i4\pi jm/N}.$$

Definiendo $\mathbf{y}, \mathbf{z} \in \mathbb{C}^{N/2}$ como $y_j = x_j + x_{j+N/2}$, $z_j = (x_j - x_{j+N/2}) e^{-i2\pi j/N}$, $j = 0, \dots, N/2 - 1$, tenemos

$$\hat{x}_k = \frac{1}{2} \begin{cases} \hat{y}_m & k = 2m, \\ \hat{z}_m & k = 2m + 1, \end{cases}$$

Raíces de la unidad Para calcular la transformada rápida de Fourier se necesita conocer las raíces de la unidad $e^{i2\pi k/N}$ con $k = 0, \dots, N/2 - 1$. Como $w^N - 1$ es un polinomio real, si w es raíz, \bar{w} también lo es. Además, si N es un múltiplo de 4, $(iw)^N = w^N$. Como se gráfica en la Figura 6.15, si w es una raíz de la unidad, también resultan ser raíces $i\bar{w}$, iw , $-\bar{w}$. Supongamos que $N = 2^k$ con $k \geq 3$, entonces $w_k = 0, (1+i)/\sqrt{2}, i, (-1+i)/\sqrt{2}$ son raíces de la unidad correspondientes a $k = 0, N/8, N/4, 3N/8$ (marcadas en negro en la Figura 6.15). Debido a las simetrías mencionadas anteriormente, basta conocer las raíces w_k para $k = 1, \dots, N/8 - 1$.

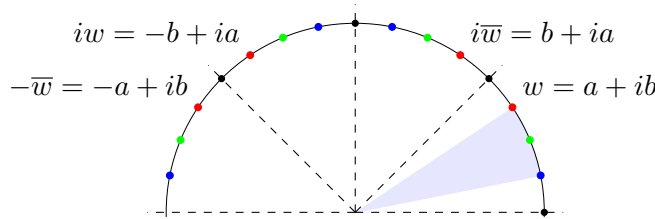


Fig. 6.15: Raíces de la unidad para $N = 32$.

6.2.7. Espectro de potencia. Dada $\mathbf{x} \in \mathbb{C}^N$, definimos como el espectro de potencia al vector $\mathbf{s} = (|\hat{x}_0|^2, |\hat{x}_1|^2, \dots, |\hat{x}_{N-1}|^2)$. Por la igualdad de Parseval, tenemos

$$\sum_{k=0}^{N-1} s_k = \sum_{k=0}^{N-1} |\hat{x}_k|^2 = \frac{1}{N} \sum_{j=0}^{N-1} |x_j|^2$$

Si para $\mathbf{x} \in \mathbb{C}^N$, $\tilde{\mathbf{x}} \in \mathbb{C}^N$ está dado por $\tilde{\mathbf{x}} = (\bar{x}_0, \bar{x}_{N-1}, \dots, \bar{x}_1)$, entonces su transformada discreta de Fourier vale

$$\begin{aligned} \hat{\tilde{x}}_k &= \frac{1}{N} \sum_{j=0}^{N-1} \tilde{x}_j e^{-i2\pi jk/N} = \frac{1}{N} \sum_{j=0}^{N-1} \bar{x}_{N-j} e^{-i2\pi jk/N} \\ &= \frac{1}{N} \sum_{j=0}^{N-1} \bar{x}_{N-j} e^{i2\pi(N-j)k/N} = \frac{1}{N} \sum_{j=0}^{N-1} \bar{x}_j e^{i2\pi jk/N} = \overline{\frac{1}{N} \sum_{j=0}^{N-1} x_j e^{-i2\pi jk/N}} = \bar{\hat{x}}_k. \end{aligned}$$

Si definimos $\mathbf{r} = \mathbf{x} * \tilde{\mathbf{x}}$, es decir

$$(6.12) \quad r_j = \frac{1}{N} \sum_{l=0}^{N-1} x_{j-l} \tilde{x}_l = \frac{1}{N} \sum_{l=0}^{N-1} x_{j-l} \bar{x}_{N-l} = \frac{1}{N} \sum_{l=0}^{N-1} x_{j+l} \bar{x}_l,$$

tenemos $\hat{\mathbf{r}} = (|\hat{x}_0|^2, |\hat{x}_1|^2, \dots, |\hat{x}_{N-1}|^2) = \mathbf{s}$. Al vector \mathbf{r} se lo suele denominar vector de autocorrelación.

Ejemplo 6.12. Si $\mathbf{x} = (-3 + i2, 6 + i7, i4, -4 + i4, 7) \in \mathbb{C}^5$, su espectro de potencia es

$$\mathbf{s} = (13, 5.3120, 4.9372, 15.4865, 0.2642).$$

Por otro lado podemos calcular el vector de autocorrelación usando (6.12)

$$\mathbf{r} = (39, -1.8 - i1.4, 14.8 + i13, 14.8 - i13, -1.8 + i1.4),$$

vemos que $\hat{\mathbf{r}} = \mathbf{s}$.

Es usual graficar al espectro de potencia en escala logarítmica, en muchos casos tomando un valor de referencia s_0 como valor 0 y expresando s en decibelios: $s(\text{dB}) = 10 \log_{10}(s/s_0)$. Por ejemplo, para tensión eléctrica se toma como 0 dBu el valor de 0.7746 V;¹¹ para presión sonora se toma la referencia de 20 μPa .¹² Al gráfico del espectro de potencia en dB se lo conoce como periodograma.

Determinación del período

Sea $x(t)$ es una señal periódica con frecuencia angular ω desconocida, queremos ver como obtener el período a partir del vector $\mathbf{x} \in \mathbb{C}^N$ obtenido tomando una muestra de $x(t)$. Consideremos en el intervalo $[0, T]$ los valores $x_j = x(t_j)$ para $j = 0, \dots, N-1$, donde $\Delta t = T/N$. Supongamos inicialmente $x(t) = \exp(i\omega t)$, entonces $x_j = \exp(i\omega j \Delta t)$, su transformada discreta es

$$\begin{aligned} \hat{x}_k &= \frac{1}{N} \sum_{j=0}^{N-1} x_j e^{-i2\pi jk/N} = \frac{1}{N} \sum_{j=0}^{N-1} e^{i(\omega T - 2\pi k)j/N} \\ &= \frac{1}{N} \frac{1 - e^{i(\omega T - 2\pi k)N/N}}{1 - e^{i(\omega T - 2\pi k)/N}} = \frac{e^{i(N-1)\beta_k} \text{sen}(N\beta_k)}{N \text{sen}(\beta_k)} \end{aligned}$$

donde $\beta_k = \frac{\omega T - 2\pi k}{2N} = \frac{\omega \Delta t}{2} - \pi \frac{k}{N}$. Por lo tanto el espectro de potencia $\mathbf{s} = (s_0 \dots s_{N-1})$ está dado por

$$s_k = |\hat{x}_k|^2 = \frac{\text{sen}^2(N\beta_k)}{N^2 \text{sen}^2(\beta_k)} = \frac{1}{N} F_N(2\beta_k),$$

donde F_N se denomina núcleo de Fejér y su gráfico se muestra en la Figura 6.16 para distintos valores de N . Como $\beta_{\min} \leq \beta_k \leq \beta_{\max}$, con $\beta_{\min} = \frac{\omega \Delta t}{2} - (1 - \frac{1}{N})\pi$ y $\beta_{\max} = \frac{\omega \Delta t}{2}$. Si $\omega \cong 2\pi k/T$

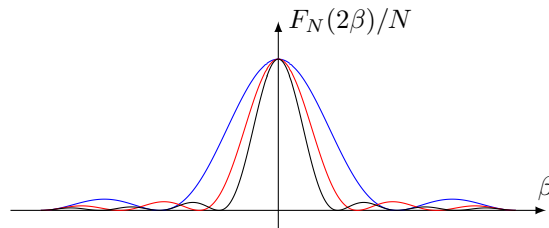


Fig. 6.16: Núcleo de Fejér para $N = 4, 6, 8$.

entonces $|\beta_k| \ll 1$ y por lo tanto $s_k = 1$, en otro caso $s_k = O(N^{-2})$. A partir de esto, podemos definir $\hat{\omega} = 2\pi \hat{k}/T$ donde $\hat{k} = \arg \min_k |\beta_k|$. Si $x(t) = c_1 \exp(i\omega_1 t) + c_2 \exp(i\omega_2 t)$, entonces

$$\hat{x}_k = c_1 \frac{e^{i(N-1)\beta_k^{(1)}} \text{sen}(N\beta_k^{(1)})}{N \text{sen}(\beta_k^{(1)})} + c_2 \frac{e^{i(N-1)\beta_k^{(2)}} \text{sen}(N\beta_k^{(2)})}{N \text{sen}(\beta_k^{(2)})}$$

¹¹Este valor proviene de las antiguas líneas de telegrafía y corresponde a la tensión necesaria para disipar 1 mW sobre una impedancia de 600 Ω (impedancia típicas de las líneas).

¹²Este nivel se fija por ser aproximadamente la fracción $1/5 \times 10^9$ de la presión atmosférica normal.

Ejemplo 6.13. Consideremos $x(t) = 0.8 + 0.8 \cos(\omega t) + 0.25 \cos(2\omega t) - 0.75 \sin(3\omega t)$ con $\omega = 4$, su gráfico en Figura 6.17(a) en el intervalo $[0, T]$ con $T = 10$. Si $\Delta t = 0.2$, tomamos $t_j = j\Delta t$, $j = 0, \dots, 49$. El espectro de potencia se muestra en la Figura 6.18(a), aparecen máximos locales se alcanzan en $k = 6, 13, 19, 31, 37, 44$, como $N = 50$ tenemos que las frecuencias asociadas a $k = 6$ y $k = 44$ corresponden a la frecuencia fundamental, $k = 13$ y $k = 31$ al segundo armónico y $k = 19$ y $k = 37$ al tercero. Tenemos entonces $\omega = 2\pi 6/T = 3.77$ para el primer caso, $2\omega = 2\pi 13/T = 8.17$ para el segundo y $3\omega = 2\pi 19/T = 11.94$ para el último, obteniendo los valores $\omega = 3.77, 4.08, 3.98$ respectivamente. Una forma más precisa se obtiene tomando los valores de s_k para los valores cercanos al primer máximo y ajustar la función

$$s_k = \frac{A \sin^2\left(\frac{\omega T - 2\pi k}{2}\right)}{N^2 \sin^2\left(\frac{\omega T - 2\pi k}{2N}\right)}$$

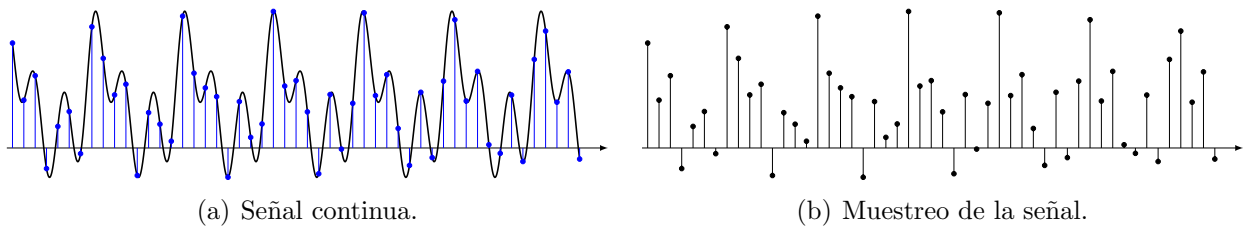


Fig. 6.17: Obtención de la señal discreta x .

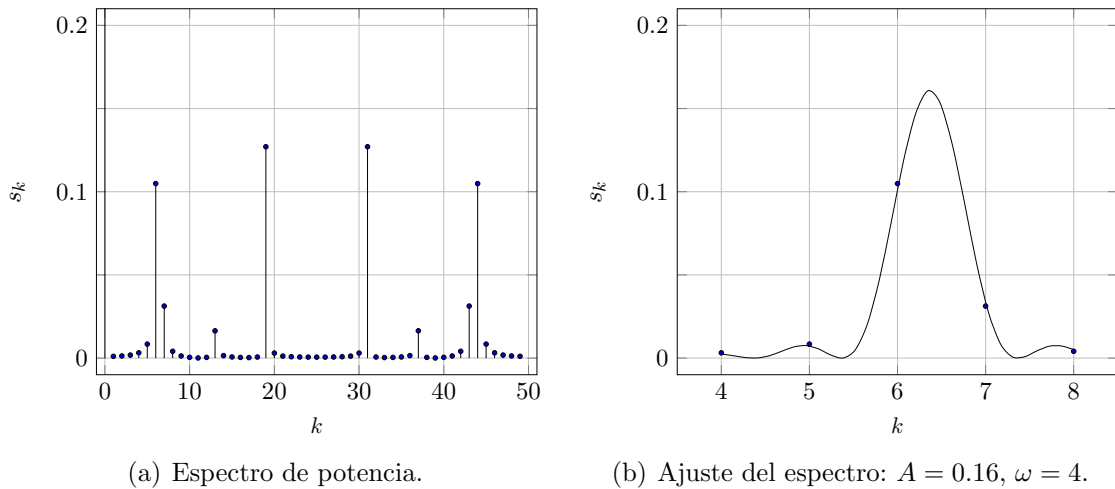
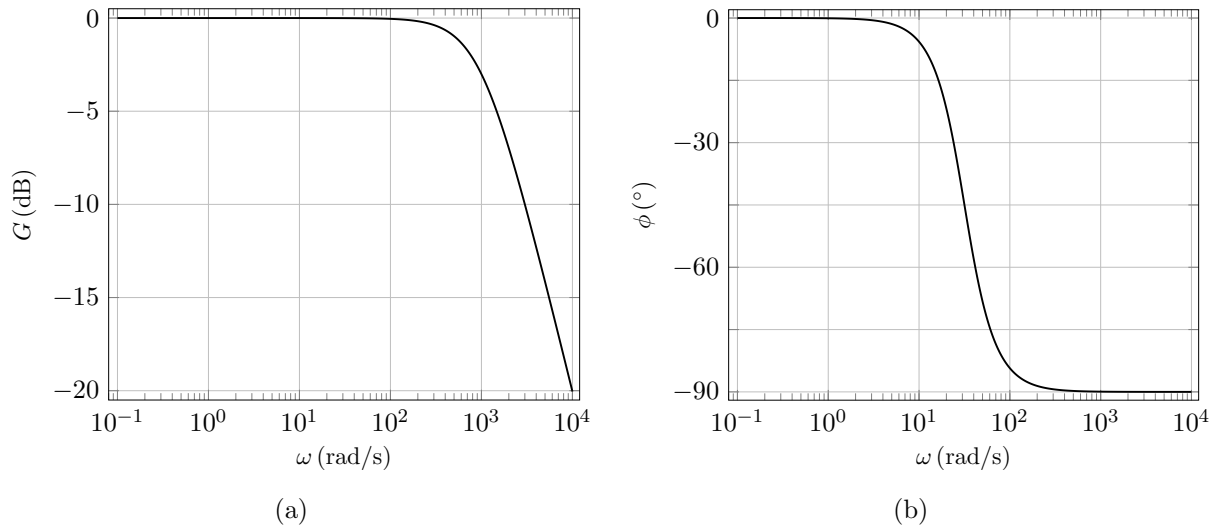


Fig. 6.18: Detección del período por el espectro de potencia.

6.2.8. Escala logarítmica. Decibeles.

6.2.9. Decibeles. Es usual expresar G en decibeles, $G(\omega)(\text{dB}) = 20 \log((1 + \omega_1^2 \tau^2)^{-1/2}) = -10 \log(1 + \omega^2 \tau^2)$. El gráficos de $G(\text{dB})$ y ϕ como función de ω en escala semi-logarítmica se llama diagrama de Bode. Los diagramas de Bode correspondientes al circuito 6.12(a) con $\tau = 1 \text{ ms}$ se muestran en la Figura 6.19.

6.3. Señales aperiódicas. Muchas señales de interés no presentan comportamiento periódico, nos gustaría ver de qué forma se pueden extender las ideas de Fourier a este caso.

Fig. 6.19: Diagramas de Bode del circuito de la Figura 6.12(a) para $\tau = 1$ ms.

6.3.1. Transformada de Fourier. Vamos a empezar suponiendo que las señales empiezan en un instante y terminan, es decir $x(t) = 0$ si $t \notin (t_i, t_f)$. Consideremos T suficientemente grande de forma que $(t_i, t_f) \subset (-T/2, T/2)$, consideramos las aproximaciones $x_n(t)$ de $x(t)$ dadas por 6.6 en el intervalo $(-T/2, T/2)$. Obviamente $x_n(t)$ es T -periódica, por lo tanto converge a la señal que se obtiene repitiendo $x(t)$ en cada intervalo $((l - 1/2)T, (l + 1/2)T)$ con l entero (ver Figura 6.20).

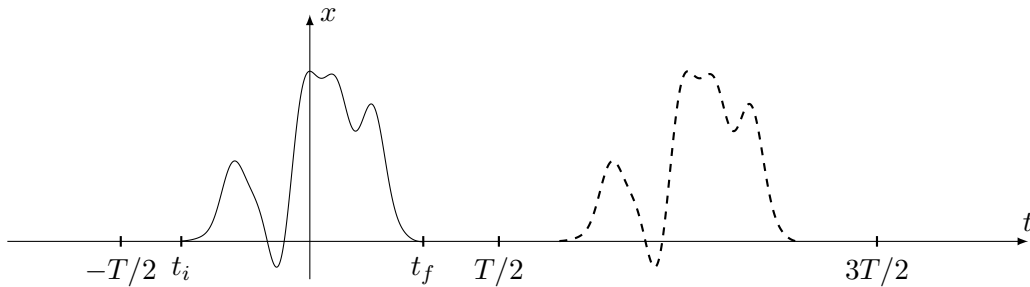


Fig. 6.20

Para $t \in (-T/2, T/2)$ se verifica

$$\begin{aligned} x(t) &= \lim_{n \rightarrow \infty} \sum_{k=-n}^n c_k e^{i2\pi kt/T} = \lim_{n \rightarrow \infty} \sum_{k=-n}^n \frac{e^{i2\pi kt/T}}{T} \int_{-T/2}^{T/2} x(t') e^{-i2\pi kt'/T} dt' \\ &= \lim_{n \rightarrow \infty} \sum_{k=-n}^n \frac{e^{i2\pi kt/T}}{T} \int_{t_i}^{t_f} x(t') e^{-i2\pi kt'/T} dt'. \end{aligned}$$

Si definimos la transformada

$$(6.13) \quad \hat{x}(\nu) = \int_{-\infty}^{\infty} x(t') e^{-i2\pi \nu t'/T} dt' = \int_{t_i}^{t_f} x(t') e^{-i2\pi \nu t'/T} dt',$$

entonces

$$x(t) = \lim_{n \rightarrow \infty} \sum_{k=-n}^n \frac{e^{i2\pi kt/T}}{T} \hat{x}(k/T).$$

La suma del lado izquierdo se puede pensar como una aproximación por trapecios de la integral

$$\int_{-\infty}^{\infty} e^{i2\pi t\nu} \hat{x}(\nu) d\nu \cong \lim_{n \rightarrow \infty} \frac{1}{T} \sum_{k=-n}^n e^{i2\pi t k/T} \hat{x}(k/T).$$

Se convierte en una igualdad cuando $T \rightarrow \infty$, es decir

$$(6.14) \quad x(t) = \int_{-\infty}^{\infty} e^{i2\pi t\nu} \hat{x}(\nu) d\nu$$

para $t \in \mathbb{R}$. Aunque el razonamiento anterior sea cuestionable desde el rigor matemático, el resultado es absolutamente válido. Observemos que si $y(\nu) = \hat{x}(\nu)$, entonces

$$(6.15) \quad \hat{y}(t) = \int_{-\infty}^{\infty} e^{-i2\pi t\nu} \hat{x}(\nu) d\nu = \int_{-\infty}^{\infty} e^{i2\pi(-t)\nu} \hat{x}(\nu) d\nu = x(-t).$$

Es decir, que si calculamos la transformada de Fourier de la transformada de Fourier es equivalente a revertir el tiempo. Si la aplicamos cuatro veces, obtenemos la función original.

Ejemplo 6.14. Si $\chi(t) = 1$ si $t \in (-\tau, \tau)$ y nula fuera de ese intervalo, entonces

$$\hat{\chi}(\nu) = \int_{-\tau}^{\tau} e^{-i2\pi t\nu} dt = \frac{e^{-i2\pi t\nu}}{-i2\pi\nu} \Big|_{t=-\tau}^{t=\tau} = \frac{\text{sen}(2\pi\tau\nu)}{\pi\nu},$$

si definimos la función $\text{sinc}(y) = \text{sen}(y)/y$, entonces $\hat{\chi}(\nu) = 2\tau \text{sinc}(2\pi\tau\nu)$.

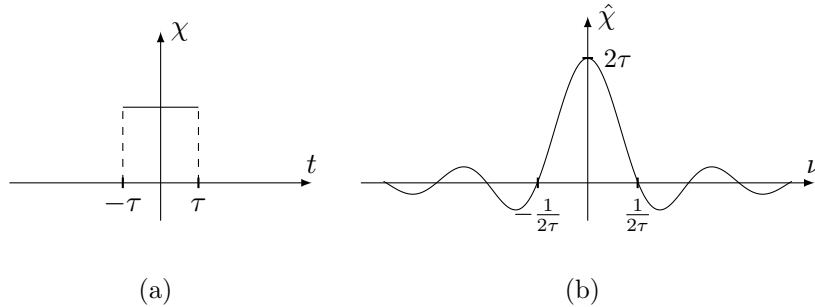


Fig. 6.21

Ejemplo 6.15. Si $\psi(t) = 1 - |t|/\tau$ si $t \in (-\tau, \tau)$ y nula fuera de ese intervalo, entonces

$$\hat{\psi}(\nu) = \int_{-\tau}^{\tau} e^{-i2\pi t\nu} (1 - |t|/\tau) dt = \frac{1 - \cos(2\pi\nu\tau)}{2\pi^2\nu^2\tau} = \tau \text{sinc}^2(\pi\nu\tau)$$

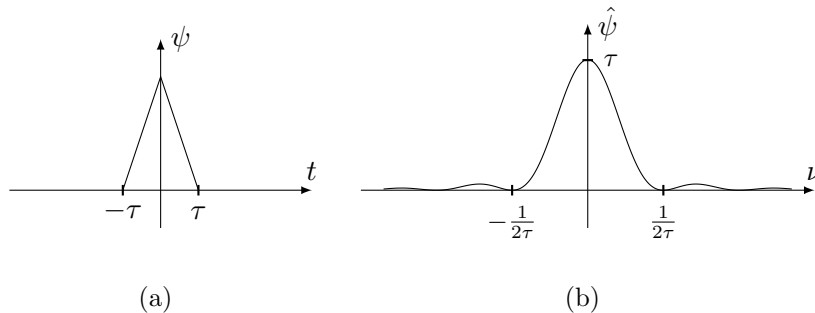


Fig. 6.22

Para los ejemplos anteriores, es difícil calcular (6.14) directamente. Métodos indirectos permiten mostrar la validez de esta igualdad.¹³ Pero podemos calcular dicha integral mediante el método de trapecios:

$$\psi(t) \cong \frac{\eta}{N} \sum_{k=-n}^n \hat{\psi}(k\eta/N) e^{i2\pi k\eta t/N},$$

donde η es una frecuencia suficientemente grande como para que podamos considerar $\hat{\psi}(\nu)$ despreciable fuera del intervalo $(-\eta, \eta)$, $N = 2n + 1$. Para $\tau = 1$, tomamos $\eta = 5$ y los tiempos

6.3.2. Propiedades de la transformada de Fourier. Claramente la transformada de Fourier es una operación lineal, esto quiere decir que si $x(t)$ y $y(t)$ son dos señales y $z(t)$ la señal que se obtiene por superposición, $z(t) = \alpha x(t) + \beta y(t)$ ($\alpha, \beta \in \mathbb{C}$), entonces $\hat{z}(\nu) = \alpha \hat{x}(\nu) + \beta \hat{y}(\nu)$. De la definición (6.13) y la fórmula de cambio de variable para integrales, se pueden probar las siguientes afirmaciones:

1. Si $y(t) = x(t - \tau)$, entonces $\hat{y}(\nu) = e^{-i2\pi\nu\tau} \hat{x}(\nu)$.
2. Si $y(t) = e^{i2\pi\eta t} x(t)$, entonces $\hat{y}(\nu) = \hat{x}(\nu - \eta)$.
3. Si $y(t) = x(-t)$, entonces $\hat{y}(\nu) = \hat{x}(-\nu)$.
4. Si $y(t) = \bar{x}(t)$, entonces $\hat{y}(\nu) = \bar{\hat{x}}(-\nu)$.
5. Si $y(t) = x(\lambda t)$, entonces $\hat{y}(\nu) = \lambda^{-1} \hat{x}(\lambda^{-1}\nu)$.

Supongamos que $x(t)$ es derivable y $x(t) \rightarrow 0$ cuando $t \rightarrow \pm\infty$, si $y(t) = \dot{x}(t)$, usando la fórmula de integración por partes vemos

$$\hat{y}(\nu) = \int_{-T}^T \dot{x}(t) e^{-i2\pi\nu t} dt = x(t) e^{-i2\pi\nu t} \Big|_{t=-T}^{t=T} + i2\pi\nu \int_{-T}^T x(t) e^{-i2\pi\nu t} dt,$$

tomando $T \rightarrow \infty$ obtenemos $\hat{y}(\nu) = i2\pi\nu \hat{x}(\nu)$.

Por otro lado, si aceptamos que se puede intercambiar la derivada con respecto a ν con la integral, vemos que

$$\hat{x}'(\nu) = \frac{d}{d\nu} \int_{-\infty}^{\infty} x(t) e^{-i2\pi\nu t} dt = -i2\pi \int_{-\infty}^{\infty} t x(t) e^{-i2\pi\nu t} dt,$$

entonces, si $y(t) = t x(t)$, vale $\hat{x}'(\nu) = -i2\pi \hat{y}(\nu)$. Estas expresiones, muestran que si una señal es varias veces derivable, su transformada de Fourier decae rápido en el infinito. Por otro lado, mientras más rápido decaiga una señal, más suave será su transformada de Fourier.

Consideremos la señal del ejemplo 6.14, $\chi(t)$ decae más rápido que cualquier potencia, es decir $\chi(t) = O(|t|^{-p})$ cuando $t \rightarrow \infty$ para todo $p > 0$, esto se refleja en su transformada de Fourier, que resulta indefinidamente diferenciable. Pero siendo que $\chi(t)$ no es derivable (solo es continua a trozos), el decaimiento de $\hat{\chi}(\nu) = O(|\nu|^{-1})$ cuando $\nu \rightarrow \infty$. En cambio, la señal $\psi(t)$ del ejemplo 6.15 es continua y derivable a trozos, su transformada de Fourier es también indefinidamente diferenciable y decae como $\hat{\psi}(\nu) = O(|\nu|^{-2})$ cuando $\nu \rightarrow \infty$.

En particular, si una función es indefinidamente derivable y decae más rápido que cualquier potencia, su transformada de Fourier tendrá estas características. Al conjunto de estas funciones se lo conoce como espacio de Schwartz. Por ejemplo, la señal gaussiana $\varphi_a(t) = e^{-at^2}$ verifica estas condiciones, por lo tanto también su transformada. Se puede probar que la transformada

¹³El cálculo de integrales por residuos es el más efectivo.

de Fourier es también una función gaussiana dada por $\hat{\varphi}_a(\nu) = \sqrt{\pi/a} e^{-\pi^2 t^2/a}$. En la Figura 6.23 se muestra la función gaussiana con dos valores distintos del parámetro $a = \pi, \pi/2$ (Figura 6.23(a)) y sus respectivas transformadas (Figura 6.23(b)).

Para calcular $\hat{\varphi}_a(\nu)$ empecemos con el caso $a = \pi$ ($\varphi_\pi(t) = \varphi(t)$), derivando obtenemos

$$\ddot{\varphi}(t) - 4\pi^2 t^2 \varphi(t) + 2\pi \varphi(t) = 0,$$

calculando la transformada de Fourier de cada término, resulta

$$-4\pi^2 \nu^2 \hat{\varphi}(\nu) + \hat{\varphi}''(\nu) + 2\pi \hat{\varphi}(\nu) = 0.$$

Sabemos de la teoría de ecuaciones diferenciales, que el conjunto de soluciones de la ecuación diferencial lineal de segundo orden es un espacio vectorial de dimensión dos. Recordemos que dos funciones son linealmente independientes si y solo si el wronskiano, $w(t) = y_1(t)y_2'(t) - y_1'(t)y_2(t)$ no es idénticamente nulo. Pero si $y_1(t), y_2(t)$ son soluciones de $\ddot{y}(t) + V(t)y(t) = 0$, $w(t)$ es constante ($\dot{w}(t) = 0$). Como $\varphi(t), \hat{\varphi}(t)$ y sus derivadas tienden a 0 cuando $t \rightarrow \pm\infty$, entonces $\lim_{t \rightarrow \infty} w(t) = 0$, siendo constante $w(t) = 0$ para todo t . Por lo tanto, existe $c \in \mathbb{R}$ tal que $\hat{\varphi}(t) = c\varphi(t)$. Aplicando transformada de Fourier a esta última igualdad, se deduce $\varphi(-t) = c^2 \varphi(t)$ y, como $\varphi(t)$ es una función par, tenemos $c = \pm 1$. Usando que $\varphi(0) = 1$ obtenemos

$$c = c\varphi(0) = \hat{\varphi}(0) = \int_{-\infty}^{\infty} \varphi(t) dt > 0,$$

por lo tanto $c = 1$. Para $a > 0$, podemos tomar $\lambda = \sqrt{a/\pi}$ y $\varphi_a(t) = e^{-at^2} = \varphi(\lambda t)$. Por lo tanto, su transformada de Fourier es $\hat{\varphi}_a(\nu) = \sqrt{\pi/a} e^{-\pi^2 t^2/a}$.

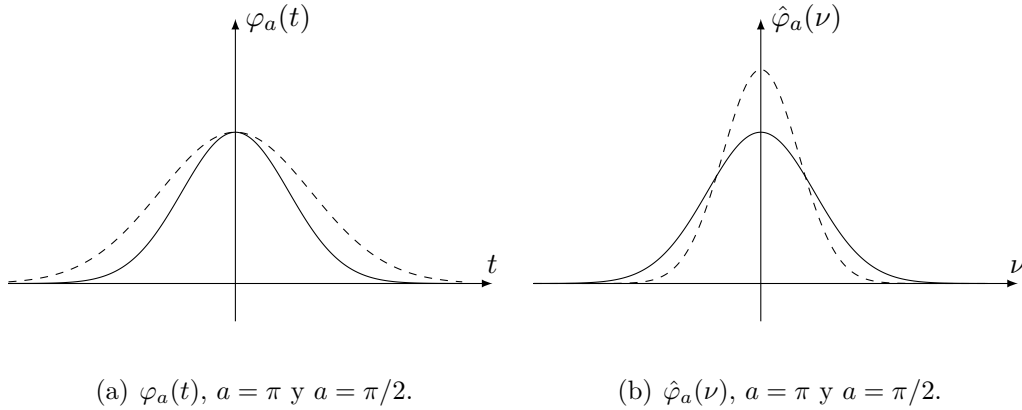


Fig. 6.23

6.3.3. Convolución de señales aperiódicas. Igual que en el caso periódico, podemos definir la operación de convolución entre señales. Si tomamos la convolución definida en este caso como $z(t) = (x * y)(t)$

$$z(t) = \int_{-\infty}^{\infty} x(t-t')y(t')dt',$$

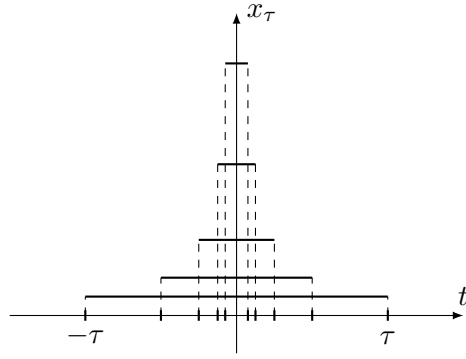
vemos que se verifican todas las propiedades del caso periódico. También para la transformada de Fourier, se verifica $\hat{z}(\nu) = \hat{x}(\nu)\hat{y}(\nu)$. Si para $\tau > 0$, tomamos la señal $x_\tau(t)$ dada

$$x_\tau(t) = \begin{cases} \tau^{-1} & t \in (-\tau/2, \tau/2), \\ 0 & t \notin (-\tau/2, \tau/2), \end{cases}$$

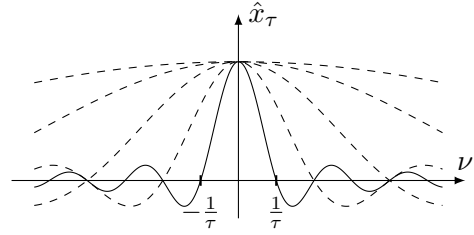
entonces podemos probar que $z(t) \xrightarrow{\tau \rightarrow 0} y(t)$. En efecto, la diferencia $z(t) - y(t)$ se puede escribir

$$|z(t) - y(t)| \leq \frac{1}{\tau} \int_{t-\tau/2}^{t+\tau/2} |y(t') - y(t)| dt' \leq \max_{|t'-t| \leq \tau/2} |y(t') - y(t)| \xrightarrow{\tau \rightarrow 0} 0.$$

En forma equivalente, por lo visto en el ejemplo 6.14, $x_\tau(\nu) = \text{sinc}(\pi\tau\nu) \xrightarrow{\tau \rightarrow 0} 1$, de donde se deduce $\hat{z}(\nu) \rightarrow \hat{y}(\nu)$. A este proceso límite se le puede asignar el objeto $\delta(t) = \lim_{\tau \rightarrow 0} x_\tau(t)$, que si bien no es una función, se puede operar como si lo fuera, se la denomina *función delta de Dirac*.¹⁴ Extendemos la noción de transformada a la función delta de Dirac tomando $\hat{\delta} = 1$, en la Figura 6.24 mostramos el paso al límite, en la Figura 6.24(a) se grafica $x_\tau(t)$ para $\tau \rightarrow 0$ y en 6.24(b) sus transformadas $\hat{x}(\nu)$.



(a) Aproximaciones a la delta de Dirac.



(b) Aproximaciones a la función 1.

Fig. 6.24: Función delta de Dirac y su transformada.

Si definimos $\delta_\tau(t) = \delta(t - \tau)$, vemos que $(\delta_\tau * x)(t) = x(t - \tau)$ y $\hat{\delta}_\tau(\nu) = e^{-i2\pi\tau\nu}$. Podemos ver

$$\begin{aligned} \psi(t) &= \delta_\tau(t), & \hat{\psi}(\nu) &= e^{-i2\pi\tau\nu}, \\ \psi(t) &= \frac{\delta_\tau(t) + \delta_{-\tau}(t)}{2}, & \hat{\psi}(\nu) &= \cos(2\pi\tau\nu), \\ \psi(t) &= \frac{\delta_\tau(t) - \delta_{-\tau}(t)}{i2}, & \hat{\psi}(\nu) &= -\text{sen}(2\pi\tau\nu). \end{aligned}$$

Usando la identidad (6.15) obtenemos

$$\begin{aligned} \psi(t) &= e^{i2\pi\eta t}, & \hat{\psi}(\nu) &= \delta_\eta(\nu), \\ \psi(t) &= \cos(2\pi\eta t), & \hat{\psi}(\nu) &= \frac{\delta_\eta(\nu) + \delta_{-\eta}(\nu)}{2}, \\ \psi(t) &= \text{sen}(2\pi\eta t), & \hat{\psi}(\nu) &= \frac{\delta_\eta(\nu) - \delta_{-\eta}(\nu)}{i2}. \end{aligned}$$

6.3.4. Teorema de muestreo. Las señales generadas por sistemas físicos, tiene espectro limitado. Por ejemplo, la voz humana se extiende desde 20 Hz hasta 20 kHz. En estos casos el denominado Teorema de muestreo nos dice que conociendo la señal a intervalos $\Delta t < \frac{1}{2\nu}$ donde ν es la máxima frecuencia, podemos recuperar la señal completa sin pérdida de información. En el ejemplo de la voz humana, tomando los valores de la señal a intervalos menores de 25 μs podríamos reproducir exactamente un sonido.

6.3.5. Criterio de Nyquist.

6.3.6. Aproximación por funciones sinc.

¹⁴Para darle sentido la delta de Dirac hay que considerar la teoría de distribuciones.

6.4. Filtros.

6.4.1. Filtros pasa bajos, pasa banda y pasa altos.

6.4.2. Filtros en el dominio de la frecuencia.

6.4.3. Filtros en el dominio del tiempo.

6.4.4. Respuesta unitaria.

6.4.5. Función de transferencia.

6.5. Aplicación.

6.5.1. Detección de ondas Alfa y Beta en señales EEG.

APÉNDICE A

Conceptos Básicos

“Varios gauchos en la pulpería conversan sobre temas de escritura y de fonética. El santiagueño Albarracín no sabe leer ni escribir, pero supone que la palabra trara no puede escribirse. Crisanto Cabrera, también analfabeto, sostiene que todo lo que se habla puede ser escrito. -Pago la copa para todos -le dice el santiagueño- si escribe trara. -Se la juego -contesta Cabrera; saca el cuchillo y con la punta traza unos garabatos en el piso de tierra. De atrás se asoma el viejo Álvarez, mira el suelo y sentencia: -Clarito, trara.”

Jorge Luis Borges y Adolfo Bioy Casares

A.1. Trigonometría y números complejos.

A.1.1. Identidades trigonométricas. Se verifican

$$\begin{aligned}\cos(\alpha \pm \beta) &= \cos(\alpha) \cos(\beta) \mp \sin(\alpha) \sin(\beta), \\ \sin(\alpha \pm \beta) &= \cos(\alpha) \sin(\beta) \pm \sin(\alpha) \cos(\beta).\end{aligned}$$

Sumando y restando las identidades anteriores, vemos que valen

$$\begin{aligned}2 \cos(\alpha) \cos(\beta) &= \cos(\alpha - \beta) + \cos(\alpha + \beta), \\ 2 \sin(\alpha) \sin(\beta) &= \cos(\alpha - \beta) - \cos(\alpha + \beta), \\ 2 \cos(\alpha) \sin(\beta) &= \sin(\alpha + \beta) - \sin(\alpha - \beta).\end{aligned}$$

Tomando $\alpha = \beta$ obtenemos

$$\begin{aligned}\cos^2(\alpha) - \sin^2(\alpha) &= \cos(2\alpha), \\ 2 \cos(\alpha) \sin(\alpha) &= \sin(2\alpha), \\ 2 \cos^2(\alpha) &= 1 + \cos(2\alpha), \\ 2 \sin^2(\alpha) &= 1 - \cos(2\alpha).\end{aligned}$$

Mediante un cálculo directo, vemos que para $k \in \mathbb{N}$

$$\begin{aligned}\int_0^1 \cos(2k\pi t) dt &= \frac{1}{2k\pi} \sin(2k\pi t) \Big|_{t=0}^{t=1} = 0, \\ \int_0^1 \sin(2k\pi t) dt &= -\frac{1}{2k\pi} \cos(2k\pi t) \Big|_{t=0}^{t=1} = 0,\end{aligned}$$

resulta entonces que para $k, m \in \mathbb{N}$ con $k \neq m$

$$\begin{aligned}\int_0^1 \cos(2k\pi t) \cos(2m\pi t) dt &= \frac{1}{2} \int_0^1 \cos(2(k-m)\pi t) + \cos(2(k+m)\pi t) dt = 0, \\ \int_0^1 \sin(2k\pi t) \sin(2m\pi t) dt &= \frac{1}{2} \int_0^1 \cos(2(k-m)\pi t) - \cos(2(k+m)\pi t) dt = 0, \\ \int_0^1 \sin(2k\pi t) \cos(2m\pi t) dt &= \frac{1}{2} \int_0^1 \sin(2(k+m)\pi t) - \sin(2(k-m)\pi t) dt = 0,\end{aligned}$$

y en el caso $k = m$ se verifica

$$\begin{aligned}\int_0^1 \cos^2(2k\pi t) dt &= \frac{1}{2} \int_0^1 1 + \cos(4k\pi t) dt = \frac{1}{2}, \\ \int_0^1 \sin^2(2k\pi t) dt &= \frac{1}{2} \int_0^1 1 - \cos(4k\pi t) dt = \frac{1}{2}, \\ \int_0^1 \sin(2k\pi t) \cos(2k\pi t) dt &= \frac{1}{2} \int_0^1 \sin(4k\pi t) dt = 0,\end{aligned}$$

Podemos resumir las relaciones anteriores

$$\begin{aligned}\int_0^1 \cos(2k\pi t) \cos(2m\pi t) dt &= \frac{1}{2} \delta_{k,m}, \\ \int_0^1 \sin(2k\pi t) \sin(2m\pi t) dt &= \frac{1}{2} \delta_{k,m}, \\ \int_0^1 \sin(2k\pi t) \cos(2m\pi t) dt &= 0.\end{aligned}$$

donde $\delta_{k,k} = 1$, $\delta_{k,m} = 0$ si $k \neq m$.

A.2. Números complejos.

A.2.1. Exponenciales complejas. Queremos darle sentido a la expresión e^z , donde $z \in \mathbb{C}$. Vamos a asumir que la función exponencial mantiene algunas propiedades del caso real, por ejemplo

- $e^0 = 1$.
- $e^{x_1+x_2} = e^{x_1} e^{x_2}$.

Entonces, si $z = x + iy$ tenemos $e^z = e^x e^{iy}$. Por lo tanto, vamos a centrarnos en exponenciales imaginarias. Consideremos el desarrollo en polinomios de Taylor de la función exponencial

$$e^x \cong 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots + \frac{x^n}{n!},$$

reemplazando $x \rightarrow iy$, obtenemos

$$\begin{aligned}e^{iy} &\cong 1 + iy - \frac{y^2}{2} - i\frac{y^3}{6} + \frac{y^4}{24} + \cdots + \frac{(iy)^n}{n!} \\ &= (1 - \frac{y^2}{2} + \frac{y^4}{24} - \cdots) + i(y - \frac{y^3}{6} + \cdots) \cong \cos(y) + i \sin(y).\end{aligned}$$

En función de lo anterior, definimos $e^{iy} = \cos(y) + i \operatorname{sen}(y)$ y $e^{x+iy} = e^x(\cos(y) + i \operatorname{sen}(y))$. Tenemos que $e^0 = e^0(\cos(0) + i \operatorname{sen}(0)) = 1$, si $z = (x_1 + x_2) + i(y_1 + y_2)$

$$\begin{aligned} e^z &= e^{x_1+x_2}(\cos(y_1 + y_2) + i \operatorname{sen}(y_1 + y_2)) \\ &= e^{x_1}e^{x_2}(\cos(y_1)\cos(y_2) - \operatorname{sen}(y_1)\operatorname{sen}(y_2) + i \operatorname{sen}(y_1)\cos(y_2) + i \cos(y_1)\operatorname{sen}(y_2)) \\ &= e^{x_1}(\cos(y_1) + i \operatorname{sen}(y_1))e^{x_2}(\cos(y_2) + i \operatorname{sen}(y_2)) = e^{x_1+iy_1}e^{x_2+iy_2} \end{aligned}$$

Observemos que $|e^z| = e^x$ y para $t \in \mathbb{R}$, $e^{i2\pi t} = \cos(2\pi t) + i \operatorname{sen}(2\pi t)$ es un punto del plano complejo sobre la circunferencia unitaria con centro en el origen. Además vale

$$\frac{d}{dt}e^{i2\pi t} = -2\pi \operatorname{sen}(2\pi t) + i2\pi \cos(2\pi t) = i2\pi e^{i2\pi t},$$

que coincide con lo esperado.

Identidad de Euler Tomando $t = 1/2$, obtenemos $e^{i\pi} = \cos(\pi) + i \operatorname{sen}(\pi) = -1$, que es equivalente a la identidad de Euler $e^{i\pi} + 1 = 0$, que relaciona los cinco números más importantes de la matemática.

Las raíces n -ésimas de la unidad, son los complejos que verifican la ecuación $z^n - 1 = 0$. Son exactamente n números distintos (raíces simples) y se escriben $z_k = e^{i2\pi k/n}$, con $k = 1, \dots, n$. En efecto, $z_k^n = e^{i2\pi k} = 1$.

A.3. Serie geométrica. Dado un número complejo $z \in \mathbb{C}$, vamos a considerar las sumas de las sucesivas potencias

$$\sigma_n(z) = 1 + z + z^2 + \dots + z^{n-1} = \sum_{k=0}^{n-1} z^k,$$

vemos que $\sigma_n(0) = 1$, $\sigma_n(1) = n$. Mostraremos que $\sigma_n(z) = (1 - z^n)/(1 - z)^1$, si $z \neq 1$. En efecto,

$$\begin{aligned} (1 - z)\sigma_n(z) &= 1 + z + z^2 + \dots + z^{n-1} - z(1 + z + z^2 + \dots + z^{n-1}) \\ &= 1 + z + z^2 + \dots + z^{n-1} - (z + z^2 + \dots + z^n) = 1 - z^n, \end{aligned}$$

notemos que $\lim_{z \rightarrow 1} (1 - z^n)/(1 - z) = n$

A.4. Delta de Kronecker. La delta de Kronecker² es la función definida en dos variables enteras que toma valores 0, 1. En forma más precisa, se define $\delta : \mathbb{Z} \times \mathbb{Z} \rightarrow \{0, 1\}$ como

$$\delta_{j,k} = \begin{cases} 1 & , \text{ si } j = k, \\ 0 & , \text{ si } j \neq k \end{cases}$$

La matriz definida por $I = (\delta_{j,k})_{1 \leq j,k \leq n}$ es la matriz identidad, es decir $IA = AI = A$ para toda $A \in \mathbb{R}^{n \times n}$.

A.5. Notación de Landau. Dadas dos funciones $f(x), g(x)$ definidas en el conjunto $I \subset \mathbb{R}$, la notación $f(x) = O(g(x))$ significa que existe una constante $C > 0$ tal que

$$(A.1) \quad |f(x)| \leq C|g(x)|, \quad \text{para todo } x \in I.$$

Por ejemplo, $\operatorname{sen}(x) = O(x)$ para $x \in \mathbb{R}$. En general nos va interesar el comportamiento de estas funciones cerca de un punto x_0 , decimos que $f(x) = O(g(x))$ cuando $x \rightarrow x_0$ si existe $\delta > 0$ y $C = C(\delta) > 0$ tal que se verifica (A.1) para $I = \{x \in \mathbb{R} : 0 < |x - x_0| < \delta\}$. Para $x \rightarrow x_0^+$, tomamos $I = (x_0, x_0 + \delta)$, y en el caso $x \rightarrow \infty$ consideramos intervalos de la forma (a, ∞) . Una función acotada se puede escribir como $f(x) = O(1)$.

¹A esta identidad se la suele denominar *sexto caso de factorio*.

²En honor al matemático Leopold Kronecker.

A.6. Operaciones aritméticas.

A.6.1. Error relativo y dígitos significativos. Supongamos que calculamos aproximadamente un número real y por un valor aproximado \hat{y} . Una forma de medir la precisión de nuestro cálculo es considerar la distancia entre ambos valores, llamaremos a esta distancia error absoluto, $E_{\text{abs}} = |y - \hat{y}|$. Más útil es en general estudiar el error relativo definido como

$$E_{\text{rel}} = \frac{|y - \hat{y}|}{|y|},$$

asumiendo $y \neq 0$. Puede parecer que las definiciones anteriores carecen de cualquier sentido práctico, para obtener los errores necesitamos conocer y . Pero si lo conociéramos, no habría necesidad de calcularlo. Más adelante vamos a ver el sentido de la definición de error. En muchos problemas aplicados no conocemos el valor del error, pero tenemos buenas estimaciones que nos permiten dar una intervalo donde se encuentra y .

Nota: De alguna manera, el error relativo es el criterio de precisión que aplicamos en general. Si en una carpintería alguien cortara una tabla con un centímetro de error respecto a lo previsto, seguro que debería rehacer el trabajo. Por otro lado, cuando la sonda espacial Giotto en 1986 llegó a estar a una distancia un poco menor que 600 km del cometa Halley, la misión fue calificada como un éxito. Es fácil entender esta diferencia en términos de error relativo, mientras que un centímetro en un metro representa un error de 1 %, una distancia de 600 km es el 0.001 % de la distancia más cercana del cometa a la tierra (58×10^6 km).

Es común hablar de dígitos significativos correctos como equivalente a error relativo. Si bien es razonable, es necesario hacer algunas observaciones. Empecemos por aclarar que queremos decir con dígitos significativos: es el número de dígitos a la derecha del primero no nulo contando también a este. Por ejemplo, $y = 32.17$ tiene cuatro dígitos significativos, igualmente $y = 0.00003217$. Por otro lado, $y = 32.1700$ tiene seis dígitos significativos, los últimos dos ceros dan idea de la precisión con que el número y fue calculado o medido, es decir $|y - 32.1700| < 5.0 \times 10^{-5}$. Como dijimos, el concepto de error relativo y dígitos significativos correctos están vinculados, pero el primero encierra más información. Por ejemplo, si $y = 0.13002$, el valor $\hat{y} = 0.12999$ es una muy buena aproximación dado que el error relativo es $E_{\text{rel}} = 0.00023$, mientras que tenemos un solo dígito significativo correcto. En [11] se da la siguiente definición de dígitos significativos correctos: decimos que \hat{y} es una aproximación con k dígitos significativos correctos si coinciden al considerar el redondeo a k dígitos. En el ejemplo anterior, si tomamos el redondeo de ambos a 4 dígitos obtenemos $\text{round}_4(y) = \text{round}_4(\hat{y}) = 0.1300$. Pero existen estas anomalías, si $y = 0.19949$ y $\hat{y} = 0.19951$ obtenemos

$$\begin{aligned}\text{round}_4(y) &= \text{round}_4(\hat{y}) = 0.1995, \\ \text{round}_3(y) &= 0.199 \neq \text{round}_3(\hat{y}) = 0.200, \\ \text{round}_2(y) &= \text{round}_2(\hat{y}) = 0.20,\end{aligned}$$

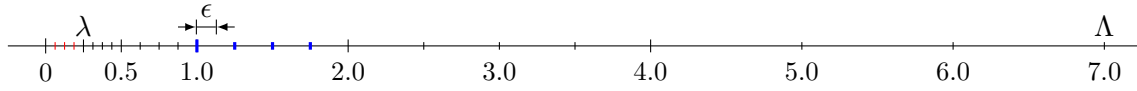
podríamos definir entonces la cantidad de dígitos significativos correctos como el máximo entero k tal que $\text{round}_k(y) = \text{round}_k(\hat{y})$. De cualquier forma, el error relativo tiene la ventaja de no depender de la base en la que se representan los números.

A.6.2. Representación en punto flotante. La representación de punto flotante (floating point en inglés) es la forma de notación científica que se usa en las computadoras digitales para representar números reales y con la que se realizan las operaciones aritméticas. En base 2 (la que usan todas las computadoras actuales), definimos el conjunto \mathbf{F}_+ formado por los números $x = m \times 2^{e-t}$, con m, e enteros, $m \in [2^{t-1}, 2^t - 1]$, llamada mantisa y el exponente $e \in [e_{\min}, e_{\max}]$. El conjunto $\mathbf{F} = \mathbf{F}_+ \cup (-\mathbf{F}_+) \cup \{0\}$ es el conjunto de los números de punto flotante normalizados, que queda determinado por los enteros t, e_{\min}, e_{\max} .

		mantisa			
		4	5	6	7
exponente	-1	0.25	0.3125	0.375	0.4375
	0	0.5	0.625	0.75	0.875
	1	1.0	1.25	1.5	1.75
	2	2.0	2.5	3.0	3.5
	3	4.0	5.0	6.0	7.0

Tabla A.1: Números de punto flotante con $t = 3$, $e_{\min} = -1$ y $e_{\max} = 3$.

Ejemplo A.1. Si consideramos $t = 3$, $e_{\min} = -1$ y $e_{\max} = 3$, entonces $m = 4, \dots, 7$ y los números positivos representables se muestran en la Tabla A.1. En la Figura A.1 mostramos \mathbf{F}_+ en la recta real. Se puede observar la distribución no uniforme de los mismos, de hecho la distancia entre números consecutivos se duplica en cada potencia de 2. En rojo marcamos los números no normalizados, que corresponden a $m \in (0, 2^{t-1})$ y $e = e_{\min}$.

Fig. A.1: Conjunto \mathbf{F}_+ correspondiente al Ejemplo A.1.

El menor de los números normalizados positivos es $\lambda = 2^{e_{\min}-1}$ y el mayor es $\Lambda = 2^{e_{\max}}(1-\epsilon)$, donde $\epsilon = 2^{-t}$. El número 1 se representa con $m = 2^{t-1}$ y $e = 1$, el número consecutivo es $x = (2^{t-1} + 1) \times 2^{1-t} = 1 + 2\epsilon$. En el ejemplo anterior, $\lambda = 0.25$, $\Lambda = 7$ y $\epsilon = 0.125$. En la Figura A.1, se muestran estos números sobre la recta real.

Sea $\mathbf{F}_+ = \{m \times 2^{e-t} : m, e \in \mathbb{N}, 2^{t-1} \leq m \leq 2^t - 1, e_{\min} \leq e \leq e_{\max}\}$, definimos la función $fl : [\lambda, \Lambda] \rightarrow \mathbf{F}_+$ como el elemento de \mathbf{F}_+ donde se realiza la mínima distancia a x , si se alcanza en dos puntos distintos, definimos $fl(x)$ como el mayor de ambos. Por ejemplo, si $x = 1.125$, $|1.0 - 1.125| = |1.25 - 1.125| = 0.125$, por lo tanto $fl(1.125) = 1.25$.

Proposición A.1. Para todo $x \in [\lambda, \Lambda]$, existe δ con $|\delta| \leq \epsilon/(1+\epsilon)$, tal que $fl(x) - x = \delta x$.

Por ejemplo, si $x = 1.125$, $fl(1.125) - 1.125 = 0.125 = \delta \times 1.125$, con $\delta = 0.1\hat{1} = \epsilon/(1+\epsilon)$; para $x = 0.9375$, $fl(0.9375) - 0.9375 = 0.0625 = \delta \times 0.9375$, con $\delta = 0.0\hat{6} = \epsilon/(2-\epsilon)$.

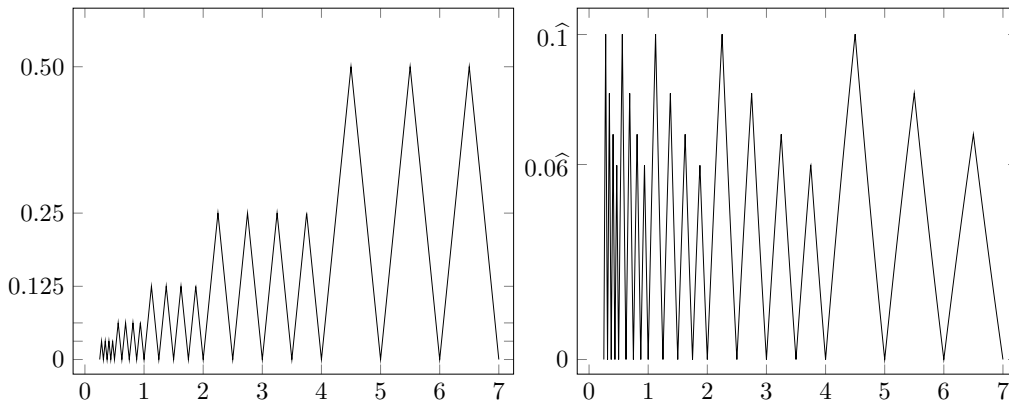
(a) Error absoluto: $E_{\text{abs}} = |fl(x) - x|$. (b) Error relativo: $E_{\text{rel}} = |fl(x) - x|/|x|$.

Fig. A.2: .

Bibliografía

- [1] D. Adams, *El Restaurante del Fin del Mundo*, Editorial Anagrama, 1984.
- [2] R. Arlt, *Los Lanzallamas*, CreateSpace Independent Publishing Platform, 2017.
- [3] B. Bereiter, S. Eggleston, J. Schmitt, C. Nehrbass-Ahles, T. F. Stocker, H. Fischer, S. Kipfstuhl, and J. Chappellaz, *Revision of the EPICA Dome C CO₂ record from 800 to 600 kyr before present*, Geophysical Research Letters **42** (2014), no. 2, 542–549.
- [4] J. L. Borges, *El Idioma Analítico de John Wilkins. Otras inquisiciones*, Emecé, 1964.
- [5] J. W. Cooley and J. W. Tukey, *An algorithm for the machine calculation of complex Fourier series*, Math. Comp. **19** (1965), 297–301.
- [6] H. Frings and M. Frings, *The effects of temperature on chirp-rate of male cone-headed grasshoppers, neoconocephalus ensiger*, Journal of Experimental Zoology **134** (1957), no. 3, 411–425.
- [7] D. Goldberg, *What every computer scientist should know about floating point arithmetic*, ACM Computing Surveys **23** (1991), no. 1, 5–48.
- [8] R. Gul and S. Bernhard, *Local sensitivity analysis of cardiovascular system parameters*, vol. 17, 04 2013.
- [9] C. G. Hewitt, *The Conservation of the Wild Life of Canada*, New York : C. Scribner's Sons, 1921, <https://www.biodiversitylibrary.org/bibliography/25399>.
- [10] M. Hierro Franco and M. Guijarro Garvi, *Un estudio mediante cadenas de Markov de la dinámica de los movimientos migratorios interterritoriales en España (1990-2003) desde un planteamiento de estimación dinámico*, Revista Asturiana de Economía **35** (2006), 145–161.
- [11] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, second ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002.
- [12] M. Kline, *El Pensamiento Matemático de la Antigüedad a Nuestros Días*, Alianza Universidad, no. v. 2, Alianza, 1994.
- [13] R. H. Landau and M. J. P. Mejía, *Computational Physics: problem solving with computers*, Physics textbook, no. v. 1, Wiley, 1997.
- [14] E. N. Lorenz, *Deterministic nonperiodic flow*, Journal of the Atmospheric Sciences **20** (1963), no. 2, 130–141.
- [15] J. D. Murray, *Mathematical Biology I. An introduction*, 3 ed., Interdisciplinary Applied Mathematics, vol. 17, Springer, New York, 2002.

- [16] J. D. Murray, *Mathematical Biology II : Spatial models and biomedical applications* , third edition, Interdisciplinary Applied Mathematics, Springer, 2005.
- [17] R. Palaniappan, *Biological Signal Analysis*, BookBoon, 2011.
- [18] D. Pauly, *Fish population dynamics in tropical waters: a manual for use with programmable calculators*, vol. 8, WorldFish, 1984.
- [19] J. G. Roederer, *Electromagnetismo Elemental*, Eudeba, 2015.
- [20] J. G. Roederer and G. D. Pozzati, *Acústica y Psicoacústica de la Música*, Melos, Buenos Aires, 2007.
- [21] H. Rossi, *Mathematics is an edifice, not a toolbox*, Notices Amer. Math. Soc. **43** (1996), no. 10.
- [22] E. Schrödinger, *¿Qué es la Vida?*, Metatemas, TusQuest, 2015.
- [23] W. B. Streett, *Pressure-volume-temperature data for neon from 80-130.deg.k and pressures to 2000 atmospheres*, Journal of Chemical & Engineering Data **16** (1971), no. 3, 289–292.
- [24] L. Von Bertalanffy, *A quantitative theory of organic growth (inquiries on growth laws. ii)*, Human biology **10** (1938), no. 2, 181–213.
- [25] N. Wiener, *Cibernética*, Metatemas, TusQuets, 1998.