Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Matemáticas

La aproximación óptima polinomial en la norma L1: Existencia, unicidad y algoritmos

Autor: Miguel Sanz Sanz
Tutor: Luis María Abia Llera

2023-2024

Índice general

Prefacio

Esta memoria se centra en resultados fundamentales de la teoría de aproximación óptima en la norma L_1 (o en media) de funciones continuas por elementos de un subespacio de dimensión finita; en particular, por funciones polinómicas. En contraste con la aproximación óptima mínimos cuadrados y la aproximación óptima en la norma de Chebyshev, la aproximación óptima en L_1 tiene poca presencia en los textos de la teoría general de la aproximación. Una referencia clásica es el texto de Watson [**6**], que hemos seguido para presentar resultados generales en el capítulo 1 y para el desarrollo de la solución L_1 de sistemas lineales sobredeterminados en el capítulo 2. Aunque la solución L_1 de un sistema sobredeterminado se puede formular como un problema de programación lineal, hemos optado por desarrollar un algoritmo de descenso directo, cuyos elementos se estudian también en este capítulo. Hemos implementado una función Matlab de este algoritmo con la que presentamos una experimentación numérica en el capítulo 4.

El capítulo 3 se centra en la aproximación L_1 de funciones continuas en un intervalo compacto. Aunque la teoría considera subespacios generales de dimensión finita que satisfacen la condicion de Haar, la situación predominante es la aproximación por polinomios. Para los resultados de existencia, unicidad y caracterización de la solución hemos seguido los textos de Powell [**5**] y Cheney [**3**]. Algoritmos para la construcción efectiva de la aproximación óptima L_1 de una función continua no están tan desarrollados como para el caso de la aproximación Chebyshev. Es por ello que el capítulo 3 aborda la discretización en una malla de puntos del problema de aproximación L_1. Cuando se discretiza el problema, se recae en la solución L_1 de un sistema lineal sobredeterminado. El capítulo 3 ilustra el buen funcionamiento de las aproximaciones discretas así construidas.

Capítulo 1

Introducción

La aproximación óptima polinomial en la norma L_1 es un tema de gran relevancia en el ámbito de las matemáticas y la optimización. Este enfoque tiene una aplicación significativa en la resolución de problemas matemáticos fundamentales, que involucran la minimización de errores o la aproximación de funciones. En este Trabajo de Fin de Grado (TFG), exploraremos los aspectos teóricos y prácticos de la aproximación óptima polinomial en la norma L_1, con un enfoque específico en la existencia, unicidad y los algoritmos asociados.

La norma L_1, también conocida como la norma del valor absoluto, es una métrica de distancia que se utiliza para medir la magnitud de un vector o la diferencia entre dos funciones. A diferencia de la norma L_2 (norma euclidiana), que utiliza el cuadrado de los valores absolutos, la norma L_1 se basa en la suma de los valores absolutos. Esto tiene implicaciones importantes en la optimización y la aproximación, ya que a menudo refleja mejor la estructura y propiedades de ciertos problemas matemáticos.

En este TFG, nos centraremos en tres aspectos clave:

1. Existencia: Investigaremos las condiciones bajo las cuales podemos garantizar la existencia de un polinomio que minimiza la distancia L_1 entre una función dada y su aproximación. Formularemos teoremas y resultados que caracterizan la aproximación óptima en la norma L_1 y establecen la existencia de dichas aproximaciones en diversos contextos y problemas.

2. Unicidad: La unicidad es otro aspecto crucial de la aproximación óptima en la norma L_1. A diferencia de la norma L_2, la norma L_1 no es estrictamente convexa y es necesario imponer condiciones adicionales para obtener la unicidad de las aproximaciones óptimas. Examinaremos bajo qué condiciones el polinomio que minimiza la distancia L_1 es único, lo que nos ayudará a comprender cuándo podemos estar seguros de que la aproximación obtenida es la mejor posible en términos de la norma L_1.

3. Algoritmos: Además de estudiar la teoría subyacente, nos sumergiremos en la práctica revisando la tipología de los algoritmos utilizados para encontrar la aproximación óptima en la norma L_1. Examinaremos métodos numéricos y algoritmos eficientes que se aplican en diversos contextos matemáticos y que juegan un papel crucial en la optimización.

En resumen, este TFG se enfoca en un tema matemático fundamental que tiene aplicaciones significativas en la optimización, la aproximación de funciones y el análisis estadístico de datos [**2**]. Con un enfoque en la existencia, unicidad y algoritmos, esperamos proporcionar una comprensión sólida de la aproximación óptima polinomial en la norma L_1 y su relevancia en el ámbito matemático.

1.1. Concepto de aproximación óptima

Todos los problemas de aproximación son casos especiales de la siguiente formulación abstracta general:

Se tiene un espacio métrico S, un subconjunto $M \subseteq S$ y un elemento $g \in S$. Se trata entonces de encontrar un elemento de M que esté a distancia mínima de g. En términos de la distancia del espacio métrico S, se formula

$$\text{Encontrar } f \in M \quad \text{tal que } d(f,g) = d(f,M).$$

El caso en que S es un espacio vectorial normado es de gran importancia y vamos a limitarnos a él.

La condición de compacidad, que implica que todas las sucesiones en M tienen subsucesiones convergentes, resulta ser la propiedad mínima que garantizará la existencia de una mejor aproximación.

Teorema 1.1.1. *Sea M un conjunto compacto en un espacio vectorial normado. Entonces, a cada punto g del espacio le corresponde un punto en M que es el más cercano a g.*

Demostración. Sea $\delta = \text{ínf}\{\|g-x\|, \quad x \in M\}$. Por la definición de ínfimo existe una sucesión de puntos $x_1, x_2, ...$ en M tal que

$$\|g - x_n\| \to \delta \quad n \to \infty$$

Como M es compacto, implica que existe una subsucesión $x_1, x_2, ...$ que converge a $x^* \in M$. Entonces

$$\|g - x^*\| \leq \|g - x_n\| + \|x_n - x^*\|$$

y tomando $n \to \infty$, $\|g - x^*\| \leq \delta$ el lado izquierdo de la desigualdad no depende de n. Entonces, como $x^* \in M$,

$$\|g - x^*\| \geq \delta$$

y entonces $\|g - x^*\| = \delta$ y x^* es el punto de mínima distancia desde g a M. $\square$

Lema 1.1.1. *Todo conjunto cerrado, acotado y de dimensión finita en un espacio normado es compacto.*

Demostración. Sea M un conjunto de dimensión n. Sea un conjunto linealmente independiente $\{m_1, m_2, ..., m_n\}$ tal que todo elemento de M debe estar escrito únicamente como

$$m = \sum_{i=1}^{n} \lambda_i m_i.$$

Consideremos $\mathbb{R}^n$ dotado con la norma L_∞ y sea T la aplicación de $\lambda \in \mathbb{R}^n$ a $m \in M$. Entonces

$$\begin{aligned} \|T\lambda - T\mu\| &= \|\sum_{i=1}^{n} \lambda_i m_i - \sum_{i=1}^{n} \mu_i m_i\| \\ &= \|\sum_{i=1}^{n} (\lambda_i - \mu_i) m_i\| \\ &\leq \sum_{i=1}^{n} |\lambda_i - \mu_i| \|m_i\| \\ &\leq \|\lambda - \mu\|_\infty \sum_{i=1}^{n} \|m_i\| \end{aligned}$$

por lo tanto T es continua. Ahora tomamos $X = \{\lambda : T\lambda \in M\}$. La compacidad de M resulta de la compacidad de X, ya que una aplicación continua de un espacio métrico en otro conserva la compacidad. Por tanto necesitamos ver que X es cerrado y acotado. Para ver que X es cerrado, tomamos $\lambda^{(k)} \in X, \lambda^{(k)} \to \lambda$. Entonces por la continuidad

$$T\lambda = T\left(\lim_{k\to\infty} \lambda^{(k)}\right) = \lim_{k\to\infty} T(\lambda^{(k)}).$$

Por ser M cerrado, $T\lambda \in M$ y $\lambda \in X$. Por tanto X es cerrado.
Ahora tomamos el conjunto $\{\lambda : \|\lambda\|_\infty = 1\}$ que es compacto, y la continuidad de T

$$\alpha = \inf_{\|\lambda\|_\alpha = 1} \|T\lambda\| = \inf_{\|\lambda\|_\alpha = 1} \|\sum_{i=1}^{n} \lambda_i m_i\|$$

se consigue. Como las funciones $\{m_i\}$ son linealmente independientes, tenemos que $\alpha > 0$. Por tanto para algún $\lambda \neq 0$,

$$\|T\lambda\| = \|T(\lambda/\|\lambda\|_\infty)\|\lambda\|_\infty\| \geq \alpha\|\lambda\|_\infty.$$

Entonces $\|T\lambda\|$ es acotado para $\lambda \in X$ porque M lo es, y $\|\lambda\|_\infty$ es acotado por $\lambda \in X$. Por tanto X es acotado.

□

Teorema 1.1.2. *Sea M un subespacio de dimensión finita de un espacio lineal normado S. Entonces existe una mejor aproximación en M de cualquier punto de S.*

Demostración. Sea M un subespacio de dimensión finita de un espacio lineal normado S y sea g un punto de S. Entonces si m' es un punto de M, el punto buscado se encuentra en el conjunto

$$\{m : m \in M, \|m - g\| \leq \|m' - g\|\}. \tag{1.1}$$

Este conjunto es cerrado y acotado y compacto por el Lema 1.1.1. Por tanto por el Teorema 1.1.1 tenemos que existe una mejor aproximación.

□

Ya hemos visto entonces la existencia de una mejor aproximación, ahora vamos a centrarnos en la unicidad de dichas mejores aproximaciones. Para ello vamos a comenzar introduciendo el concepto de convexidad.

Definición 1.1.1. *Un conjunto M de un espacio vectorial S es convexo si dados $x, y \in M$ se cumple que*

$$\lambda x + (1 - \lambda) \cdot y \in M, \qquad \forall \lambda \quad tal\ que \quad 0 \leq \lambda \leq 1$$

Definición 1.1.2. *Un conjunto M de un espacio vectorial S es estríctamente convexo si dados $x, y \in M$ y $a \in S$ que se encuentran en la frontera de la esfera cerrada de radio r, se cumple que*

$$\|\lambda x + (1 - \lambda)y - a\| < r, \qquad 0 < \lambda < 1 \tag{1.2}$$

Teorema 1.1.3. *En un espacio vectorial normado S estríctamente convexo, un subespacio de dimensión finita M tiene una única mejor aproximación para cualquier punto $g \in S$.*

Demostración. Podemos asegurar que al menos existe una mejor aproximación gracias al Teorema 1.1.2. Sea m_1 y m_2 elementos de M de mínima distancia desde g con distancia r. Entonces para un $\lambda \in \mathbb{R}$ tal que $0 \leq \lambda \leq 1$,

$$\begin{aligned} \|\lambda m_1 + (1 - \lambda)m_2 - g\| &\leq \|\lambda(m_1) - g\| + \|(1 - \lambda)(m_2 - g)\| \\ &\leq r \end{aligned}$$

y si

$$\|m_1 - g\| = \|m_2 - g\| = \|\lambda m_1 + (1 - \lambda)m_2 - g\|$$

entonces contradice a (1.2) a no ser que $m_1 = m_2$.

□

Teorema 1.1.4. *Sea D un subconjunto cerrado y convexo de $\mathbb{R}^n$. Entonces D no contiene el origen si y solo si existe $z \in \mathbb{R}^n$ tal que*

$$d^T z > 0, \qquad \forall d \in D$$

Demostración. Asumimos que $\mathbf{0} \notin D$ y consideramos el problema

$$\text{encontrar } \mathbf{d} \in D \text{ para minimizar } \|\mathbf{d}\|_2.$$

Debemos buscar aproximaciones a partir del conjunto

$$\{d \in D : \|\mathbf{d}\|_2 \leq \|\hat{\mathbf{d}}\|_2\}$$

donde $\mathbf{d} \in D$ es arbitrario y, dado que este conjunto es compacto, se garantiza la existencia de un punto $\mathbf{z} \in D$ en el cual se alcanza el mínimo. Ahora, si $\mathbf{d} \in D$, debido a la convexidad de D

$$\gamma\mathbf{d} + (1-\gamma)\mathbf{z} \in D, \qquad 0 \leq \gamma \leq 1$$

así pues

$$\begin{aligned} 0 &\leq \|\gamma\mathbf{d} + (1-\gamma)\mathbf{z}\|_2^2 - \|\mathbf{z}\|_2^2 \\ &= \gamma^2\|\mathbf{d}-\mathbf{z}\|_2^2 + 2\gamma(\mathbf{d}-\mathbf{z})^T\mathbf{z}. \end{aligned}$$

esta desigualdad no puede ser válida para valores pequeños de γ a menos que

$$(\mathbf{d}-\mathbf{z})^T \geq 0$$

y por lo tanto

$$\mathbf{d}^T\mathbf{z} \geq \mathbf{z}^T\mathbf{z} > 0$$

lo cual, dado que $\mathbf{d}$ es arbitrario, proporciona la conclusión requerida. Ahora, asumamos que existe un valor $\mathbf{z} \in \mathbb{R}^n$ tal que

$$\mathbf{d}^T\mathbf{z} > 0, \qquad \forall \mathbf{d} \in D$$

si D contiene el origen, es imposible.

□

1.2. Caracterización de aproximaciones óptimas

En esta sección vamos a presentar un resultado de caracterización general para una clase importante de problemas de aproximación lineal, que incluye todos los problemas que se pueden esperar en la práctica y aquellos que se tratan en detalle más adelante.

El requisito básico aquí es una función (posiblemente no lineal) $f(\mathbf{a})$ que va desde $\mathbb{R}^n$ a un subconjunto M de un espacio vectorial normado arbitrario S. Entonces, sin pérdida de generalidad, el problema de encontrar una mejor aproximación desde M hacia un elemento de S se puede expresar como:

$$\text{Minimizar } \|f(\mathbf{a})\|, \qquad \text{tal que } \mathbf{a} \in \mathbb{R}^n \tag{1.3}$$

Antes de formular el teorema necesitamos introducir algunos conceptos fundamentales sobre el espacio dual de un espacio normado.
Sea S^* el espacio dual de S, es decir, el conjunto de funcionales lineales continuos $v(f)$ definidos en S. Para mayor comodidad, escribiremos

$$v(f) = \langle f, v\rangle,$$

por lo tanto, definiendo la funcional lineal como un producto interno entre los elementos de S y los de S^*. La norma dual en S^* viene dada por

$$\|v\|^* = \sup_{\|f\|\leq 1} \langle f, v\rangle.$$

Ahora definimos, dado $f \in S$, el conjunto $V(f)$ por

$$V(f) = \{v \in S^* : \|f\| = \langle f, v\rangle, \|v\|^* \leq 1\}. \tag{1.4}$$

Observamos que $V(f)$ es un subconjunto convexo de S^*. Entonces, asumiendo que la derivada parcial de f con respecto a a_j existe y se denota como $g_j(\mathbf{a}), j = 1, 2, ..., n,$, podemos dar la siguiente condición necesaria para que $\mathbf{a}$ sea solución de (1.3).

Teorema 1.2.1. *Sea $\boldsymbol{a} \in \mathbb{R}^n$ que resuelve (1.3) y supongamos que existe un entorno de $\boldsymbol{a}$ en el que podemos escribir*

$$f(\boldsymbol{a} + z) = f(\boldsymbol{a}) + \sum_{j=1}^{n} z_j g_j(\boldsymbol{a}) + o\left(\|z\|_n\right) \tag{1.5}$$

donde $\|\cdot\|_n$ representa cualquier norma en $\mathbb{R}^n$. Entonces existe $v \in V\left(f(\boldsymbol{a})\right)$ tal que

$$\langle g_j, v\rangle = 0 \quad j = 1, 2, ..., n. \tag{1.6}$$

Demostración. Supongamos que $\mathbf{a}$ resuelve (1.3) y que se cumple (1.5) pero no se cumple (1.6). El conjunto $D \subset \mathbb{R}^m$, definido por

$$D = \{\mathbf{d} : d_j = \langle g_j, v\rangle / \, j = 1, 2, ..., n, \ v \in V\left(f(\mathbf{a})\right)\}.$$

V es un subconjunto convexo ya que dados $v_1, v_2 \in V(f)$. Tenemos que:

$$\begin{aligned}
\|(1-\lambda)v_1 + \lambda v_2\|^* &= \sup_{\|f\|\leq 1} \langle f, (1-\lambda)v_1 + \lambda v_2\rangle \\
&\leq \sup_{\|f\|\leq 1} (1-\lambda)\langle f, v_1\rangle + \lambda\langle f, v_2\rangle \\
&\leq (1-\lambda) \sup_{\|f\|\leq 1} \langle f, v_1\rangle + \lambda \sup_{\|f\|\leq 1} \langle f, v_2\rangle \\
&= (1-\lambda)\|v_1^*\| + \lambda\|v_2^*\| \leq 1.
\end{aligned}$$

Y por ser S un subconjunto convexo y cerrado de $\mathbb{R}^m$, entonces por el Teorema 1.1.4, existe $\mathbf{z} \in \mathbb{R}^n$ tal que

$$\mathbf{d}^T\mathbf{z} > 0, \qquad \forall \mathbf{d} \in D$$

ó

$$\sum_{j=1}^{n} \mathbf{z}_j \langle g_j, \mathbf{v} \rangle \leq -\delta \quad \forall \mathbf{v} \in V\left(f(\mathbf{a})\right)$$

para algún $\delta > 0$. Para cualquier $v(\gamma) \in V\left(f(\mathbf{a}+\gamma z)\right)$ con $\gamma > 0$

$$\begin{aligned} \|f(\mathbf{a}+\gamma z)\| &= \langle f(\mathbf{a}+\gamma z), v(\gamma)) \rangle \\ &= \langle f(\mathbf{a}), v(\gamma)) \rangle + \gamma \sum_{j=1}^{n} z_j \langle g_j, v(\gamma) \rangle + o(\gamma) \\ &< \langle f(\mathbf{a}), v(\gamma)) \rangle - \delta\gamma + \gamma \sum_{j=1}^{n} z_j \langle g_j, v(\gamma - v) \rangle + \ o(\gamma) \end{aligned} \tag{1.7}$$

para todo $v \in V\left(f(\mathbf{a})\right)$. Existe una secuencia positiva $\{\gamma_j\} \longrightarrow 0$ tal que

$$\langle u, v(\gamma_j) - v^* \rangle \to 0 \ \ \text{como}\, j \to \infty$$

para todo $u \in S$. Además

$$\begin{aligned} 0 &\leq \|f(\mathbf{a})\| - \langle f(\mathbf{a}), v(\gamma) \rangle \\ &\leq \|f(\mathbf{a}+\gamma z)\| - \langle f(\mathbf{a}), v(\gamma) \rangle \\ &= o(1) \end{aligned}$$

y así $v^* \in V\left(f(\mathbf{a})\right)$. Hacemos que γ tienda a 0 en la desigualdad (1.7) en la sucesión $\{\gamma_j\}$ y tomando $v = v^*$ llegamos a la contradicción que a es un mínimo. □

Teorema 1.2.2. *Sea $f(\boldsymbol{a})$ una función lineal de $\boldsymbol{a}$. Entonces $\boldsymbol{a}$ resuelve (1.3) si y solo si existe $v \in V\left(f(\boldsymbol{a})\right)$ tal que*

$$\langle g_j, v \rangle = 0 \quad j = 1, 2, ..., n. \tag{1.8}$$

Demostración. La implicación necesaria se deriva del Teorema (1.2.1). Supongamos que se cumplen las condiciones de este teorema y consideremos $\mathbf{b}$ como cualquier otro vector en $\mathbb{R}^n$. En ese caso

$$\langle g_j, v \rangle = 0$$

donde g_j es independiente de a. Entonces

$$\langle f(\mathbf{b}), v \rangle \ = \ \langle f(\mathbf{a}), v \rangle \ = \ \|f(\mathbf{a})\|$$

para $v \in V\left(f(\mathbf{a})\right)$ que satisface (1.8). Y esto nos da

$$\|f(\mathbf{a})\| \leq \|f(\mathbf{b})\|.$$

□

1.3. Conceptos básicos y generalidades

Vamos a comenzar dando unas primeras nociones básicas necesarias que utilizaremos de manera recurrente a lo largo de todo el proyecto. Son nociones las cuales la mayoría ya se han trabajado durante el transcurso del grado.

Definición 1.3.1. *(Convexidad) Un conjunto $C \subset \mathbb{R}^n$ es convexo si para todo par de puntos $a, b \in C$ el segmento que los une*

$$[a, b] = \{x = (1 - t)a + tb, \quad \forall t \in [0, 1]\}.$$

Definición 1.3.2. *Sea V un espacio vectorial. Un funcional lineal en V sobre un campo F es una aplicación lineal $V \longrightarrow F$.*

Definición 1.3.3. *Sea V un espacio vectorial sobre un campo F. El espacio dual de V denotado por V^* se define como el conjunto de todos los funcionales lineales $V \longrightarrow F$, con operaciones lineales definidas punto a punto:*

$$(\psi + \phi)(x) := \phi(x) + \psi(x), \quad (\lambda\psi)(x) := \lambda\psi(x).$$

El enfoque de los espacios métricos ofrece una manera general de evaluar la calidad de una aproximación, ya que una de las características fundamentales de un espacio métrico es que incluye una función de distancia. En concreto, la función de distancia $d(x, y)$ de un espacio métrico $\mathbf{T}$ es una función que asigna valores reales y está definida para todos los pares de puntos (x, y) en $\mathbf{T}$.

En la mayoría de los problemas de aproximación, existe un espacio métrico adecuado que contiene tanto a f como al conjunto de aproximaciones A. Entonces, es normal decidir que $a_0 \in A$ es una mejor aproximación que $a_1 \in A$ si se cumple la desigualdad

$$d(a_o, f) \leq d(a_1, f) \tag{1.9}$$

y vamos a definir a $a^* \in A$ es la mejor aproximación si se cumple que

$$d(a^*, f) \leq d. \tag{1.10}$$

Puede ser importante saber si existe o no una mejor aproximación. Una razón es que muchos métodos de cálculo se derivan de propiedades que se obtienen a partir de una mejor aproximación. El siguiente teorema muestra la existencia en el caso en que A es compacto.

Teorema 1.3.1. *Si A es un conjunto compacto en un espacio métrico T, entonces, para cada f en T, existe un elemento $a^* \in T$ tal que la condición (1.10) se cumple para todos los $a \in A$.*

Demostración. Sea d^*

$$d^* = \inf_{a \in A} d(a, f).$$

Si hay un elemento $a \in T$ para el cual se logra este límite en la distancia, entonces se llega a la demostración. De lo contrario,existe una sucesión $\{a_i; i = 1, 2, ...\}$ de puntos de A que dan tiene por límite

$$\lim_{i\to\infty} d(a_i, f) = d^*. \tag{1.11}$$

Por la compacidad, la sucesión tiene al menos un punto límite en A, llamémoslo a^+. Por (1.11) y la definición de a^+, implica que para cualquier $\epsilon > 0$, existe un $k \in \mathbb{Z}$ tal que

$$d(a_k, f) < d^* + \frac{\epsilon}{2}$$

y

$$d(a_k, a^+) < \frac{\epsilon}{2}$$

se cumple. Y por la desigualdad triangular obtenemos que

$$\begin{aligned} d(a^+, f) &\leq d(a^+, a_k) + d(a_k, f) \\ &< d^* + \epsilon \end{aligned}$$

como ϵ puede ser lo suficientemente pequeño que deseemos, la distancia $d(a^+.f)$ no es mayor que la de d^*. Por tanto a^* es la mejor aproximación.

□

Las propiedades de los espacios métricos no son lo suficientemente fuertes para la mayoría de nuestro trabajo, por lo que se asume que A y f están contenidos en un espacio lineal normado, al que también llamamos T cuando queremos referirnos a él. La norma es una función de valores reales $\|x\|$ que está definida para todo $x \in T$. Sus propiedades son tales que

$$d(x, y) = \|x - y\| \tag{1.12}$$

está definida como una función de distancia.

Teorema 1.3.2. *Si A es un espacio vectorial de dimensión finita en un espacio vectorial normado S, entonces, para cada f en S, existe un elemento en A que es la mejor aproximación de A a f.*

Demostración. Sea el subconjunto A_0 que incluye los elementos de A que cumplen con la condición

$$\|a\| \leq 2\|f\| \tag{1.13}$$

es compacto porque es un subconjunto cerrado y acotado de un espacio de dimensión finita. No está vacío, por ejemplo, contiene el elemento cero. Por lo tanto, según el Teorema 1.3.1, existe una mejor aproximación desde A_0 a f, la que llamaremos a_0^*. Por definición, se cumple la desigualdad

$$\|a - f\| \geq \|a_0^* - f\|, \quad a \in A_0. \tag{1.14}$$

Si un elemento **a** está en A pero no está en A_0, entonces, por la condición (1.13) tenemos la cota

$$\begin{aligned} \|a - f\| &\geq \|a\| - \|f\| \\ &> \|f\| \\ &\geq \|a_0^* - f\|, \end{aligned}$$

donde la última desigualdad hace uso del hecho de que el elemento cero está en A_0. Por tanto la expresión (1.14) se cumple para todo $a \in A$ y por tanto a_0^* es la mejor aproximación. □

Capítulo 2

La solución L_1 de un sistema sobredeterminado de ecuaciones lineales

En este capítulo nos centramos en la solución L_1 de un sistema sobredeterminado de ecuaciones lineales

$$A\mathbf{a} = \mathbf{b} \tag{2.1}$$

con A una matriz $m \times n$, $m \geq n$ y $\mathbf{b} \in \mathbb{R}^m$ dado. Denotaremos a las filas de A como $\boldsymbol{\alpha}_1^T, ..., \boldsymbol{\alpha}_m^T$ De modo que $A^T = [\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_m]$. También denotamos a las componentes de $\mathbf{b} = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m]^T$.

Si denotamos con $\mathbf{r}(\mathbf{a}) := \mathbf{b} - A\mathbf{a}$ el vector residuo, se trata de determinar $\mathbf{a} \in \mathbb{R}^n$ que minimiza

$$\|\mathbf{r}(\mathbf{a})\| = \sum_{i=1}^{m} |r_i(\mathbf{a})|. \tag{2.2}$$

Este problema es un caso particular del problema general considerado en el capítulo anterior.

Ahora S es el subespacio de $\mathbb{R}^m$ generado por los vectores columna de la matriz A, el vector $\mathbf{b} \in \mathbb{R}^m$ es un vector arbitrario y se trata de determinar el vector $\mathbf{a}$ de $S = \langle \mathbf{a}_1, ..., \mathbf{a}_n \rangle$ que hace mínima la distancia en la norma L_1 al vector $\mathbf{b}$.

2.1. Caracterización de las soluciones

Dado $\mathbf{a} \in \mathbb{R}^n$, definimos $Z = Z(\mathbf{a})$ como el conjunto de índices "i" para los que $r_i(\mathbf{a}) = 0$ y definimos

$$V(\mathbf{a}) = \{\mathbf{v} \in \mathbb{R}^m / \, |v_i| \leq 1; v_i = sign(r_i(\mathbf{a})), \quad i \notin Z\}.$$

Teorema 2.1.1. *Un vector* $\mathbf{a} \in \mathbb{R}^n$ *satisface (2.2) si y solo si existe un conjunto* $\mathbf{v} \in V(\boldsymbol{a})$ *tal que:*

$$A^T\mathbf{v} = 0. \tag{2.3}$$

Demostración. Supongamos que las condiciones del teorema se cumplen para **a**. Entonces

$$\|\mathbf{r}(\mathbf{a})\| = \mathbf{v}^T\mathbf{r} = \mathbf{v}^T\mathbf{b}$$

para algún $\mathbf{d} \in \mathbb{R}^n$

$$\|\mathbf{r}(\mathbf{d})\| = \sum_{i=1}^{m} |r_i(\mathbf{d})| \geq \sum_{i=1}^{m} v_i r_i(\mathbf{d}) = \mathbf{v}^T\mathbf{b}.$$

Entonces **a** resuelve (2.2).

Ahora, consideremos que **a** es una solución para (2.2) y asumamos que las condiciones del teorema no se cumplen. Sea

$$D = \{\mathbf{d} \in \mathbb{R}^n : \mathbf{a} = A^T\mathbf{v}, \, \mathbf{v} \in V\}$$

el cual es cerrado y convexo de $\mathbb{R}^n$ donde $\mathbf{0} \notin D$. Por lo tanto usando el Teorema 1.1.4, existe $\mathbf{c} \in \mathbb{R}^n$ tal que $\forall \mathbf{d} \in D$, $\mathbf{d}^T\mathbf{c} > 0$, entonces

$$\mathbf{v}^T A\mathbf{c} > 0, \qquad \forall \mathbf{v} \in V. \tag{2.4}$$

Recordando que α_i^T denota la fila "i-ésima"de A,

$$\begin{aligned} \|\mathbf{r}(\mathbf{a} + \gamma\mathbf{c})\| &= \sum_{i=1}^{m} |r_i(\mathbf{a} + \gamma\mathbf{c})| \\ &= \sum_{i=1}^{m} |r_i(\mathbf{a}) - \gamma\alpha_i^T\mathbf{c})| \\ &= \sum_{i \notin Z} v_i(r_i(\mathbf{a}) - \gamma\alpha_i^T\mathbf{c}) + \gamma \sum_{i \in Z} |\alpha_i^T\mathbf{c}|. \end{aligned}$$

Para $\gamma > 0$ suficientemente pequeño, digamos $\gamma < G$ para el cual el signo de $r_i(\mathbf{a} + \gamma\mathbf{c})$ permanece constante y es igual al signo de $r_i(\mathbf{a})$ para todo $i \notin Z$.

Escogemos

$$v_i = -\text{signo}(\alpha_i^T\mathbf{c}) \quad i \in Z.$$

Entonces si $0 < \gamma < G$

$$\begin{aligned} \|\mathbf{r}(\mathbf{a} + \gamma\mathbf{c})\| &= \|\mathbf{r}(\mathbf{a})\| - \gamma\mathbf{v}^T A\mathbf{c} \\ &< \|\mathbf{r}(\mathbf{a})\| \end{aligned}$$

usando (2.4). Esto contradice la suposición de que **a** resuelve (2.2) y demuestra el resultado. □

Teorema 2.1.2. *Si la matriz A tiene rango t, siempre existe una solución* $\boldsymbol{a}$ *para (2.2) tal que Z contiene al menos t índices.*

Demostración. Tomemos A con rango t y $\mathbf{a}$ una solución de (2.2) con $Z(\mathbf{a})$ que contiene $s < t$ indices. Sea $\mathbf{c} \in V(\mathbf{a})$, tal que $\|\mathbf{c}\|_2$ satisface

$$\alpha_i^T \mathbf{c} = 0, \quad i \in Z \tag{2.5}$$

y tomamos $\mathbf{v}$ que satisfaga (2.3). Entonces para un $\gamma > 0$ suficientemente pequeño

$$\begin{aligned} \|\mathbf{r}(\mathbf{a}+\gamma\mathbf{c})\| &= \sum_{i=1}^{m} |r_i(\mathbf{a}) - \gamma\alpha_i^T\mathbf{c}| \\ &= \sum_{i\notin Z} v_i(r_i(\mathbf{a}) - \gamma\alpha_i^T\mathbf{c}) \\ &= \|\mathbf{r}(\mathbf{a})\| - \gamma\sum_{i\notin Z} v_i\alpha_i^T\mathbf{c} \\ &= \|\mathbf{r}(\mathbf{a})\|. \end{aligned}$$

Por la hipótesis del rango de A, existe $\mathbf{c}$ satisfaciendo (2.5) con $\alpha_j^T\mathbf{c} \neq 0$ para algún $j \notin Z$. Por lo tanto, es posible aumentar gradualmente los valores de $|\gamma|$ desde cero hasta que la primera componente de $\mathbf{r}_j(\mathbf{a}+\gamma\mathbf{c})$ $j \notin Z$ se vuelva cero, mientras $\mathbf{a}+\gamma\mathbf{c}$ como solución de (2.2).
La cantidad de componentes igual a cero puede ir aumentando de esta manera hasta que no sea posible encontrar ninguna solución no trivial para (2.5), momento en el que Z contiene al menos t índices.

□

Teorema 2.1.3. *Denotemos a S como el conjunto de soluciones para (2.2) y a* $\boldsymbol{K}$ *como la envolvente convexa de todos los* $\boldsymbol{a}$ *para los cuales* $Z(\boldsymbol{a})$ *contiene t índices. Entonces* $S \equiv \boldsymbol{K}$.

Demostración. Claramente $\mathbf{K} \subset S$. Sea $\mathbf{a}^* \in S$, pero $\mathbf{a}^* \notin K$. Entonces el conjunto convexo y cerrado $\{\mathbf{k}\text{-}\mathbf{a}^*, \mathbf{k} \in \mathbf{K}\}$ no contiene el origen y por el *Teorema* 1.1.4 existe $\mathbf{u} \in \mathbb{R}^n$, tal que para todo $\mathbf{k} \in K$,

$$\mathbf{u}^T\mathbf{k} < \mathbf{u}^T\mathbf{a}^* = \beta.$$

Para cualquier $\mathbf{a} \in S$ el cual $Z(\mathbf{a})$ contenga $s < t-1$ índices, podemos escoger un $\mathbf{c} \in \mathbb{R}^n$, tal que $\|\mathbf{c}\|_2 = 1$, que cumpla

$$\alpha_i^T\mathbf{c} = 0 \quad i \in Z(\mathbf{a})$$

$$\mathbf{u}^T\mathbf{c} = 0.$$

Por lo tanto, según el Teorema 2.1.2, podemos encontrar $\mathbf{d} \in S$ con $Z(\mathbf{d})$ que contiene $(t-1)$ índices y que satisface $\mathbf{u}^T\mathbf{d} = \beta$. Sea $\mathbf{c} \in \mathbb{R}^n$, $\|\mathbf{c}\|_2 = 1$, de manera que

$$\alpha_i^T\mathbf{c} = 0 \quad i \in Z\,(d)$$

$$\mathbf{u}^T\mathbf{c} \geq 0.$$

Luego, al aumentar el valor de γ en una dirección positiva, podemos obtener $\mathbf{d}+\gamma\mathbf{c} \in S$ con $Z(\mathbf{d}+\gamma\mathbf{c})$ que contiene t índices. Además, $\mathbf{u}^T(\mathbf{d}+\gamma\mathbf{c}) \geq \beta$ lo cual lleva a una contradicción. Por tanto $S \subset K$.

□

El anterior teorema nos dice que si A tiene rango máximo es posible computar una solución óptima en el sentido L_1 encontrando todos los vectores $\mathbf{a}$ que hacen mínimo la norma L_1 del residuo y a la vez hacen n de los residuos del sistema iguales a cero.

En particular esto reduce el problema a calcular la solución L_1 del sistema (2.1) a un problema combinatorio:
Se resuelven todos los posibles subsistemas lineales formados por n ecuaciones (en las n incógnitas de $\mathbf{a}$), y se determinan entre esas soluciones las que tienen norma L_1 mínima. La envolvente convexa de esas soluciones es el conjunto de soluciones del problema.

Naturalmente este procedimiento no es el más efectivo, pero indica que el proceso de solución puede determinarse en un número finito de pasos.

2.2. El problema L_1 y la programación lineal

Para comenzar debemos reformular (2.2) como

$$\text{minimizar} \quad e^T w = \sum_{i=1}^{w} \omega_i \tag{2.6}$$

$$\text{sujeto a } w_i \geq |a_i^T x - \beta_i|, \quad i = 1, ..., m,$$

donde $e = [1, ..., 1]^T$ y $w = [\omega_1, ..., \omega_m]^T$ que es equivalente a

$$\text{minimizar} \quad e^T w \tag{2.7}$$

$$\text{sujeto a } w_i \geq (a_i^T x - \beta_i), \quad i = 1, ..., m,$$
$$\text{y} \quad w_i \geq (\beta_i - a_i^T x), \quad i = 1, ..., m.$$

Y si lo ponemos en notación matricial esto se convierte en

$$\text{minimizar} \begin{bmatrix} e^T, & 0^T \end{bmatrix} \begin{bmatrix} w \\ x \end{bmatrix} \tag{2.8}$$

$$\text{sujeto a } \begin{bmatrix} I & -A^T \\ I & A^T \end{bmatrix} \begin{bmatrix} w \\ x \end{bmatrix} \geq \begin{bmatrix} -b \\ b \end{bmatrix}.$$

Podemos obtener el problema dual de (2.8) y es:

$$\text{maximizar} \quad -u^T b + v^T b \tag{2.9}$$

$$\begin{aligned} &\text{sujeto a} \quad u + v = e, \\ &Av - Au = 0, \\ &\text{y} \quad u \geq 0, \quad v \geq 0. \end{aligned}$$

Y después de realizar las sustituciones de $v = e - u$ e $y = 2u - e$, tenemos el problema

$$\begin{aligned} &\text{minimizar} \quad b^T y \\ &\text{sujeto a} \quad Ay = 0, \\ &\text{y} \quad -e \leq y \leq e. \end{aligned} \tag{2.10}$$

2.3. Métodos de optimización por descenso directo

En este capítulo hablaremos de manera concreta de la implementación del algoritmo de descenso directo en Matlab y tomando algún ejemplo para ver los resultados finales obtenidos.

La esencia del enfoque se encuentra en la demostración de la necesidad de las condiciones del Teorema 2.1.1, ya que si podemos obtener un vector $\mathbf{c}$ que cumpla con (2.4), entonces $\mathbf{c}$ representa una dirección de descenso para $\|\mathbf{r}\|$. La dirección de descenso particular utilizada en el algoritmo se define de manera directa y conveniente de la siguiente manera.

Para mantener la unidad en la notación, deseamos considerar que todos los vectores sean vectores columna. En consecuencia, formularemos el problema de una manera ligeramente diferente a la que se utiliza comúnmente. Sea A una matriz real $n \times m$ $(m > n \geq 2)$ con $\alpha_1, \alpha_2, ..., \alpha_m$ columnas. Sea $\mathbf{b}$ un vector con m componentes reales $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_m$. Queremos resolver el sistema sobredeterminado de ecuaciones lineales

$$A^T \mathbf{a} = \mathbf{b} \tag{2.11}$$

en la norma L_1. Es decir, deseamos encontrar un vector real con componentes que resuelva el problema:

$$\text{minimizar} \quad \phi(\mathbf{a}) = \|r(\mathbf{a})\| = \|A\mathbf{a} - b\|_1 = \sum_{i=1}^{m} |\boldsymbol{\alpha}_i^T \mathbf{a} - \boldsymbol{\beta}_i|. \tag{2.12}$$

Por el momento, no se imponen restricciones en $\mathbf{a}$. Sin embargo, las restricciones de desigualdad lineal se pueden acomodar fácilmente en el problema. Para cualquier punto $\mathbf{a}$, denotamos por Z al conjunto de índices donde el residuo es cero

$$Z = \{i / \boldsymbol{\alpha}_i^T \mathbf{a} - \boldsymbol{\beta}_i = 0\} = \{i_1, ..., i_k\}. \tag{2.13}$$

Podemos estar interesados en el conjunto de ecuaciones que se satisfacen y su conjunto complementario (las ecuaciones que no se satisfacen). A este conjunto, asociamos la matriz

$$A_Z^T = [\boldsymbol{\alpha}_{i_1}, \boldsymbol{\alpha}_{i_2}, ..., \boldsymbol{\alpha}_{i_k}] \tag{2.14}$$

que estará formada por las filas de A que están relacionadas con las ecuaciones satisfechas, y también denotaremos con N al núcleo de A_z

$$N = N(A_Z^T) = \{y / \boldsymbol{\alpha}_i^T y = 0 \quad \forall i \in Z\}. \tag{2.15}$$

El proyector ortogonal en el conjunto N se denotará como P_N. En cada punto $\mathbf{a}$ que estamos considerando, los vectores

$$r = \mathbf{b} - A\boldsymbol{\alpha} = [\rho_1, ..., \rho_m]^T \tag{2.16}$$

y

$$s = \text{signo}(\mathbf{r}) = [\text{signo}(\rho_1), ..., \text{signo}(\rho_m)]^T = [\sigma_1, ..., \sigma_m]^T. \tag{2.17}$$

Frecuentemente haremos referencia a estos vectores. Finalmente, consideraramos el conjunto

$$D = \{\eta_l / l \in Z^c \quad \text{y} \quad \eta_l = -\rho_l / \boldsymbol{\alpha}_l^T \mathbf{c} \quad \text{y} \quad \eta_l > 0\} \tag{2.18}$$

para un vector $\mathbf{c}$ dado.

En las secciones siguientes, propondremos un algoritmo para el problema (2.12). Debemos anticipar, como ocurre con muchos algoritmos, que tendremos que considerar el nuestro desde dos perspectivas distintas: la matemática y la computacional. Desde ninguna de estas perspectivas se requerirá que A tenga rango completo. Sin embargo, las dependencias lineales en la matriz A asociadas con ciertos puntos $\mathbf{a}$ requerirán una consideración especial desde el punto de vista matemático, pero pueden ser tratadas de manera menos importante desde el punto de vista computacional. Las dependencias que nos preocupan corresponden en concepto a vértices degenerados en los problemas de programación lineal asociados.

Como vimos en la sección 2.2, el problema (2.12) es precisamente equivalente al problema de programación lineal. Dado que el punto de vista matemático exige cierta consideración de la degeneración cuando se aborda el problema de programación lineal, la degeneración debe ser al menos un problema pasajero para nosotros en el problema.

2.3.1. Cambios lineales en el argumento de ϕ

Tomemos cualquier punto $\mathbf{a}$ y consideremos el vector $\mathbf{c}$ de n-componentes como indicativo de una dirección en la que nos alejamos de $\mathbf{a}$. Nos interesa determinar el valor de ϕ en un punto trasladado $\boldsymbol{\alpha} + \gamma\mathbf{c}, \quad \gamma > 0$.

$$\begin{aligned}\phi(\mathbf{a}+\gamma\mathbf{c}) &= \sum_{i=1}^{m}\left[\boldsymbol{\alpha}_i^T(\mathbf{a}+\gamma\mathbf{c})-\boldsymbol{\beta}_i\right]\cdot \text{signo}\left[\boldsymbol{\alpha}_i^T(\mathbf{a}+\gamma\mathbf{c})-\boldsymbol{\beta}_i\right] \\ &= \sum_{i\in Z^c}\left[\rho_i+\gamma\boldsymbol{\alpha}_i^T\mathbf{c}\right]\cdot \text{signo}\left[\rho_i+\gamma\boldsymbol{\alpha}_i^T\mathbf{c}\right]+\gamma\sum_{i\in Z}\left[\boldsymbol{\alpha}_i^T\mathbf{c}\right]\cdot \text{signo}\left[\boldsymbol{\alpha}_i^T\mathbf{c}\right]\end{aligned} \tag{2.19}$$

donde estamos usando el hecho de que $\rho_i = \boldsymbol{\alpha}_i^T\mathbf{a} - \boldsymbol{\beta}_i$ es cero cuando $i \in Z$. Si $\gamma > 0$ es lo suficientemente pequeño, si se cumple que

$$0 < \gamma < \text{mín}\, D \tag{2.20}$$

tal que D está dado en (2.18), entonces signo $\left[\rho_i + \gamma\boldsymbol{\alpha}_i^T\mathbf{c}\right] = \text{signo}\left[\rho_i\right] = \sigma_i \quad \forall i \in Z^c$ y (2.19) sse reduce a

$$\phi(\mathbf{a}+\gamma\mathbf{c}) = \sum_{i\in Z^c}|\rho_i| + \gamma\left(\sum_{i\in Z^c}\sigma_i\boldsymbol{\alpha}_i^T\mathbf{c} + \sum_{i\in Z}\left|\boldsymbol{\alpha}_i^T\mathbf{c}\right|\right). \tag{2.21}$$

Definimos el vector $\mathbf{h}$ como

$$\mathbf{h} = \sum_{i\in Z^c}\sigma_i\boldsymbol{\alpha}_i \tag{2.22}$$

entonces, la expresión (2.21) adquiere una forma que se asemeja a una expansión de Taylor

$$\phi(\mathbf{a}+\gamma\mathbf{c}) = \phi(\mathbf{a}) + \gamma h^T\mathbf{c} + \gamma\sum_{i\in Z}\left|\boldsymbol{\alpha}_i^T\mathbf{c}\right|, \quad \forall\mathbf{c}\in N. \tag{2.23}$$

2.3.2. Proyección

Si consideramos solo aquellos vectores de dirección $\mathbf{c}$ que están en el núcleo de A_Z^T, según se define en (2.14) y (2.15), entonces el último término de (2.23) será cero.

$$\phi(\mathbf{a}+\gamma\mathbf{c}) = \phi(\mathbf{a}) + \gamma h^T\mathbf{c}, \quad \forall p \in N. \tag{2.24}$$

Necesitamos determinar algún $\mathbf{c} \in N$ tal que $h^T\mathbf{c} < 0$ para el punto $\mathbf{a} + \gamma\mathbf{c}$ en el cual su valor en ϕ es estrictamente menor que $\phi(\mathbf{a})$. De hecho, γ podría ser elegido igual al límite superior en (2.20) para lograr la mayor reducción en el valor de ϕ con respecto a todos los $\gamma > 0$ para los cuales (2.24) se cumple. Al romper la limitación de este límite superior, como veremos en una sección posterior, se pueden realizar elecciones aún mejores de γ.

La elección más sencilla de $\mathbf{c}$ es

$$\mathbf{c} = -P_N h, \tag{2.25}$$

que es la proyección de $-h$ en el núcleo de A_Z^T mientras esta proyección no sea cero. Si la proyección es cero, el punto $\mathbf{a}$ se le llamará *punto muerto*.

2.3.3. Optimización y puntos muertos

Consideremos vectores de dirección generales $\mathbf{c}$, es decir, volvamos a considerar las ecuaciones (2.21) y (2.23). Claramente $\phi(\mathbf{a}+\gamma\mathbf{c}) \geq \phi(\mathbf{a}), \quad \forall\gamma > 0$ lo suficientemente cerca del cero si y solo si

$$h^T\mathbf{c} + \sum_{i\in Z}|\boldsymbol{\alpha}_i^T\mathbf{c}| \geq 0, \quad \forall\mathbf{c}. \tag{2.26}$$

Si este es el caso, entonces $\mathbf{a}$ proporciona un mínimo local para la función ϕ.

Como $\phi(\mathbf{a}) = \|A\mathbf{a} - b\|_1$ es una función convexa, los mínimos locales son mínimos globales. Así que $\mathbf{a}$ es óptimo para (2.12) si y solo si (2.26) se cumple. Debe destacarse que la solución al problema L_1 no necesita ser única.

Supongamos que $P_N\mathbf{h}$ no es cero. Entonces, el vector $\mathbf{c}$ elegido como en (2.25) violará (2.26), por lo tanto, el correspondiente $\mathbf{a}$ no puede ser óptimo. Por otro lado, asumamos que $P_N h = 0$, decimos entonces que el $\mathbf{a}$ correspondiente es un punto muerto. Entonces $\mathbf{h}$ está en el espacio generado por las columnas A_Z:

$$\mathbf{h} = A_Z\mathbf{w} = \sum_{j=1}^{k}\mathbf{w}_j\boldsymbol{\alpha}_{i_j} \tag{2.27}$$

para algún $\mathbf{w} = [w_1, .., w_k]^T$ y para $i_j \in Z, \quad i = 1, ...k$. Por (2.26) se puede reescribir cómo

$$\sum_{j=1}^{k}\left[1 + \text{signo}(\boldsymbol{\alpha}_{i_j}^T\mathbf{c})\omega_j\right]|\boldsymbol{\alpha}_{i_j}^T\mathbf{c}|. \tag{2.28}$$

Vemos que $\boldsymbol{\omega}$ es único si y solo si las columnas de A_Z^T son linealmente independientes. El punto $\mathbf{a}$ en el que esto ocurre se llama punto muerto no degenerado. Asumimos por tanto que $\boldsymbol{\omega}$ es único.
Notemos que (2.28) solo puede ser negativo si $|\omega_{j_0}| > 1$ para algún j_0. Si $|\omega_j| \leq 1$ para todo j, entonces, todos los términos de la suma deben ser no negativos sin importar qué vector $\mathbf{c}$ se considere. Supongamos que un valor como el mencionado anteriormente existe y tomamos

$$\mathbf{c} = -\text{signo}(\omega_{j_0})P_{N_{j_0}}\boldsymbol{\alpha}_{i_{j_0}}, \tag{2.29}$$

donde N_{j_0} es el núcleo nulo de A_Z con las columnas $\boldsymbol{\alpha}_{i_{j_0}}$ eliminadas. Para esta elección de $\mathbf{c}$ tenemos $\mathbf{c}^T\boldsymbol{\alpha}_i = 0$ si $i \in Z - \{i_{j_0}\}$ y obtenemos

$$\text{signo}(\mathbf{c}^T\boldsymbol{\alpha}_{i_{j_0}})\omega_{j_0} = -|\omega_{j_0}|. \tag{2.30}$$

Esto proporciona un valor negativo para (2.28). Por lo tanto, un punto muerto no degenerado es óptimo si y solo si $|\omega_j| \leq 1, \quad \forall j = 1, ..., k$.

2.3.4. El gradiente restringido y la elección de γ

Implícito en todo lo anterior es el resultado de que o bien a es óptimo, o bien hay un $\mathbf{c}$ tal que

$$\phi(\mathbf{a}+\gamma\mathbf{c}) = \phi(\mathbf{a}) + \gamma\mathbf{c}^T g, \quad \mathbf{c}^T\mathbf{g} < 0 \tag{2.31}$$

para todo los γ que cumpla (2.20). En el caso de que $\mathbf{a}$ no sea un punto muerto, tomamos

$$\mathbf{g} = h \quad y \quad \mathbf{c} = -P_N\mathbf{h}. \tag{2.32}$$

En el caso de que $\mathbf{a}$ sea un punto muerto no degenerado, podemos tomar

$$\mathbf{g} = \mathbf{h} + \text{signo}(\mathbf{c}^T\boldsymbol{\alpha}_{i_{j_0}}) \quad y \quad \mathbf{c} = -\text{signo}(w_{j_0})P_{N_{j_0}}\boldsymbol{\alpha}_{i_{j_0}}, \tag{2.33}$$

donde j_0 es el índice de una componente de $\mathbf{w}$, como se muestra en (2.27), para la cual $|w_{j_0}| > 1$. Nos referiremos a esto como el gradiente restringido de la función ϕ.
Cuando el gradiente restringido existe bajo las reglas anteriores, es posible producir un punto $\mathbf{a}+\gamma\mathbf{c}$ en el cual el valor de ϕ es menor que $\phi(\mathbf{a})$. Nuestras suposiciones nos permitirán hacer $\mathbf{a}$ no más grande que la cota dada en (2.20).

Sin embargo, si $\mathbf{a}$ se eligiera ligeramente mayor que esta cota, entonces es fácil de verificar que el gradiente restringido en el punto $\mathbf{a}+\gamma\mathbf{c}$ para ese valor de γ es

$$g' = g - 2\cdot\sum_{i\in L}\sigma_i\boldsymbol{\alpha}_i. \tag{2.34}$$

Donde L conjunto de índices es tomado de los índices de Z^c, para los cuales se alcanza la cota superior de (2.20). De hecho, en problemas no degenerados, L consistirá en un solo índice l', de modo que

$$g' = g - 2\cdot\sigma_{l'}\eta_{l'} \tag{2.35}$$

donde l' proporciona el mínimo de las razones positivas descritas en (2.18). Todavía puede ocurrir que $\mathbf{c}^T g' < 0$; es decir, que

$$\mathbf{c}^T\mathbf{g} < 2\mathbf{c}^T\left(\sum_{i\in L}\sigma_i\boldsymbol{\alpha}_i\right), \tag{2.36}$$

En este caso, γ puede incrementarse al siguiente valor más grande en el conjunto D para lograr una disminución adicional en ϕ. Podemos continuar con valores sucesivos en el conjunto D, ajustando el gradiente restringido de acuerdo a (2.34) y aplicandolo en (2.36). Este proceso debe terminar el criterio en algún valor $\eta \in D$, ya que si pudiéramos aumentar a η indefinidamente, ϕ podría disminuir indefinidamente, no puede ocurrir.
Una vez que se ha determinado un valor η máximo en D, podemos sustituir $\mathbf{a}$

por $\mathbf{a}+\gamma\mathbf{c}$ y reconsiderar la búsqueda de una dirección de descenso para ϕ desde este nuevo punto.

2.3.5. Algoritmo (Caso no degenerado)

1. Seleccionar cualquier punto $\mathbf{x}_0$

2. *a*) Identificar $Z_a = \{i_1, ..., i_k\}$, (Índices de ecuaciones con residuo 0)
   Tomamos A_Z como hemos definido en (2.14) y N como en (2.15).

   b) Calculamos $\mathbf{h}$ como se indica en (2.22).

   c) Calculamos $\mathbf{c} = -P_N\mathbf{h}$
   Si $\mathbf{c} \neq 0$, tomamos $\mathbf{g} = \mathbf{h}$ y vamos al paso **(3)**, si no

   d) Calculamos $\boldsymbol{\omega}$ como en (2.27).

   e) Si $|w_j| \leq 1$ para todo $j = 1, ..., k$ enconces $\mathbf{x}_0$ es **óptimo** y paramos el algoritmo. **(STOP)**

   Si no ocurre esto,

   f) Buscamos $i_{j_0} \in Z$ tal que $|w_{j_0}| > 1$.

   g) Reemplazamos Z por $Z - \{i_{j_0}\}$ y hacemos los correspondientes cambios en A_Z y en N.

   h) Calculamos
   $$\mathbf{c} = -\text{signo}(w_{j_0})P_N\boldsymbol{\alpha}_{i_{j_0}},$$
   y tomamos
   $$\mathbf{g} = h - \text{signo}(w_{j_0})\boldsymbol{\alpha}_{i_{j_0}}$$

3. Determinamos los elementos de D como en (2.18) y los ordenamos:
   $$0 < \eta_{l_1} < ... < \eta_{l_t},$$
   donde η_{l_j} corresponden cada uno a $-\sigma_{l_j}/\boldsymbol{\alpha}_{l_j}^T\mathbf{c}$. Las desigualdades estrictas se derivan al considerar el caso de puntos muertos no degenerados, es decir, puntos para los que las columnas de A_z son linealmente independientes.
   Tomamos $k = 1$.

4. Si $\mathbf{c}^T\mathbf{g} \geq 2\sigma_{l_k}\mathbf{c}^T\boldsymbol{\alpha}_{l_k}$, entonces vamos al paso número **(6)**. Si no:

5. Sustituimos $\mathbf{g}$ por $\mathbf{g} - 2\boldsymbol{\alpha}_{l_k}\eta_{l_k}$.
   Sumamos 1 a k, es decir, pasamos de k a $k+1$. Mientras satisfaga la condición $\mathbf{c}^T\mathbf{g} \geq 2\sigma_{l_k}\mathbf{c}^T\boldsymbol{\alpha}_{l_k}$.
   Vamos al paso **(4)**.

6. Reemplazamos $\mathbf{a}$ por $\mathbf{a} + \boldsymbol{\eta}_{l_k} \cdot \mathbf{c}$ y volvemos a empezar en **(2)** una nueva iteración.

2.3.6. Implementación del proyector P_N

Si la matriz A_Z^T tiene columnas independientes es posible factorizarla (ortogonalización de Gram-Schmidt) en la forma

$$A_z^T = Q \cdot \begin{bmatrix} R \\ - \\ 0 \end{bmatrix} = \begin{bmatrix} Q_1 & | & Q_2 \end{bmatrix} \cdot \begin{bmatrix} R \\ - \\ 0 \end{bmatrix} \tag{2.37}$$

donde $Q = \begin{bmatrix} Q_1 & | & Q_2 \end{bmatrix}$ es una matriz $1 \times n$ ortogonal y R es un matriz $k \times k$ triangular superior y no singular. Q_1 denota las primeras k columnas de Q y Q_2 las $n - k$ columnas restantes.

El proyector sobre N el núcleo de A_z, está dado por la matriz

$$Q_2 Q_2^T \tag{2.38}$$

y el vector $\mathbf{w}$ de (2.27) viene dado por

$$\mathbf{w} = R^{-1} Q_1^T \mathbf{h}. \tag{2.39}$$

La implementación efectiva del algoritmo considera la actualización de la factorización (2.37) cuando a A_Z^T se añade o elimina una columna.

Hay técnicas matriciales muy efectivas para esta actualización que no consideramos. La implementación del algoritmo en Matlab que presentamos en el apéndice recomputa partiendo de cero la factorización QR en cada iteración.

2.3.7. Convergencia en un número finito de pasos

Comenzando en cualquier punto **a**, el algoritmo genera una secuencia de puntos adicionales, cada miembro de la secuencia se obtiene de su predecesor **a'** en el paso (**6**). El parámetro η_{l_v} en el paso (**6**) se elige de manera que el nuevo punto cumple que la ecuación asociada con el índice l_v del paso (**4**). Si el punto **a** no es un punto muerto, entonces el vector de dirección **c** generará el punto siguiente **a"** el cual se elige en el paso (**2**, (c)). Esto garantizará el nuevo punto **a"** generado a partir de **a'** cumplirá todas las ecuaciones que ya cumplía **a'** Este proceso se realiza para asegurarse de que cada nuevo punto generado mantenga las condiciones satisfechas por su predecesor, lo cual es crucial para el progreso del algoritmo y la convergencia hacia la solución óptima. Además, **c** se habrá elegido de manera que $\phi(\mathbf{a}'') < \phi(\mathbf{a}')$.

Así que, a menos que se encuentre un punto muerto, el conjunto de ecuaciones cumplidas por cada nuevo punto contiene estrictamente al conjunto de ecuaciones cumplidas por su predecesor, y se produce una secuencia estrictamente decreciente de valores desde ϕ.
Una consecuencia de lo anterior es el hecho de que los pasos (**2**) hasta el paso (**6**) no pueden ejecutarse más de m veces consecutivas antes de llegar a un punto

muerto $\hat{\mathbf{a}}$. Para puntos muertos, la determinación de si existe un vector $\mathbf{c}$ que servirá como dirección de descenso se da en los pasos (**2**, (d)) hasta el paso (**2**, h).
Vemos que la submatriz $\hat{A}_Z$ de A está asociada al punto muerto $\hat{\mathbf{a}}$. POr tanto se también tiene asociado dicho punto el subespacio $N(\hat{A}_Z^T)$. Vemos que (2.24) y (2.27) implica que ϕ es constante en todo el núcleo, por lo que debemos considerar constante al valor $\hat{\phi} = \phi(\hat{\mathbf{a}})$.

Supongamos que se encuentra un vector $\mathbf{c}$ en el paso (**2**) el cual viola a (2.26). Entonces $\phi(\hat{\mathbf{a}} + \gamma\mathbf{c}) < \hat{\phi}$ para algún $\gamma > 0$ y y para todos los puntos que el algoritmo genere a partir de este punto en adelante, el valor de ϕ será estrictamente menor que $\hat{\phi}$. Esto significa que nunca se encontrará ningún punto muerto asociado a $\hat{A}_Z$ después de que hayamos pasado el punto $\hat{\mathbf{a}}$ en el algoritmo.

Por lo tanto, podemos concluir que solo puede haber un número finito de ejecuciones de los pasos (**2**) a (**6**) entre cualquier par de puntos muertos, y que solo se puede generar un número finito de puntos muertos, ya que cada uno está asociado con una submatriz de $\mathbf{A}$ que solo debemos encontrar una vez. Por otro lado, algún punto muerto es óptimo y el algoritmo no lo terminará hasta que se haya encontrado un punto muerto óptimo.

Capítulo 3

Aproximación óptima L_1 de funciones continuas

Una aproximación óptima L_1 desde un subconjunto A del espacio $C[a,b]$ de las funciones continuas en el intervalo $[a,b]$ a una función $f \in C[a,b]$ es un elemento $p^* \in A$ tal que minimiza la expresión:

$$\|f-p\|_1 = \int_a^b |f(x)-p(x)|dx, \quad p \in A. \tag{3.1}$$

El caso de mayor interés en el que nos centraremos, es cuando el conjunto A es un subespacio vectorial de dimensión finita.
En este caso si $\{\phi_1, ..., \phi_n\}$ es una base del subespacio A, los resultados generales del capítulo 1 garantizan la existencia de aproximaciones óptimas a f en A, es decir, que minimizan aproximaciones $p^* = \sum_{i=1}^n a_i\phi_i(x)$

$$\|r(\mathbf{x},\mathbf{a})\|_{L_1} = \|f(x) - \sum_{i=1}^n a_i\phi_i(\mathbf{x})\|_{L_1}, \quad a = (a_1, ..., a_n) \in \mathbb{R}^n.$$

Las condiciones necesarias y suficientes que caracterizan que la función p^* en A sea la mejor aproximación L_1 a f se dan en la siguiente sección.
Esta es la situación, por ejemplo, cuando $A \subseteq \mathbb{P}_n$ es el espacio de los polinomios de grado menor o igual que n en $\mathbf{x}$. Estas condiciones tienen la propiedad interesante de que toda la dependencia de f está contenida en la función de signo:

$$s^*(x) = \begin{cases} -1 & si \quad f(x) < p^*(x). \\ 0 & si \quad f(x) = p^*(x). \qquad a \le x \le b. \\ 1 & si \quad f(x) > p^*(x). \end{cases} \tag{3.2}$$

Por lo tanto, se deduce que si p^* es la mejor aproximación a f, y si f se modifica de alguna manera sin que se altere la función de signo (3.2), entonces p^*

seguirá siendo aproximación óptima. También es relevante el caso discreto de aproximación a f por elementos de A. Aquí requerimos que $p^* = \sum_{i=1}^{n} a_i\phi_i(x)$ minimice

$$\sum_{t=1}^{m}|f(x_t) - p(x_t)|, \quad p \in A \tag{3.3}$$

donde $\{x_t/t = 1, 2, ..., m\}$ es un conjunto finito de puntos en el intervalo $[a, b]$. Como motivación consideramos la aproximación óptima por una constante de una función f continua monótona creciente en $[a, b]$. Se trata de:

$$\text{minimizar} \quad \varphi(\mathbf{c}) = \int_a^b |c - f(x)|\, dx.$$

La integral anterior la podemos expresar tal que

$$\begin{aligned}\int_a^b |c - f(x)|\, dx &= \int_{f(a)}^{c} dy \left[\int_a^{f^{-1}(y)} dx\right] + \int_c^{f(b)} dy \left[\int_{f^{-1}(y)}^{b} dx\right] \\ &= \int_{f(a)}^{c} \left(f^{-1}(y) - a\right) dy + \int_c^{f(b)} \left(b - f^{-1}(y)\right) dy.\end{aligned}$$

Derivamos $\varphi(c)$ y obtenemos

$$\varphi'(c) = \left(f^{-1}(c) - a\right) - \left(b - f^{-1}(c)\right) = 0,$$

lo que implica que

$$f^{-1}(c) = \frac{a+b}{2},$$

es decir

$$c = f\left(\frac{a+b}{2}\right).$$

3.1. Caracterización de las soluciones

La caracterización de las soluciones de (3.1) puede estar dada de diferentes formas, de las cuales la que se proporciona aquí es quizás la más conveniente. Un papel importante es desempeñado por el conjunto $Z(\equiv Z(\mathbf{a}))$ de puntos $\mathbf{x} \in X$ para los cuales el error $f(x) - p(x)$ es cero. También utilizaremos $Z(\epsilon)$ para denotar el conjunto

$$Z(\epsilon) = \{\mathbf{x} : |f(x) - p(x)| \leq \epsilon\},$$

que está definido para todo $\epsilon > 0$.

Denotamos $\theta(\mathbf{x}, \mathbf{a})$ a la función signo del residuo

$$\theta(\mathbf{r}, \mathbf{a}) = \text{signo}(f(x) - \sum_{i=1}^{n} a_i\phi_i(x)), \quad \mathbf{x} \in [a, b] \tag{3.4}$$

y con $V(\mathbf{a})$ al conjunto

$$V(\mathbf{a}) = \{v : [a,b] \to \mathbb{P}_n / \max_{\mathbf{x}\in[a,b]} |v(\mathbf{x})| \leq 1 \quad \text{y} \quad v(x) = \theta(x,a),\, x \notin Z\}.$$

Teorema 3.1.1. *Un elemento $\boldsymbol{p}^* = \sum_i^n a_i\phi_i(x) \in A$ resuelve (3.1) si y solo si existe $v(\boldsymbol{x}) \in V(a)$ tal que*

$$\int_a^b v(\boldsymbol{x})\phi_j(\boldsymbol{x})dx = 0, \quad j = 1,2,...,n. \tag{3.5}$$

El decir, la función signo del error es ortogonal al subespacio A.

Demostración. Supongamos que las condiciones del teorema se cumplen en el elemento **a**. Entonces se tiene

$$\begin{aligned}
\|r(\mathbf{x},\mathbf{a})\| &= \int_{[a,b]-Z} |r(\mathbf{x},\mathbf{a})|dx \\
&= \int_{[a,b]-Z} \theta(\mathbf{x},\mathbf{a}) \cdot \left(f(\mathbf{x}) - \sum_{j=1}^{n} a_j\phi_j(\mathbf{x}) \right) dx \\
&= \int_{[a,b]-Z} \theta(\mathbf{x},\mathbf{a})\,(f(\mathbf{x}))\,dx + \int_Z v(\mathbf{x}) \cdot f(\mathbf{x})dx.
\end{aligned} \tag{3.6}$$

Tomemos $\mathbf{d} \in \mathbb{R}^n$ arbitrario ($\mathbf{d} \neq \mathbf{a}$)

$$\begin{aligned}
\|r(\mathbf{x},\mathbf{d})\| &\geq \int_{[a,b]-Z} \theta(\mathbf{x},\mathbf{a})\; r(\mathbf{x},\mathbf{d})dx + \int_{[a,b]} v(\mathbf{x}) \cdot r(\mathbf{x},\mathbf{d})dx \\
&= \int_{[a,b]-Z} \theta(\mathbf{x},\mathbf{a}) f(\mathbf{x})dx + \int_Z v(\mathbf{x}) \cdot f(\mathbf{x})dx \\
&= \|r(\mathbf{x},\mathbf{a})\|
\end{aligned}$$

usando (3.6). Por tanto $\mathbf{p}^* = \sum_{i=1}^n a_i\phi_i(x)$ resuelve 3.1.

Ahora tomamos **p** que sea solución de (3.1) y suponemos que las condiciones del teorema no se cumplen. Tomamos

$$D = \{\mathbf{d} \in \mathbb{R}^n; d_j = \int_{[a,b]} v(\mathbf{x})\phi_j(\mathbf{x})dx, \qquad v(\mathbf{x}) \in V, \quad j = 1,2,...,n\}.$$

Entonces D es cerrado y es un subconjunto convexo de $\mathbb{R}^n$ tal que $0 \notin D$. Entonces por el teorema 1.1.4, existe $\mathbf{c} \in \mathbb{R}^n, \delta > 0$, tal que

$$\mathbf{d}^T c \geq \delta > 0, \quad \forall \mathbf{d} \in D$$

ó

$$\sum_{j=1}^{n} c_j \int_{[a,b]} v(\mathbf{x})\phi_j(\mathbf{x})dx \geq \delta > 0, \quad \forall v(\mathbf{x}) \in V. \tag{3.7}$$

Si $\gamma > 0$ entonces

$$\begin{aligned}\|r(\mathbf{x},\mathbf{a}+\gamma\mathbf{c})\| &= \int_{[a,b]} \left| r(\mathbf{x},\mathbf{a}) - \gamma \sum_{j=1}^{n} c_j \phi_j(\mathbf{x}) \right| dx \\ &= \gamma \int_{[a,b]} \left| \sum_{j=1}^{n} c_j \phi_j(\mathbf{x}) \right| dx + \int_{[a,b]-Z} \left| r(\mathbf{x},\mathbf{a}) - \gamma \sum_{j=1}^{n} c_j \phi_j(\mathbf{x}) \right| dx.\end{aligned}$$

Ponemos

$$v(\mathbf{x}) = \begin{cases} -\text{signo}\left(\sum_{j=1}^{n} c_j \phi_j(\mathbf{x})\right) & si \qquad x \in Z \\ \\ \theta(\mathbf{x},\mathbf{a}) & si \quad x \in [a,b] - Z. \end{cases}$$

entonces $v(\mathbf{x}) \in V$ y por (3.7)

$$\int_Z \left| \sum_{j=1}^{n} c_j \phi_j(\mathbf{x}) \right| dx \leq \int_{[a,b]-Z} \theta(\mathbf{x},\mathbf{a}) \sum_{j=1}^{n} c_j \phi_j(\mathbf{x}) dx - \delta.$$

Tomamos $M = \text{máx}_{x\in[a,b]} = \left| \sum_{j=1}^{n} c_j \phi_j(\mathbf{x}) \right|$ y tomamos $\epsilon = M\gamma$
Entonces

$$\begin{aligned}\|r(\mathbf{x},\mathbf{a}+\gamma\mathbf{c})\| &= -\gamma\delta + \gamma \int_{[a,b]-Z} \theta(\mathbf{x},\mathbf{a}) \sum_{j=1}^{n} c_j \phi_j(\mathbf{x}) dx \\ &\quad + \int_{[a,b]-Z} \left| r(\mathbf{x},\mathbf{a}) - \gamma \sum_{j=1}^{n} c_j \phi_j(\mathbf{x}) \right| dx \\ &= -\gamma\delta + \gamma \int_{[a,b]-Z(\epsilon)} \theta(\mathbf{x},\mathbf{a}) \sum_{j=1}^{n} c_j \phi_j(\mathbf{x}) dx \\ &\quad + \gamma \int_{Z(\epsilon)-Z} \theta(\mathbf{x},\mathbf{a}) \sum_{j=1}^{n} c_j \phi_j(\mathbf{x}) dx + \int_{[a,b]-Z(\epsilon)} |r(\mathbf{x},\mathbf{a})| dx \\ &\quad - \gamma \int_{[a,b]-Z(\epsilon)} \theta(\mathbf{x},\mathbf{a}) \sum_{j=1}^{n} c_j \phi_j(\mathbf{x}) dx \\ &\quad + \gamma \int_{Z(\epsilon)-Z} \left| r(\mathbf{x},\mathbf{a}) - \gamma \sum_{j=1}^{n} c_j \phi_j(\mathbf{x}) \right| dx.\end{aligned}$$

Si escogemos un $\gamma > 0$ suficientemente pequeño, tal que $0 < \gamma < G$ entonces

$$\text{signo}\left(r(\mathbf{x},\mathbf{a}) - \gamma \sum_{j=1}^{n} c_j \phi_j(\mathbf{x}) \right) = \theta(\mathbf{x},\mathbf{a}) \quad x \in [a,b] - Z(\epsilon).$$

Por tanto, si $\gamma < G$,

$$\|r(\mathbf{x},\mathbf{a}+\gamma\mathbf{c})\| \leq -\gamma\delta + \|r(\mathbf{x},\mathbf{a})\| + \int_{Z(\epsilon)-Z} \left\{ \left| r(\mathbf{x},\mathbf{a}) - \gamma \sum_{j=1}^{n} c_j\phi_j(\mathbf{x}) \right| - |r(\mathbf{x},\mathbf{a})| + \gamma\theta(\mathbf{x},\mathbf{a}) \sum_{j=1}^{n} c_j\phi_j(\mathbf{x}) \right\} dx.$$

El integrando está acotado por el módulo $4\epsilon = 4\gamma M$. Entonces

$$\|r(\mathbf{x},\mathbf{a}+\gamma\mathbf{c})\| \geq \|r(\mathbf{x},\mathbf{a})\| + \gamma(4M\mu(Z(\epsilon)-Z) - \delta)$$

donde $\mu(Z(\epsilon)-Z)$ denota la medida de Lebesgue de $Z(\epsilon)-Z$, que tiende a cero cuando $\epsilon \to 0$. Si $\gamma < G$ es tal que

$$4M\mu(Z(\epsilon)-Z) < \delta$$

tenemos la contradicción de que $\mathbf{a}$ reselve (3.1) y terminamos la demostración. □

Corolario 3.1.1. *Supongamos que* $\boldsymbol{p} = \sum_i^n \alpha_i\phi_i(x) \in A$ *es tal que* $\mu(Z) = 0$. *Entonces* $\boldsymbol{p}$ *resuelve (3.1) si y solo si*

$$\int_{[a,b]} \theta(\boldsymbol{x}, \boldsymbol{a})\phi_j(\boldsymbol{x})dx = 0, \quad j = 1, 2, ..., n. \tag{3.8}$$

Este corolario es importante, ya que en general podríamos esperar que $\mu(Z) = 0$; por ejemplo, esto es cierto si $f(x) - p(x)$ se anula solo en un número finito de puntos en $[a, b]$.

En general, si $\mu(Z) > 0$, entonces la caracterización de aproximación óptima puede formularse en los siguientes términos:

Teorema 3.1.2. *Sea A un subespacio vectorial de* $C[a, b]$. *Sea f cualquier función en* $C[a, b]$ *y sea p cualquier elemento de A, de manera que el conjunto*

$$I = \{x : f(x) = p^*(x), a \leq x \leq b\} \tag{3.9}$$

esté vacío o esté compuesto por un número finito de intervalos y puntos discretos. Entonces, p es una mejor aproximación L_1 *desde A hacia f, si y solo si se satisface la desigualdad*

$$\left| \int_a^b s^*(x) \cdot p(x)dx \right| \leq \int_I |p(x)|dx \tag{3.10}$$

para todos los p en A, donde $s^* = signo(f - p^*)$ *es la función (3.2).*

Demostración. Si la condición (3.10) no se cumple para todas las funciones p en A, tomamos p como un elemento de A tal que el número

$$\mu = \int_a^b s^*(x) \cdot p(x)dx - \int_I |p(x)|dx \tag{3.11}$$

sea positivo y la condición de norma

$$\|p\|_\infty = 1 \tag{3.12}$$

se cumpla. Demostramos que p^* no es una mejor aproximación L_1 desde A hacia f al mostrar que, si el número θ es suficientemente pequeño y positivo, entonces la desigualdad

$$\|f - (p^* + \theta p)\|_1 < \|f - p^*\|_1 \tag{3.13}$$

se obtenga. El límite superior de θ depende del conjunto

$$I_\theta = \{x : 0 < |f(x) - p^*(x)| \leq \theta, a \leq x \leq b\}. \tag{3.14}$$

Necesitamos que θ sea tan pequeño que la condición

$$\int_{I_\theta} dx < \frac{1}{2}\mu \tag{3.15}$$

se cumpla, lo cual es posible debido a las restricciones impuestas en I que se dan en el enunciado del teorema. Dejamos que I_R sea el conjunto que contiene los puntos de $[a, b]$ que no están ni en I ni en I_θ. La desigualdad (3.13) se demuestra dividiendo el rango de integración en la definición

$$\|f - (p^* + \theta p)\|_1 = \int_a^b |f(x) - p^*(x) - \theta p(x)|dx \tag{3.16}$$

en tres partes, I, I_θ y I_R. La definición (3.9) nos da esta igualdad

$$|f(x) - p^*(x) - \theta p(x)| = \theta|p(x)|, \tag{3.17}$$

la condición (3.12) ofrece la cota

$$|f(x)-p^*(x)-\theta p(x)| \leq |f(x)-p^*(x)|+\theta|p(x)| \leq |f(x)-p^*(x)|+\theta[2-s^*(x)\cdot p(x)], \tag{3.18}$$

con $x \in I_\theta$ y las ecuaciones (3.12) y (3.14) implica que, cuando x está en I_R el signo de $\{f(x) - p^*(x) - \theta p(x)\}$ es el mismo que el de $\{f(x) - p^*(x)\}$ lo cual nos da

$$|f(x) - p^*(x) - \theta p(x)| = |f(x) - p^*(x)| - \theta \cdot s^*(x) \cdot p(x) \quad x \in I_R. \tag{3.19}$$

Por lo tanto, se deduce de las ecuaciones (3.2) y (3.16) que se obtiene la condición

$$\|f-(p^*+\theta p)\|_1 \leq \|f-p^*\|_1+\theta \int_I |p(x)|dx-\theta \int_a^b s^*(x)\cdot p(x)dx+2\theta \int_{I_\theta} dx. \tag{3.20}$$

La desigualdad (3.13) es ahora una consecuencia de las expresiones (3.11) y (3.15), lo que demuestra la primera mitad del teorema. Para demostrar la segunda parte del teorema, tomamos q como un elemento general de A, luego, dejamos que p sea la función $(p^* - q)$ que también está en A y deducimos a partir de la desigualdad (3.10) que la distancia $\|p^* - q\|_1$ no es menor que la distancia $\|f - p*\|_1$. Específicamente, de las expresiones (3.2), (3.9) y (3.10) obenemos la relación

$$\begin{aligned}\int_a^b |f(x) - q(x)|dx &\geq \int_a^b s^*(x)[f(x) - q(x)]dx + \int_I |f(x) - q(x)|dx \\ &= \int_a^b s^*(x)[f(x) - p^*(x)]dx + \int_a^b s^*(x)[p^*(x) - q(x)]dx + \\ &\quad + \int_I |p^*(x) - q(x)|dx \\ &= |f - p^*|_1 + \int_a^b s^*(x) \cdot p(x) + \int_I |p(x)|dx \\ &\geq \|f - p^*\|_1\end{aligned} \tag{3.21}$$

donde la primera línea depende de la propiedad $\{s^*(x) = 0, x \in I\}$. Debido a que esta desigualdad muestra que q no es una mejor aproximación L_1 que p^*, por tanto el teorema queda demostrado.

□

Notemos que cuando el conjunto I definido por la ecuación (3.9) contiene solo un número finito de puntos discretos. En este caso, debido a que el lado derecho de la expresión (3.10) es cero, p^* es una mejor aproximación L_1 a f en A si y solo si se cumple la condición

$$(s^*, p) = 0, \quad p \in A \tag{3.22}$$

se cumple, donde s^* es la función (3.2), y donde (s^*, p) es el producto escalar

$$(s^*, p) = \int_a^b s^*(x) \cdot p(x)dx. \tag{3.23}$$

3.1.1. Unicidad y la condición de Haar

Si el subespacio vectorial $A = \langle \phi_1, ..., \phi_n \rangle$ satisface la condición de Haar es posible probar la unicidad de la aproximación óptima en L_1 a una función f. Esta propiedad generaliza a subespacios $A = \langle \phi_1, ..., \phi_n \rangle$ propiedades específicas que se dan con los polinomios de grado menor o igual que n.

Es útil expresar el teorema en una forma que aplique a una clase de funciones que incluye a los polinomios como un caso especial. La forma habitual de

definir esta clase es identificar las propiedades de dichos polinomios. Estas propiedades hacen referencia a si cumplen o no la condición de Haar.

Definición 3.1.1. *Sea A un subespacio vectorial de dimensión n de $C[a,b]$, entonces se dice que A satisface la condición de Haar:*

1. *Si ϕ es cualquier elemento de A que no es identicamente cero, entonces el número de raíces de la ecuación $\{\phi(x) = 0; a \leq x \leq b\}$ es a lo sumo n.*

2. *Si $\{\phi_i; i = 1, ..., n\}$ es cualquier base de A y si $\{\xi_i; j = 1, ..., n\}$ es cualquier cualquier conjunto de n puntos distintos en $[a,b]$, entonces el determinante de la matriz de dimensión $n \times n$ que tiene por elementos $\{\phi_i(\xi_j); i = 1, ..., n; j = 1, ..., n\}$ es no nulo.*

Ahora vamos a demostrar que se cumple la equivalencia entre ambas condiciones.

Demostración. Supongamos que se cumple la condición (**1**) pero no la condición (**2**). Entonces existen n puntos distintos $\{\xi_j; j = 1, ..., n\}$ en $[a,b]$, tales que la matriz $\{\phi_i(\xi_j); i = 1, ..., n; j = 1, ..., n\}$ tiene determinante nulo siendo $\{\phi_i; i = 1, ..., n\}$ una base de A. Por lo tanto existen escalares $\{\lambda_i; i = 1, ..., n\}$ que no son todos ceros y que cumplen con

$$\sum_{i=1}^{n} \lambda_i \phi_i(\xi_j) = 0, \quad j = 1, ..., n. \tag{3.24}$$

Por tanto la función

$$\phi(x) = \sum_{i=1}^{n} \lambda_i \phi_i(x) \quad a \leq x \leq b, \tag{3.25}$$

tiene ceros en los puntos $\{\xi_j; j = 1, ..., n\}$, lo que contradice la condición (**1**). Por otra parte si la condición (**1**) no se cumple, entonces existe una función (3.25) que es idénticamente nula, y los ceros están en los puntos $\{\xi_j; j = 1, ..., n\}$. Entonces la ecuación (3.24) se cumple, lo que implica que la matriz $\{\phi_i(\xi_j); i = 1, ..., n; j = 1, ..., n\}$ tiene determinante nulo. Por tanto es una contradicción si la condición (**1**) no se cumple pero la (**4**) si, lo que completa la demostración de que (**1**) y (**4**) son equivalentes.

□

Dos aplicaciones del Teorema 3.1.2 son las siguientes. La primera parte de la demostración del teorema proporciona un método constructivo para obtener una aproximación a f en A que es mejor que p^* si la condición (3.10) no se satisface. En segundo lugar, el teorema a veces se puede utilizar para calcular directamente la mejor aproximación.

Teorema 3.1.3. *Sea A un subespacio vectorial de dimensión n en $C[a,b]$ que cumple con la condición de Haar, y sea f cualquier función en $C[a,b]$. Si p^* es una mejor aproximación L_1 a f en A, y si el número de ceros de la función de error*

$$e^*(x) = f(x) - p^*(x), \quad a \leq x \leq b \tag{3.26}$$

es finito, entonces e^* cambia de signo al menos n veces.

Demostración. Supongamos que e^* tiene un número finito de ceros y que cambia de signo menos de $(n+1)$ veces. Entonces, según la propiedad (**2**) vista previamente, existe una función p en A tal que el producto $s(x) \cdot p(x)$ es positivo para todos los valores de $x \in [a,b]$, excepto en los ceros de e^*, donde s^* es la función (3.2). Por lo tanto, la integral (3.23) es positiva, pero el lado derecho de la expresión (3.10) es cero, porque I tiene medida cero. Por lo tanto, p^* no satisface el teorema de caracterización. Esta contradicción demuestra finalmente el teorema.

□

Ahora vamos a presentar un resultado muy importante, el **Teorema de unicidad**.

Teorema 3.1.4. *Sea A un subespacio vectorial de dimensión finita de $C[a,b]$ que cumple con la condición de Haar. Entonces, para cualquier función f en $C[a,b]$, existe una única mejor aproximación L_1 a f en A.*

Demostración. Sean q^* y r^* las mejores aproximaciones a f en A, y sea p^* la función $\frac{1}{2}(q^* + r^*)$. Consideramos la desigualdad

$$\begin{aligned}\int_a^b |f(x) - p^*(x)|dx &= \int_a^b |\frac{1}{2}[f(x) - q^*(x)] + \frac{1}{2}[f(x) - r^*(x)|dx \\ &\leq \frac{1}{2}\int_a^b |f(x) - q^*(x)|\, dx + \frac{1}{2}\int_a^b |f(x) - r^*(x)|dx \\ &= \frac{1}{2}\|f - q^*\|_{L_1} + \frac{1}{2}\|f - r^*\|_{L_1} \\ &= d(f, A).\end{aligned} \tag{3.27}$$

Dado que el lado derecho es la distancia mínima a f en A, y porque p^* está en A, esta desigualdad se cumple como una ecuación. Por lo tanto, debido a que todas las funciones están en $C[a,b]$, la identidad

$$|f(x) - p^*(x)| = \frac{1}{2} \cdot |f(x) - q^*(x)| + \frac{1}{2} \cdot |f(x) - r^*(x)| \tag{3.28}$$

se cumple para todos los $x \in C[a,b]$. En particular, cuando $f(x)$ es igual a $p^*(x)$, entonces tanto $q^*(x)$ como $r^*(x)$ deben ser iguales a $f(x)$. Se sigue del Teorema 3.1.3 que la función

$$q^*(x) - r^*(x), \quad a \leq x \leq b$$

tiene al menos n ceros. Por lo tanto, la condición de Haar implica que las funciones $q^*(x)$ y $r^*(x)$ son iguales.

□

Teorema 3.1.5. *Supongamos que A es un subespacio vectorial de dimensión n en $C[a,b]$ que cumple la condición de Haar, y sea f una función en $C[a,b]$ tal que la función de error (3.26) tiene exactamente n ceros, donde p^* es la mejor aproximación L_1 a f en A. Entonces, las posiciones de los ceros no dependen de f.*

Demostración. Sea s^* la función (3.2) y consideremos los ceros de la función de error $\{f(x)-p^*(x); a \leq x \leq b\}$ en los puntos $\{\xi_i; i = 1, ..., n\}$. Sea g una función en $C[a,b]$ tal que la función error

$$d^*(x) = g(x) - q^*(x), \quad a \leq x \leq b \tag{3.29}$$

tiene n ceros, donde q^* es la mejor aproximación L_1 de A a f. Supongamos que los ceros se encuentren en los puntos $\{\eta_i; i = 1, ..., n\}$, y sea

$$t^* = \begin{cases} -1 & si \quad g(x) < q^*(x). \\ 0 & si \quad g(x) = q^*(x). \qquad a \leq x \leq b \\ 1 & si \quad g(x) > q^*(x). \end{cases} \tag{3.30}$$

tenemos que demostrar que los conjuntos $\{\xi_i; i = 1, ..., n\}$ y $\{\eta_i; i = 0, 1, ..., n\}$ son el mismo. Por las condiciones de Haar y por el Teorema 3.1.2 tenemos las ecuaciones

$$\int_a^b s^*(x)p(x)dx = \int_a^b t^*(x) \cdot p(x)dx = 0, \quad p \in A. \tag{3.31}$$

También necesitamos dos consecuencias del Teorema 3.1.3, a saber, que las funciones de error (3.26) y (3.29) cambian de signo en sus ceros, y que $e^*(a)$ y $d^*(a)$ son distintos de cero.
Asumimos sin pérdida de generalidad que $\xi_0 \leq \eta_0$, y que los signos de $e^*(a)$ y $d^*(a)$ son iguales. Debido a las condiciones de Haar tenemos que p es una función en A que cambia de signo en los puntos $\{\xi_i; i = 1, ..., n\}$ y que no tiene otros ceros. Elegimos el signo general de p de manera que los signos de $p(a)$ y $e^*(a)$ sean opuestos. Consideramos la ecuación

$$\int_a^b [s^*(x) - t^*(x)] \cdot p(x)dx = 0 \tag{3.32}$$

que se sigue de la condición (3.31). El signo del integrando es importante. Nuestras suposiciones indican que $[s^*(x) - t^*(x)]$ es igual a cero cuando x está en el intervalo $[a, \xi_0)$. Además, en el intervalo $(\xi_0, b]$, el producto $s^*(x)p(x)$ es positivo, excepto en un conjunto de medida cero, es decir, en el conjunto de puntos

$\{\eta_i; i = 1, ..., n\}$.

Además, las definiciones (3.2) y (3.30) muestran que, si $s^*(x)p(x)$ es positivo, entonces el producto $[s^*(x) - t^*(x)] \cdot p(x)$ nunca es negativo negativo. Sabiendo todo esto, deducimos que la desigualdad

$$[s^*(x) - t^*(x)] \cdot p(x) \geq 0, \quad a \leq x \leq b \tag{3.33}$$

se cumple. La ecuación (3.32) implica que la función $\{s^*(x) - t^*(x); a \leq x \leq b\}$ es igual a cero en casi todos los puntos. Por tanto los conjuntos $\{\xi_i; i = 1, ..., n\}$ y $\{\eta_i; i = 1, ..., n\}$ son iguales.

□

Es de interés el concepto de **unicidad fuerte** que es más restrictivo que el de la unicidad.

Definición 3.1.2. *Sea $A \subset C[a, b]$ (no es necesariamente un subespacio) y dado $f \in C([a, b])$. Sea $p^* \in A$ una aproximación óptima a f en A, es decir tal que*

$$\|f - p^*\| \leq \|f - p\|, \quad \forall p \in A.$$

Si existe $\gamma > 0$ tal que para todo $p \in A$

$$\gamma \cdot \|p - p^*\| \leq \|f - p\| - \|f - p^*\|$$

se dice que p^* es una aproximación óptima fuertemente única.

Nótese que en particular la desigualdad anterior implica la unicidad de la aproximación óptima $p^* \in A$.

El siguiente ejemplo ilustra cómo la condición de Haar de un subespacio de dimensión finita A no es suficiente para obtener la unicidad fuerte de la aproximación óptima.

Ejemplo: Sea $x = [0, 1]$, $n = 1$, $f(x) = x^2$, $\phi_1(x) = 1$

$$\begin{aligned}
\|r(x, a)\| &= \int_0^1 |x^2 - a| dx \\
&= \int_0^{\sqrt{a}} (a - x^2) dx + \int_{\sqrt{a}}^1 (x^2 - a) dx \quad \text{si } \ 0 \leq \sqrt{a} \leq 1 \\
&= \frac{4}{3}|a|^{3/2} - a + \frac{1}{3}.
\end{aligned}$$

Esta expresión se minimiza al elegir $a = \dfrac{1}{4}$, con $\|r\| = \dfrac{1}{4}$, y claramente este es el valor mínimo de la norma. Ahora, para cualquier c tal que $\sqrt{c} \in \left[\dfrac{1}{2}, 1\right]$

$$\frac{\|r(x, c)\| - \dfrac{1}{4}}{|c - 1/4|} = \frac{\dfrac{4}{3}c^{3/2} - c + \dfrac{1}{12}}{\dfrac{1}{4}} \to 0 \qquad \text{como } \sqrt{c} \to \frac{1}{2}.$$

Por lo tanto $a = \frac{1}{4}$ no es una solución única de (3.1).

Concluimos esta sección considerando brevemente la situación cuando el requisito de continuidad en (3.1) es menor, de modo que $C[a,b]$ se reemplaza por el espacio $L^1[a,b]$ de funciones integrables de Lebesgue. Este caso es examinado en detalle por Kripke y Rivlin (1965), quienes demuestran, por ejemplo, que el importante corolario del Teorema 3.1.1 todavía se mantiene. Sin embargo, el Teorema 3.1.4 ya no es válido, en el sentido de que existen funciones $f \in L_1([a,b])$, para los que aún con A subespacio de dimensión finita satisfaciendo la condición de Haar, no se tiene unicidad de la aproximación óptima.

El siguiente ejemplo se ilustra esta situación:

Ejemplo: Sea $x = [-1,3]$, $n = 2$, $\phi_1 = 1$, $\phi_2(x) = x$

$$f(x) = \begin{cases} 0 & si \quad -1 \leq x \leq 2 \\ \\ -1 & si \quad 2 < x < 3. \end{cases}$$

Entonces los polinomios $p(x) = tx$ con $-\frac{1}{3} \leq t \leq 0$ son todos aproximaciones óptimas de f,

$$\int_{-1}^{3} 1 \cdot \text{signo}(f(x) - tx)dx = 0,$$

$$\int_{-1}^{3} x \cdot \text{signo}(f(x) - tx)dx = 0$$

y tx es la mejor aproximación para todo t tal que $-\frac{1}{3} \leq t \leq 0$.

3.2. La construcción de las mejores aproximaciones

El problema del cálculo real de las mejores aproximaciones L_1 a funciones continuas, es de interés teórico más que práctico, y no es uno que haya recibido mucha atención, al menos en general. Solo nos ocuparemos del caso particular de aproximación de funciones continuas en $[a,b]$ por elementos de $A = \langle \phi_1, ...\phi_n \rangle$, ya que esto permite utilizar algunos aspectos especiales de la teoría. Hemos visto que para el problema discreto L_1, existe una relación cercana con un problema de interpolación adecuado, y el Teorema 3.1.3 sugiere que una conexión similar puede establecerse en el caso continuo.

Suponemos que tenemos un conjunto de puntos $a = x_0 < x_1 < ... < x_n < x_{n+1} = b$ y la función signo $s(x)$ satisface que $|s(x)| = 1$, $x \in (x_i, x_{i+1})$, el cual puede cambiar de valor en cada punto x_i, $i = 1, 2, ...n$ y en ningún otro punto

en $[a, b]$. Supongamos además que

$$\int_a^b s(x)\phi_j(x)dx = 0, \quad j = 1, 2, ..., n. \tag{3.34}$$

Ahora tomamos $\mathbf{a} \in \mathbb{R}^n$ que satisfazca las condiciones de interpolación

$$\sum_{j=1}^{n} a_j\phi_j(x_i) = f(x_i), \quad i = 1, 2, ..., n. \tag{3.35}$$

Entonces, si $s(x) = \theta(x, \mathbf{a})$, se sigue de (3.34) y del Corolario del Teorema 3.1.1 que $\mathbf{a}$ resuelve (3.1). Por supuesto, no hay garantía de que una solución para (3.35) sea tal que $s(x) = \theta(x, \mathbf{a})$, por lo que incluso si se pudiera obtener un $s(x)$ apropiado que satisfaga (3.34), no necesariamente llevará directamente a una solución de (3.1).

Cuando $\phi_j(x), j = 1, 2, ..., n$ cumple las condiciones de Haar en el intervalo $[a, b]$. Entoncesla función signo es única y tiene exactamente n cambios de signo. En el caso de la aproximación por polinomios de grado menor o igual que $n-1$ de una función continua $f \in C([-1, 1])$, la función signo debe satisfacer

$$\int_{-1}^{1} x^k s(x)dx = 0, \quad k = 0, 1, ..., n-1. \tag{3.36}$$

Demostraremos que un $s(x)$ apropiado se obtiene tomando como

$$s(x) = \text{signo}\,(U_n(x))$$

donde $U_n(x)$ es el polinomio de Chebyshev de segunda clase de grado n, definido por

$$U_n(x) = \frac{\sin(n+1)\theta}{\sin\theta}$$

donde $x = \cos\theta, \quad \theta \in [0, \pi]$. Tenemos entonces

$$\begin{aligned}
\int_{-1}^{1} x^k \text{signo}\,(U_n(x))\,dx &= \int_0^{\pi} (\cos\theta)^k \text{signo}\left(\frac{\sin(n+1)\theta}{\sin\theta}\right)\sin\theta d\theta \\
&= \int_0^{\pi} (\cos\theta)^k \sin\theta \text{signo}(\sin(n+1)\theta)d\theta \\
&= \frac{1}{2}\int_{-\pi}^{\pi} (\cos\theta)^k \sin\theta \cdot \text{signo}(\sin(n+1)\theta)d\theta.
\end{aligned}$$

Como

$$(\cos\theta)^k \sin\theta = \frac{(e^{i\theta} + e^{-i\theta})^k \cdot (e^{i\theta} + e^{-i\theta})}{2^{k+1}i} = \sum_{|p|\leq k+1} \alpha_p e^{ip\theta}$$

para ciertas constantes α_p. El resultado requerido (3.36) se cumple siempre que

$$I_p = \int_{-\pi}^{\pi} e^{ip\theta}\text{signo}(\sin(n+1)\theta)d\theta = 0, \quad p = 0, 1, ..., \pm n.$$

Tomamos $\theta = \phi + \dfrac{\pi}{n+1}$, obtenemos

$$\begin{aligned} I_p &= \int_{-\pi-\pi/(n+1)}^{\pi-\pi/(n+1)} e^{ip(\phi+\pi/(n+1))} + \text{signo}\left(\sin(n+1)(\phi+\pi/(n+1))\right) d\phi \\ &= -e^{ip\pi/(n+1)} \int_{-\pi-\pi/(n+1)}^{\pi-\pi/(n+1)} e^{ip\phi}\text{signo}\left(\sin(n+1)\phi\right) d\phi \\ &= -e^{ip\pi/(n+1)} \cdot I_p \end{aligned}$$

ya que el integrando es periódico, concretamente de periodo 2π.
Como $-e^{ip\pi/(n+1)} \neq 1, |p| \leq n$, obtenemos que $I_p = 0, |p| \leq n$, Y así hemos demostrado que

$$\int_{-1}^{1} x^k \text{signo}\left(U_n(x)\right) dx = 0, \quad k = 0, 1, ..., n-1. \tag{3.37}$$

Ahora, los ceros de $U_n(x)$ en $[-1, 1]$ son los mismos que los de $\dfrac{\sin(n+1)\theta}{\sin\theta}$.
Dado que $\sin\theta = 0$ en $\theta = 0$, los ceros se obtienen en los valores de π donde $U_n(x) \neq 0$, y están dados por:

$$(n+1)\theta_k = k\pi, \quad k = 1, 2, ..., n$$

o por

$$x_k = \cos\theta_k = \cos\frac{k\pi}{n+1}, \quad k = 1, 2, ..., n. \tag{3.38}$$

Por lo tanto, si $\mathbf{a} \in \mathbb{R}^n$ es tal que

$$p(x_k) \equiv \sum_{i=1}^{n} a_i x_k^{i-1} = f(x_k), \quad k = 1, 2, ..., n \tag{3.39}$$

sin ningún otro punto de interpolación en $[-1, 1]$

$$\text{signo}(p(x) - f(x)) = \pm\text{signo}(U_n(x))$$

y por (3.37)

$$\int_{-1}^{1} x^k \text{signo}\left(p(x) - f(x)\right) dx = 0, \quad k = 0, 1, ..., n-1.$$

En otras palabras, si $p(x) - f(x)$ es igual a cero solo en los puntos $\{x_k, k = 1, 2, ..., n\}$ definidos por (3.38), entonces $p(x)$ es la mejor aproximación L_1 por

un polinomio de grado $(n-1)$ a $f(x)$ en $[-1, 1]$. Este resultado es notable porque los puntos x_k están fijos e independientes de $f(x)$.

Para el tratamiento de problemas de aproximación L_1 más generales, es esencial hacer la suposición de que $\mu(Z) = 0$ en la solución, de modo que el problema se reduce a encontrar $\mathbf{a} \in \mathbb{R}^n$ tal que satisfaga

$$\int_a^b \theta(x, \mathbf{a})\phi_j(x)dx = 0, \quad j = 1, 2, ..., n. \tag{3.40}$$

Esto es simplemente un sistema de n ecuaciones no lineales en n incógnitas, aunque de una forma algo inusual y poco conveniente, por lo que es poco probable que un enfoque directo para su solución sea exitosa. Sin embargo, si se sabe que en la solución, $r(x, \mathbf{a})$ cambia de signo exactamente **m** veces en **[a, b]**, entonces (3.40) puede ser reemplazado de manera exitosa por las ecuaciones:

$$\sum_{i=1}^{m+1} (-1)^j \int_{x_{i-1}}^{x_i} \phi_j(x)dx = 0, \quad j = 1, 2, ..., n,$$

$$\sum_{j=1}^{n} a_j\phi_j(x_i) - f(x_i) = 0, \quad i = 1, 2, ..., m,$$

donde $x_0 = a$ y x_{m+1}. Este es un sistema de $(n + m)$ ecuaciones no lineales en las $(n + m)$ incógnitas $\mathbf{a}, x_1, x_2, ..., x_m$, de una forma más manejable que la ecuación (3.40).
El valor de m y las aproximaciones a las incógnitas pueden obtenerse resolviendo una discretización del problema (3.1), donde los métodos del capítulo anterior están disponibles.

Otra forma posible de resolver (3.1) es mediante un método de descenso. Si $\mathbf{a}$ es una aproximación a la solución que no satisface (3.40), entonces se puede lograr una reducción en el valor de la norma mediante un paso lo suficientemente pequeño en cualquier dirección $\mathbf{c}$ que cumpla con

$$\sum_{j=1}^{n} c_j \int_X \upsilon(\mathbf{x})\phi_j(x)dx \geq \delta > 0, \quad \forall \upsilon(\mathbf{x}) \in V. \tag{3.41}$$

En este caso, lo podemos escribir como

$$\sum_{j=1}^{n} \int_a^b \theta(x, \mathbf{a})\phi_j(x)dx \geq \delta > 0. \tag{3.42}$$

Si se asume que $\mu(Z) = 0$ en la aproximación actual, entonces podemos elegir, por ejemplo,

$$c_j = \int_a^b \theta(x, \mathbf{a})\phi_j(x)dx, \quad j = 1, 2, ..., n,$$

lo que solo dejará de definir una dirección de descenso si $\mathbf{a}$ es una solución.

3.3. Discretización

Para las mejores aproximaciones de Chebyshev en un conjunto continuo X, se le da una gran importancia a los problemas discretos correspondientes obtenidos al reemplazar X por un conjunto discreto de puntos en X. En el caso de L_1, la relación entre estos problemas no es tan cercana.

Limitaremos el problema al caso en el que $X = C[a, b]$ y consideraremos aproximaciones en conjuntos discretos $Y = \{x_1, x_2, ..., x_m\}$ elegidos de manera que $\{x_i \in [y_{i-1}, y_i], i = 1, 2, ...m\}$, donde $a \leq y_o < y_1 < ... < y_m \leq b$.

Sea $h = \text{máx}_{1\leq i\leq m}\, h_i$ donde $\{h_i = y_i - y_{i-1}, i = 1, 2, ..., m\}$ y sea

$$\Omega(\epsilon) = \max_{1\leq i\leq m} \sup_{\|\mathbf{x}-\mathbf{y}\|_2\leq\epsilon} |\phi_i(\mathbf{x}) - \phi_i(\mathbf{y})|$$

y

$$\omega(f, h) = \sup_{|\mathbf{x}\text{-}\mathbf{y}|\leq h} |f(\mathbf{x}) - f(\mathbf{y})|.$$

Lema 3.3.1. *Sea $g(x) \in C[a, b]$. Entonces*

$$\left| \int_a^b g(x)dx - \sum_{i=1}^{m} g(x_i)h_i \right| \leq (b - a) \cdot \omega(g, h).$$

Demostración.

$$\begin{aligned} \left| \int_a^b g(x)dx - \sum_{i=1}^{m} g(x_i)h_i \right| &= \left| \sum_{i=1}^{m} \int_{y_i-1}^{y_i} [g(x) - g(x_i)]dx \right| \\ &\leq \sum_{i=1}^{m} \int_{y_i-1}^{y_i} |g(x) - g(x_i)|dx \\ &\leq \omega(g, h) \sum_{i=1}^{m} h_i = (b - a) \cdot \omega(g, h). \end{aligned}$$

□

Ahora definimos

$$M = \min_a \|r(x, \mathbf{a})\|$$

y

$$M(h) = \min_a \sum_{i=1}^{m} |r(x_i, \mathbf{a})|h_i = \sum_{i=1}^{m} |r(x_i, \mathbf{a}(h))|h_i.$$

Teorema 3.3.1. *Sea $h \to 0$. Entonces*

1. *$M(h) \to M$.*
2. *$|r(x, \boldsymbol{a}(h))| \to M$.*

3. Si $\boldsymbol{a}^*$ es la única solución de (3.1), $\boldsymbol{a}(h) \to \boldsymbol{a}^*$.

Demostración. Por continuidad, $\omega(f,h)$ y $\Omega(h)$ tienden a 0 cuando $h \to 0$. sea h tal que $\omega(f,h) \leq |f|/(b-a)$, $\Omega(h) < \delta$, donde $\delta > 0$ satisface:

$$\|\sum_{j=1}^{n} c_j \phi_j(x)\| \geq \delta \left|\sum_{j=1}^{n} c_j\right|$$

para todo $c \in \mathbb{R}^n$. Entonces

$$\begin{aligned} M &\leq \int_a^b |r(x, \mathbf{a}(h))|\, dx \\ &\leq M(h) + (b-a) \cdot \omega\left(|r(x, \mathbf{a}(h))|, h\right) \quad \text{por el Lema 3.3.1} \\ &\leq M(h) + (b-a) \cdot \omega(r(x, \mathbf{a}(h)), h) \\ &\leq M(h) + (b-a) \cdot \left|\sum_{j=1}^{n} a_j(h)\right| \cdot \Omega(h). \end{aligned} \tag{3.43}$$

Además

$$\begin{aligned} \left|\sum_{j=1}^{n} a_j(h)\right| &\leq \frac{1}{\delta} \left\|\sum_{j=1}^{n} a_j(h)\phi_j(x)\right\| \\ &\leq \frac{2}{\delta} \sum_{i=1}^{m} \left|\sum_{j=1}^{n} a_j(h)\phi_j(x_i)\right| h_i \\ &= \frac{2}{\delta} \sum_{i=1}^{m} \left|\sum_{j=1}^{n} a_j(h)\phi_j(x_i) - f(x_i) + f(x_i)\right| h_i \\ &\leq \frac{4}{\delta} \sum_{i=1}^{m} |f(x_i)| h_i \\ &\leq \frac{4}{\delta} \left(\|f\| + (b-a) \cdot \omega(f,h)\right) \\ &\leq \frac{8}{\delta} \|f\|. \end{aligned} \tag{3.44}$$

Por tanto

$$M \leq M(h) + \frac{8(b-a)}{\delta} \|f\| \Omega(h) \tag{3.45}$$

y si $\mathbf{p}^*$ es una solución de (3.1)

$$M = \sum_{i=1}^{m} |f(x_i) - \mathbf{p}^*(x_i)| h_i + E(h).$$

Donde $E(h) \to 0$ si $h \to 0$. Así pues

$$M \geq M(h) + E(h). \tag{3.46}$$

Por lo tanto, las ecuaciones (3.45) y (3.46) proporcionan la afirmación (1). Las desigualdades de (3.43) luego dan lugar a la afirmación (2), ya que las desigualdades de (3.44) muestran que $\{\mathbf{a}(h)\}$ está acotado. Finalmente, los puntos límite, según la afirmación (2), deben resolver (3.1), y de aquí se obtiene la afirmación (3).

□

El Teorema 3.3.1 muestra que se puede obtener una buena aproximación del valor mínimo de la norma en (3.3) mediante uno de los métodos del Capítulo 2. Para que dicho método sea eficiente, nos gustaría ser más selectivos en la elección del conjunto discreto en cada etapa, de manera similar a como se hace en el caso L_∞. Sin embargo, el enfoque a seguir ya no es tan directo. Una posible manera de abordar este problema sería construir una secuencia de problemas discretos, cuyas soluciones estén determinadas por n puntos de interpolación tomados del conjunto discreto.

Estos puntos deberían converger hacia el conjunto adecuado de puntos de interpolación en el intervalo $[a, b]$, los cuales caracterizan esencialmente la solución del problema continuo. Sin embargo, cualquier método de este tipo probablemente requerirá suponer que se cumple una condición de las condiciones de Haar. Además, no existe una forma obvia de colocar los nuevos puntos en cada etapa para garantizar la convergencia. Mientras estas dificultades fundamentales no se resuelvan, el problema de desarrollar un método general efectivo para resolver (3.1) sigue sin resolverse.

En cálculos de ajuste de datos, cuando buscamos el elemento de A que minimiza la expresión (3.1), hay un teorema de caracterización similar al Teorema (3.1.2). Este teorema se presenta de tal manera que permite asignar diferentes ponderaciones a los valores de la función $\{f(x_t); t = 1, 2, ...m\}$.

Teorema 3.3.2. *Supongamos que los valores de* $\{f(x_t); t = 1, 2, ...m\}$ *y las funciones peso* $\{w_t; t = 1, 2, ...m\}$ *están dados. Tomemos A como un espacio vectorial de funciones definidas en el conjunto de puntos* $\{x_t; t = 1, 2, ...m\}$. *Sea* p^* *cualquier elemento de A y L tiene los puntos de* $\{x_t; t = 1, 2, ...m\}$ *que satisfacen la condición*

$$p^*(x_t) = f(x_t) \tag{3.47}$$

y sea s^* *la función signo*

$$s^*(x) = \begin{cases} -1 & si \quad f(x_t) < p^*(x_t). \\ 0 & si \quad f(x_t) = p^*(x_t). \qquad t = 1, 2, ..., m. \\ 1 & si \quad f(x_t) > p^*(x_t). \end{cases} \tag{3.48}$$

entonces p^* es la función en A que minimiza la expresión

$$\sum_{t=1}^{m} w_t |f(x_t) - p(x_t)|, \quad p \in A \tag{3.49}$$

si y solo si la desigualdad

$$\left| \sum_{t=1}^{m} w_t s^*(x_t) p(x_t) \right| \leq \sum_{x_t \in L} w_t |p(x_t)| \tag{3.50}$$

se cumple para todo $p \in A$.

Demostración. El método de prueba es similar a la prueba del Teorema (3.1). Si la condición (3.50) no se cumple, consideramos reemplazar la aproximación $p*$ por $(p^* + \theta p)$, donde $|\theta|$ es tan pequeño que, si x_t no está en L, el signo de $\{f(x_t) - p^*(x_t) - \theta p(x_t)\}$ es el mismo que el signo de $s^*(x_t)$. Se sigue que el reemplazo cambia el valor de la expresión (3.49) por

$$-\theta \sum_{t=1}^{m} w_t s^*(x_t) p(x_t) + \theta \sum_{x_t \in L} w_t |p(x_t)|. \tag{3.51}$$

Por lo tanto, si el lado izquierdo de la expresión (3.50) es mayor que el lado derecho, se puede elegir el signo de θ de modo que $(p^* + \theta p)$ sea una mejor aproximación que p^*.

Por otro lado, si se cumple la condición (3.50) para todos los p en A, entonces, al reemplazar las integrales en la expresión (3.20) por sumas con pesos, se deduce que $p*$ es la mejor aproximación discreta de L_1 a los datos.

□

El siguiente teorema muestra que existe una función p^* en A que minimiza la expresión (3.49), y que es tal que el conjunto del Teorema 3.3.2 contiene al menos $(n+1)$ puntos, donde $(n+1)$ es la dimensión de A. Por lo tanto, muchos algoritmos para calcular las mejores aproximaciones discretas L_1 buscan un conjunto que permita obtener una función óptima p^* mediante la interpolación.

Teorema 3.3.3. *Supongamos que los valores de $\{f(x_t); t = 1, 2, ...m\}$ y las funciones peso $\{w_t; t = 1, 2, ...m\}$ están dados. Sea A un subespacio vectorial de R^m donde los componentes de cada vector $\boldsymbol{p}$ en A tienen los valores $\{p(x_t); t = 1, 2, ...m\}$. Entonces existe un elemento p^* en A que minimiza la expresión (3.49) y que tiene la propiedad de que el vector cero es el único elemento p en A que satisface las condiciones $\{p(x_t) = 0; x_t \in L\}$ donde se define el conjunto L como en el Teorema 3.3.2.*

Demostración. Sea p^* la mejor aproximación ponderada L_1 de A a los datos, pero supongamos que existe un elemento no nulo q en A que satisface la condición

$$q(x_t) = 0, \quad x_t \in L \tag{3.52}$$

consideramos la función

$$\psi(\theta) = \sum_{t=1}^{m} w_t |f(x_t) - p^*(x_t) - \theta p(x_t)|, \quad -\infty < \theta < \infty, \tag{3.53}$$

donde θ es una variable real.

Tenemos que (3.53) es una función continua, lineal por partes de θ, que tiende a infinito cuando $|\theta|$ crece, y que toma su valor mínimo cuando θ es cero, porque p^* es la mejor aproximación. Además, la ecuación (3.52) implica que dos segmentos de línea diferentes de ψ no se unen en $\theta = 0$. Por lo tanto, ψ es constante en un entorno de $\theta = 0$ Si θ se incrementa desde cero, entonces $\psi(\theta)$ permanece constante hasta que se alcance un valor de θ que cumpla las condiciones

$$f(x_t) - p^*(x_t) - \theta p(x_t) = 0 \tag{3.54}$$

y

$$q(x_t) \neq 0$$

para algún valor de t. Deja que este valor de θ sea $\hat{\theta}$. Dado que $\psi(\hat{\theta})$ es igual a $\psi(0)$, la función $(p^* + \hat{\theta}q)$ es otra mejor aproximación L_1 ponderada de A a los datos. La ecuación (3.52) implica que los residuos $\{f(x_t) - (p^* + \hat{\theta}q)(x_t)\}$ son cero. Además, otro residuo cero se obtiene de (3.54). Por lo tanto, nuestra construcción aumenta el número de ceros de una mejor aproximación. Dado que la construcción se puede aplicar recursivamente, por tanto, se sigue con que el teorema se cumple.

□

3.4. Puntos de interpolación L_1 para polinomios algebraicos

El Teorema 3.1.5 proporciona el método principal para calcular las mejores aproximaciones L_1 a funciones continuas. Se comienza asumiendo que la función de error cambiará de signo solo n veces. En este caso, debido a que los ceros de la función de error son independientes de f, pueden ser encontrados mediante una consideración detallada de A. Una aproximación a f en A se calcula mediante interpolación en estos ceros, y luego se verifica si su función de error cumple con la hipótesis. Si la hipótesis es válida, entonces se ha encontrado la aproximación que estamos buscando.

Para aplicar el algoritmo para calcular las mejores aproximaciones L_1, descrito

en el párrafo anterior, es necesario identificar los puntos de interpolación que son objeto de estudio del Teorema 3.1.5. Los puntos de interpolación para el caso especial importante cuando A es el espacio P_n se dan en el siguiente teorema.

Teorema 3.4.1. *Sea que las condiciones del Teorema 3.1.5 se cumplen y sea A el espacio P_n y (a, b) es el intervalo $[-1, 1]$. Entonces los ceros de la función de error*

$$e(x) = f(x) - p^*(x), \quad -1 \leq x \leq 1 \tag{3.55}$$

tiene por valores

$$\xi_i = cos\left[\frac{(n+1-i)\pi}{n+2}\right], \quad i = 0, 1, ..., n. \tag{3.56}$$

Demostración. El Teorema 3.1.3 implica que la función de error (3.55) cambia de signo en sus ceros. Por lo tanto, debido al Teorema de Caracterización, es suficiente demostrar que la ecuación

$$\int_{-1}^{1} s^*(x)p(x)dx = 0 \tag{3.57}$$

se cumple para todos los polinomios p en P donde s^* es la función de signo

$$s^*(x) = \begin{cases} 1 & si \quad -1 < x < \xi_0. \\ (-1)^i & si \quad \xi_{i-1} < x < \xi_i. \\ (-1)^{n+1} & si \quad \xi_n < x < 1. \\ 0 & si \quad \text{otro caso.} \end{cases} \qquad i = 1, 2, ..., n. \tag{3.58}$$

Los números $s^*(-1)$ y $s*(1)$ se definen como cero, para que la función

$$\sigma(\theta) = s^*(\cos\theta), \quad 0 \leq \theta \leq \pi \tag{3.59}$$

satisfazca algunas condiciones de periodicidad. Extendemos σ a un rango infinito definiendo $\{\sigma(-\theta) = -\sigma(\theta); 0 \leq \theta \leq \pi\}$ y al dejar σ como una función 2π-periódica. Se deduce de las ecuaciones (3.56) y (3.58) que el gráfico de $\{\sigma(\theta); -\infty < \theta < \infty\}$ es una onda cuadrada que cambia de signo cuando θ es un múltiplo entero de $\dfrac{\pi}{n+2}$. Por lo tanto, se obtiene la condición

$$\sigma\left(\theta + \frac{\pi}{n+2}\right) = -\sigma(\theta), \qquad -\infty < \theta < \infty. \tag{3.60}$$

Se ve que, si se realiza un cambio de variables $\{x = \cos\theta; 0 \leq \theta \leq \pi\}$ en la integral (3.57), entonces la condición (3.60) permite demostrar la ecuación (3.57) cuando p es uno de los polinomios de Chebyshev.

$$T_j(x) = \cos(j\cos^{-1}x), \quad -1 \leq x \leq 1, \quad j = 0, 1, ..., n. \tag{3.61}$$

Dado que estos polinomios son una base de P_n completamos la prueba del teorema estableciendo las ecuaciones

$$\int_{-1}^{1} s^*(x)T_j(x)dx = 0, \quad j = 0, 1, ..., n.$$

La identidad

$$\begin{aligned}\int_{-1}^{1} s^*(x)T_j(x)dx &= \int_0^{\pi} s^*(x)(\cos\theta)\cos(j\theta)\sin\theta dx \\ &= \frac{1}{2}\int_0^{\pi} \sigma(\theta)\{\sin[(j+1)\theta] - \sin[(j-1)\theta]\}d\theta \\ &= \frac{1}{4}\int_{-\pi}^{\pi} \sigma(\theta)\{\sin[(j+1)\theta] - \sin[(j-1)\theta]\}d\theta\end{aligned}$$

se cumple, donde la última línea depende del hecho de que σ es una función impar. Por lo tanto, es suficiente demostrar que las integrales

$$I_k = \int_{-\pi}^{\pi} \sigma(\theta)\sin(k\theta)d\theta, \qquad k = 0, 1, ..., n+1$$

son cero.

Utilizamos la periodicidad del integrando de I_k, luego la condición (3.60) y después el hecho de que σ es una función impar para deducir la ecuación

$$\begin{aligned}I_k &= \int_{-\pi}^{\pi} \sigma\left(\theta + \frac{\pi}{n+2}\right)\sin\left[k\left(\theta + \frac{\pi}{n+2}\right)\right]d\theta \\ &= -\cos\left(\frac{k\pi}{n+2}\right)\int_{-\pi}^{\pi} \sigma(\theta)\sin(k\theta)d\theta - \sin\left(\frac{k\pi}{n+2}\right)\int_{-\pi}^{\pi} \sigma(\theta)\cos(k\theta)d\theta \\ &= -\cos\left(\frac{k\pi}{n+2}\right) I_k, \qquad k = 0, 1, ..., n+1.\end{aligned}$$

Dado que el factor $-\cos\left(\frac{k\pi}{n+2}\right)$ no es igual a uno, se sigue que I_k es cero, lo que proporciona el resultado buscado.

□

Capítulo 4

Experimentación numérica

En este capítulo abordamos una experimentación numérica con una implementación en Matlab del algoritmo de descenso descrito en el capítulo 2. El algoritmo en los términos descritos se debe a R.H Bartels y A. R. Cohn [**2**], pero la implementación que se recoge en el apéndice dista mucho de un software eficiente que contemple más allá del caso no degenerado. Además tampoco optimiza alguno de las herramientas numéricas utilizadas recurriendo en general a funciones de Matlab ya existentes.

Así, por ejemplo, para el cálculo del vector $\mathbf{h}$ de la proyección ortogonal sobre el núcleo de A_Z para determinar la dirección de descenso $\mathbf{c}$ cada ciclo del algoritmo tiene que realizar una factorización QR de A_Z. Esta factorización se implementa llamando en cada iteración a la función qr de Matlab, cuando sería posible un tratamiento más eficiente teniendo en cuenta que de una iteración a otra A_Z solo cambiar en una fila, y por tanto el algoritmo de factorización puede sacar partido de este hecho.

Tampoco consideramos un tratamiento robusto del caso en el que el algoritmo tropieza con puntos muertos degenerados, situación que llevaría a modificar el cálculo de $\mathbf{w}$ descrito en el algoritmo (etapas 2d, 2e, 2f, 2g y 2h descritas en 2.4.5).

4.1. Primer test

Como primer test [**2**] mostramos los resultados del siguiente problema académico

$$A\mathbf{x} = \mathbf{b}$$

con

$$A^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}, \qquad \mathbf{b}^T = \begin{bmatrix} 1 & 1 & 2 & 3 & 2 \end{bmatrix}$$

La solución óptima en el sentido L^1 es $\mathbf{x} = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \end{bmatrix}^T$. Con aproximación inicial $\mathbf{x}_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$, obtenemos la solución óptima en 2 iteraciones.

4.2. Segundo test

Ahora mostramos un ejemplo que se ha realizado de manera personal tomando estas matrices

$$A^T = \begin{bmatrix} 1 & 2 & 3 & 2 & 3 \\ -1 & 1 & -1 & -1 & 0 \\ 1 & 1 & 5 & 1 & 0 \end{bmatrix}, \qquad \mathbf{b}^T = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

y tomando el vector de aproximación inicial es $\mathbf{x}_0 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$ y tras 5 iteraciones obtenemos la solución óptima en el sentido L^1 que es

$$\mathbf{x} = \begin{bmatrix} 0.333333333333333 \\ -0.416666666666667 \\ -0.083333333333333 \end{bmatrix}.$$

Y norma del residuo $\|r\| = 1.1667$.

4.3. Tercer test

El segundo ejemplo [**2**] considera la solución L_1 óptima del problema dado por

$$\mathbf{A} = \begin{bmatrix} 5 & 3 & 4 \\ 9 & 7 & 3 \\ 6 & 6 & 0 \\ 9 & 9 & 7 \\ 3 & 0 & 1 \\ 8 & 1 & 8 \\ 1 & 9 & 8 \\ 3 & 1 & 1 \\ 0 & 9 & 3 \end{bmatrix}, \quad \mathbf{b}^T = \begin{bmatrix} 7 & 4 & 2 & 7 & 7 & 7 & 3 & 5 & 3 \end{bmatrix}.$$

Con solución tras 4 iteraciones

$$\mathbf{x} = \begin{bmatrix} 0.299319727891156 \\ 0.034013605442177 \\ 0.571428571428572 \end{bmatrix}$$

Y norma del residuo $\|r\| = 15.9456$.

4.3.1. Tercer test linealmente dependiente

Tomando como ejemplo de partida el test previo, vamos a añadir dos columnas nuevas, las cuales son linealmente dependientes de las tres primeras columnas.

$$A = \begin{bmatrix} 5 & 3 & 4 & 12 & 4 \\ 9 & 7 & 3 & 19 & 13 \\ 6 & 6 & 0 & 12 & 12 \\ 9 & 9 & 7 & 25 & 11 \\ 3 & 0 & 1 & 4 & 2 \\ 8 & 1 & 8 & 17 & 1 \\ 1 & 9 & 8 & 18 & 2 \\ 3 & 1 & 1 & 5 & 3 \\ 0 & 9 & 3 & 12 & 6 \end{bmatrix}, \qquad \mathbf{b}^T = \begin{bmatrix} 7 & 4 & 2 & 7 & 7 & 7 & 3 & 5 & 3 \end{bmatrix}.$$

Con $x_0 = [0\ 0\ 0\ 0\ 0]^T$ se alcanza la solución óptima en 7 iteraciones siendo

$$\mathbf{x} = \begin{bmatrix} 0.165986 \\ -0.09931972 \\ 0.190476 \\ 0.257142 \\ -0.123809 \end{bmatrix}.$$

Y la norma del residuo $\|r\| = 15.9456$ que coincide con con la norma del test previo.

4.4. Cuarto test

$$A = \begin{bmatrix} 5 & 3 & 4 & 12 & 4 \\ 9 & 7 & 3 & 19 & 13 \\ 6 & 6 & 0 & 12 & 12 \\ 9 & 9 & 7 & 25 & 11 \\ 3 & 0 & 1 & 4 & 2 \\ 8 & 1 & 8 & 17 & 1 \\ 1 & 9 & 8 & 18 & 2 \\ 0 & 9 & 3 & 12 & 6 \\ 3 & 1 & 1 & 5 & 3 \\ 6 & 7 & 6 & 19 & 7 \\ 6 & 1 & 9 & 16 & -2 \\ 0 & 4 & 8 & 12 & -4 \\ 0 & 5 & 7 & 12 & -2 \\ 7 & 3 & 2 & 12 & 8 \\ 5 & 4 & 9 & 18 & 0 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 7 \\ 4 \\ 2 \\ 7 \\ 7 \\ 7 \\ 3 \\ 3 \\ 5 \\ 1 \\ 4 \\ 1 \\ 6 \\ 6 \\ 0 \end{bmatrix}.$$

Este ejemplo añade 6 ecuaciones adicionales a las ecuaciones del ejemplo anterior. Obteniendo tras 7 iteraciones la solución óptima:

$$\mathbf{x} = \begin{bmatrix} 0.224489 \\ -0.0816326 \\ 0.0272108 \\ 0.170068 \\ 0.115646 \end{bmatrix}.$$

Y norma del residuo $\|r\| = 34.2245$.

4.5. Quinto test

Como experimento más exigente en cálculo ilustrando el proceso de discretización de un problema de aproximación polinomial óptima en L_1 por polinomios consideramos $f(x) = e^x, \quad x \in [0, 2]$ [**2**]. Muestreamos la función en abscisas $x = [0.0 : 0.1 : 2.0]$. Buscamos determinar aproximaciones L_1 óptimas discretas por polinomios de grados 1 hasta 7 resolviendo el sistema lineal sobredeterminado con incógnitas en los coeficientes del polinomio de grado 7 en la base polinomial $\{1, x, x^2, .., x^7\}$.

La norma del residuo es $\|r\| = 8.366667$ y obtenemos como solución óptima tras 5 iteraciones

$$\mathbf{x} = \begin{bmatrix} 0.232237 \\ 2.832967 \end{bmatrix}.$$

Tras 7 iteraciones obtenemos la norma del residuo $\|r\| = 1.441155$ y como solución

$$\mathbf{x} = \begin{bmatrix} 1.188057 \\ 0.114670 \\ 1.415553 \end{bmatrix}.$$

Tras 13 iteraciones obtenemos la norma del residuo $\|r\| = 0.190059$ y como solución

$$\mathbf{x} = \begin{bmatrix} 0.983546 \\ 1.206081 \\ 0.054814 \\ 0.468423 \end{bmatrix}.$$

Tras 18 iteraciones obtenemos la norma del residuo $\|r\| = 0.01830$ y como solución

$$\mathbf{x} = \begin{bmatrix} 1.003167 \\ 0.956106 \\ 0.637888 \\ 0.001884 \\ 0.119062 \end{bmatrix}.$$

Tras 33 iteraciones obtenemos la norma del residuo $\|r\| = 0.001591$ y como solución

$$\mathbf{x} = \begin{bmatrix} 0.999647 \\ 1.004982 \\ 0.476336 \\ 0.211976 \\ 0.002040 \\ 0.0232998 \end{bmatrix}.$$

Tras 16 iteraciones obtenemos por norma del residuo $\|r\| = 9.98 \cdot 10^{-5}$ y como solución

$$\mathbf{x} = \begin{bmatrix} 1.000000 \\ 0.99960 \\ 0.503054 \\ 0.158046 \\ 0.0532550 \\ -0.000479 \\ 0.003840 \end{bmatrix}$$

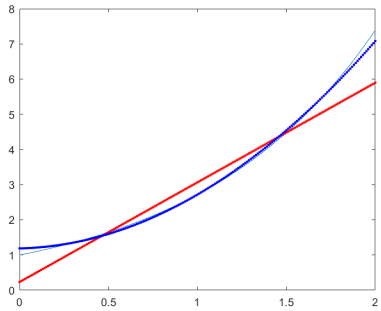

Figura 4.1: Aproximación discreta L1 de grados 1 (en rojo) y 2 (en azul)

Vamos a hacer un resumen en la que indicamos el número de iteraciones necesarios y la norma del residuo obtenida en cada caso, poniendo de manifiesto la reducción de la norma del residuo con el grado de la aproximación polinomial.

Grado	Número de iteraciones	Norma del residuo $\|r\|$
2	5	8.3667(0)
3	7	1.4412(0)
4	13	1.9006(-1)
5	18	1.8303(-1)
6	33	1.5946(-3)
7	16	9.98(-5)

Para poder realizar este experimento hemos usado ciertas funciones de Matlab. Para llevarlo a cabo hemos comenzado definiendo un vector v que esté formado por

$$v = [0.0 : 0.1 : 2.0]$$

para definir los puntos para evaluar la función. También debemos evaluar la función exponencial en dichos puntos ya calculados con

$$b = exp(v)'.$$

Definimos la matriz A

$$A = fliplr(vander(v))$$

y en este caso cómo va a ser hasta el grado 5 tomamos las 5 primeras columnas con

$$A = A(:, 1 : 5)$$

y defino el vector inicial x_0

$$x_0 = [0\,0\,0\,0\,0]'.$$

Una vez definidas estas matrices, ejecutamos nuestro algoritmo que se encuentra desglosado en el apéndice y obtenemos las soluciones nombradas previamente

$$\gg descenso(A, b, x_0).$$

4.6. Sexto test

Ahora vamos a realizar un test con 201 puntos con la función $f(x) = e^{-x}, \quad x \in [0, 2]$ [**2**]. Muestreamos la función en abscisas $x = [0.0 : 0.01 : 2.0]$. Buscamos determinar aproximaciones L_1 óptimas discretas por polinomios de grados 1 hasta 8 resolviendo el sistema lineal sobredeterminado con incógnitas en los coeficientes del polinomio de grado 8 en la base polinomial $\{1, x, x^2, .., x^8\}$. Vamos a hacer un resumen en la que indicamos el número de iteraciones necesarios y la norma del residuo obtenida en cada caso.

Grado	Número de iteraciones	Norma del residuo $\|r\|$
2	6	9.5961(2)
3	9	3.0643(2)
4	12	7.4530(1)
5	19	1.4628(1)
6	21	2.4083(0)
7	24	3.4056(-1)
8	24	4.2337(-2)

Capítulo 5

Apéndice

```
function  [xsol,res,l1norm,iter] = descenso(A,b,x0)
format long;
%
[m,n] = size(A);  % m >> n (sistema lineal
   sobredeterminado).
nb = length(b);
if ((n<1) || (m < 0) || (m < n) || (m~= nb))
    error('Dimensiones incorrectas de la matriz A o
       del vector b o del vector x0 (si se proporciona
       ).')
    return
end
if (nargin < 3)
    x0 = zeros(n,1);    % vector inicial
end
nx = length(x0);
if (nx ~= n)
    error('Dimensiones incorrectas de x0');
    return
end
%
TOLZERO = eps*sqrt(n); % Vamos a definirlo para ver si
    un valor es 0
MXSTEPS = 100;   % Numero maximo de iteraciones.
ZERO = 0.0d0;
%
x = x0;
%
% Inicialiciacion antes de empezar las iteraciones
%
```

```
27 iter = 0;
28 %
29 % Iniciamos el programa
30 %
31 nact = 0;          % Numero maximo de ecuaciones activas
32 iaddc = 0;        % Indice de la ultima ecuacion con
       residuo cero no activa
33 idelc = 0;        % Indice de la columna de AZ que sera
       eliminada
34
35 indx = 1:m;       % Vector con indices de columnas en A
       ^T: indx(1), ..., indx(nact)
36                   % son los indices de las columnas de
                        A^T activas.
37                   % indx(nact+1),...,indx(iaddc) son
                        los indices de las columnas de A^T
38                   % con residuo cero y no activas.
39                   % indx(iaddc+1),..,indx(m) son los
                        indices de las ecuaciones
40                   % con residuos no nulos
41                   % INDX da en cada iteracion la
                        informacion de ecuaciones
42                   % activas, ecuaciones con residuo
                        cero y no activas,
43                   % y ecuaciones con residuo no cero.
44
45 done = false;
46 while (~done)
47     iter = iter+1;
48     %
49     % Preparacion para el proximo la proxima iteracion
           de minimizacion
50     %
51     if (iter > MXSTEPS)
52         error('Se alcanzo el numero maximo de
               iteraciones sin encontrar una solucion.')
53         return
54     end
55     res = A*x - b;              % Residuo inicial
56     l1norm = norm(res,1);
57     mxabsres = max(abs(res));
58     if (mxabsres < TOLZERO)  % x es solucion de A*x=b
           aritmeticamente
59         xsol = x0;
60         l1norm = norm(res,1);
61         return
```

```
62      end
63
64      %
65      % COMPROBANDO RESIDUOS IGUALES A CERO. El test de
           zero residuals compara con
66      % TOLZERO la magnitud del valor absoluto de cada
           componente ix del
67      % residuo dividida por el maximo de las magnitudes
            en valor absoulto
68      % de los terminos A(ij,j)*x(j).
69      %
70      % Es un test bastante estricto del residuo cero.
           Es preferible en costo
71      % computacional ignorar un residuo cero extra que
           aceptarlo  y
72      % provocar sucesiones ciclicas de ecuaciones
           activas (estas se pueden
73      % salvar incorporando aleatoriedad en la eleccion
           de las ecuaciones
74      % activas.
75      %
76      for i = 1:iaddc      % Establecemos en cero los
           residuos que se sabe que
77          ix = indx(i);    % son cero en los indices
               indx(1) ... indx(iaddc)
78          res(ix) = ZERO;
79      end
80      iadp1 = iaddc+1;
81      for i = iadp1 : m    % Viendo nuevos residuos para
            x
82          ix = indx(i);
83          temp = res(ix);
84          test = abs(b(ix));
85          for j = 1 : n
86              prod = abs(A(ix,j)*x(j));
87              if (prod > test)
88                  test = prod;
89              end
90          end
91          test = TOLZERO*test;
92          if (abs(temp) > test)
93              res(ix) = temp;
94          else
95              iaddc = iaddc+1;
96              indx(i) = indx(iaddc);
97              indx(iaddc) = ix;
```

```
98              res(ix) = ZERO;
99          end
100     end
101     %
102     % No se elimina de la lista de ecuaciones activas
           las que
103     % son linealmente dependientes.
104     %
105     sigma = sign(res);    % signos de los residuos
106     indxact = indx(1:nact);
107     iadp1 = iaddc+1;
108     indxresn0 = indx(iadp1:m);
109
110     nresn0= length(indxresn0);     % numero ecuaciones
            con residuo no nulo
111     %
112     % Calculo del gradiente restringido (GRDX = H) o (
           GRDX = P_N (H)),
113     % proyeccion de H
114     %
115     if (nresn0 == 0)
116         h = zeros(1,n);h = h';
117     else
118         h = zeros(1,n);
119         for i = 1:nresn0
120             h = h + sigma(indxresn0(i)) * A(indxresn0(
                   i),:);
121         end
122         h = h';
123     end
124     %
125     % Identificamos las ecuaciones activas y hacemos
           la factorizacion QR a
126     % la matriz AZ: Matriz donde cada columna contiene
            todas las filas activas de A
127     %
128     if (nact == 0)
129         Q = eye(n); R = zeros(n,nact); % factorizacion
                AZ (=0) = Id * R (=0)
130         Q1 = Q(:,1:nact); Q2 = Q(:,nact+1:n);
131         proj = -h;
132     elseif (nact == n)
133
134         AZ = A(indxact,:)';
135         [Q,R] = qr(AZ);      % factorizacion QR de AZ
136         Q1 = Q(:,1:nact); Q2 = Q(:,nact+1:n);
```

```
137         proj = zeros(n,1); % suponiendo columnas
                independientes en AZ
138     else                    % para 1 <= nact < n;
139         AZ = A(indxact,:)';
140         [Q,R] = qr(AZ);      % factorizacion QR de AZ
141         Q1 = Q(:,1:nact); Q2 = Q(:,nact+1:n);
142         proj= -Q2*(Q2')*h;  % Proyeccion de h en el
                nucleo de AZ
143
144     end
145     %
146     % Prueba de que el vector proj = 0 (necesario en
            caso de que nact=n.).
147     % Prueba de la optimalidad de x cuando la proj=0.
148     %
149     if ( max(abs(proj)) < TOLZERO )
150         %
151         % Prueba de optimalidad o modificacion de
                proyeccion y gradiente restringido.
152         %
153         w = R(1:nact,1:nact)\(Q\h);  % R puede ser
                singular (no checked)
154
155         if (max(abs(w)) <= 1)         % then x is
                optimal
156             xsol = x;
157             l1norm = norm(res,1);
158             return      % TERMINATION OF THE PROGRAM
159         end
160         %
161         % Elimino del conjunto de ecuaciones activas
                la ecuacion con
162         % indice IDELC, que corresponde que el valor
                de W de mayor
163         % magnitud. Actualiza el valor de NACT, y el
                indice IDELC se
164         % permuta con el indice en INDX(IADDC). Se
                modifica de manera
165         % correspondiente AZ, su factorizacion AZ=[Q1,
                Q2]*R, GRDX y la
166         % proyeccion PROJ.
167         %
168         [~,ixw] = sort(abs(w));  % ordena de menor a
                mayor en magnitud
169         idelc = indxact(ixw(end));   % indice que
                abandona INDXACT
```

```
170         %
171         % Actualizacion de NACT, INDX, INDXACT,
               INDXRES0
172         %
173         ixwp1 = ixw(end)+1;
174         for i = ixwp1 : iaddc
175             indx(i-1) = indx(i);
176         end
177         indx(iaddc) = idelc;
178         nact = nact - 1;
179         %
180         %  Actualizacion de AZ, factorizacion QR de AZ
               , GRDX y PROJ
181         %
182         if (nact == 0)
183             Q = eye(n); R = zeros(n,nact); %
                   factorizacion AZ (=0) = Id * R (=0)
184             Q1 = Q(:,1:nact); Q2 = Q(:,1:n);
185             proj = -h;
186         else                        % necesariamente NACT <
               N
187             AZ = A(indx(1:nact),:)';
188             [Q,R] = qr(AZ);
189             Q1 = Q(:,1:nact); Q2 = Q(:,nact+1:n);
190         end
191         proj = - sign(w(ixw(end)))*Q2*(Q2')*A(indx(
               idelc),:)';
192         grdx = h - sign(w(ixw(end)))*A(indx(idelc),:)
               ';
193     else
194         grdx = h;
195     end
196     %
197     % Ordena residuos positivos de ecuaciones no
           activas
198     %
199     %
200     alpha = - ((res(indxresn0) ./ (A(indxresn0, : )*
           proj)));
201     kk = find(alpha > TOLZERO); itop = length(kk);
202     [alphasort, ixalpha] = sort(alpha(kk));
203
204
205     for i = 1 : itop
206         if ((proj'*grdx) >= 2*sigma(indxresn0(kk(
               ixalpha(i))))*A(indxresn0(kk(ixalpha(i)))
```

```
                ,:)*proj)
207             x = x + alphasort(i) * proj;
208             break
209         else
210             grdx = grdx - 2*sigma(indxresn0(kk(ixalpha
                    (i))))*A(indxresn0(kk(ixalpha(i))),:)';
211         end
212         continue
213     end
214     %
215     % Actualiza ecuaciones activas: INDX, NACT
216     %
217     nact = nact + 1;
218     ix = indx(nact);    % permuta indx(nact+1) con
            indxresn0(kk(ixalpha(i)))
219     indx(nact) = indxresn0(kk(ixalpha(i)));
220     indx(iaddc + kk(ixalpha(i))) = ix;
221     if (iaddc < nact)
222       iaddc = nact;
223     end
224
225     done = false;
226 end
227 end
```

Bibliografía

[1] Barrodale. I., Roberts. F. D. K. *An improved algorithm for discrete L_1 linear approximation*. SIAM Journal on Numerical Analysis, 15(2).

[2] Bartels, R. H., Conn, A. R., and Sinclair, J. W. (1978). *Minimization techniques for piecewise differentiable functions: The L_1 solution to an overdetermined linear system.* SIAM Journal on Numerical Analysis, 15(2), pp. 224-241.

[3] Cheney, E. W. (1966). *Introduction to Approximation Theory.* McGraw-Hill Book Company. New York

[4] Golub, G. H., Van Loan, Ch. F. (2013) *Matrix Computations* 4th edition. The Johns Hopkins University Press. Baltimore.

[5] Powell, M. J. D. (1981). *Approximation Theory and Methods.* . Cambridge University Press. Cambridge.

[6] Watson, G. A. (1980). *Approximation Theory and Numerical Methods.* John Wiley and Sons. New York.