

UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

“Métodos de la programación lineal ordinaria y semi-infinita para
la aproximación uniforme de funciones, y una aplicación a la
óptica”

Tesis

Para obtener el título de:

LICENCIADO EN MATEMÁTICAS APLICADAS

p r e s e n t a :

Raúl Nivón Santiago

DIRECTOR DE TESIS:

M. en C. Cuauhtémoc H. Castañeda Roldán

Huajuapán de León, Oax.

Septiembre de 2003.

ÍNDICE

Introducción	2
Capítulo 1 Aproximación Uniforme	4
1.1 Desarrollo histórico de la Teoría de Aproximación	4
1.2 Aproximación polinomial uniforme	6
1.3 Caracterización de los mejores aproximantes	9
1.4 Aproximación sobre un conjunto finito de puntos	14
1.5 Aproximación uniforme por polinomios generalizados	19
1.6 Aproximación por polinomios generalizados sobre un conjunto finito de puntos	28
Capítulo 2 Optimización lineal	33
2.1 El Problema de Optimización Lineal	33
2.2 Dos métodos computacionales para la resolución de Programas Lineales.	35
2.2.1 Programación Lineal Ordinaria	35
2.2.2 Programación Lineal Semi-infinita	35
2.3 El Método Simplex	38
Capítulo 3 Optimización lineal aplicada al problema de aproximación uniforme	48
3.1 Optimización lineal y aproximación uniforme	48
3.2 Resultados numéricos	56
Capítulo 4 Una aplicación a la Óptica	63
4.1 Introducción	63
4.2 Superficies cónicas	63
4.3 Cálculo de parámetros	64
4.4 Ejemplos	66
Conclusiones	70
Bibliografía	72

Introducción

Hace más de cincuenta años que George B. Dantzig diseñó el Método Simplex [1] para resolver problemas de distribución óptima de pertrechos de guerra, además de otros problemas de planeación y coordinación propios de la acción militar. Desde entonces esta herramienta ha sido aplicada a la resolución de una gran variedad de programas lineales que modelan problemas en muchas de las áreas inimaginables del quehacer humano. Asimismo, el desarrollo de la optimización lineal y no lineal ha ido de la mano con sus exitosas aplicaciones en problemas científicos y tecnológicos de todo tipo, por mencionar solo algunos: planeación de la producción de bienes y servicios, problemas de control, modelación geométrica [11]. Con respecto a este último punto cabe resaltar la importancia de la aproximación de funciones para ajustar conjuntos de datos por medio de funciones fácilmente manejables, lo cual ha sido usado para el diseño asistido por computadora.

Diferentes métodos de optimización han sido usados para encontrar el mejor aproximante [3], no solo polinomial sino en diferentes familias de funciones, y con diferentes sentidos para el concepto de aproximación, de acuerdo a la seminorma (norma) empleada. Así por ejemplo, haciendo uso de la programación lineal puede encontrarse el mejor aproximante polinomial de grado menor o igual a un número natural predeterminado en el sentido de la seminorma uniforme [13] en el espacio de las funciones definidas sobre un conjunto discreto de la recta real. Otro método de optimización surgido más recientemente puede resultar más apropiado para este tipo de problemas: la programación semi-infinita; en ella se abordan problemas con un número infinito, ya sea de restricciones o de variables, pero no de ambas.

La programación semi-infinita lineal [8] puede considerarse como una extensión de la programación lineal tradicional. Sus herramientas se pueden aplicar al problema especial de la determinación del mejor aproximante polinomial.

En este trabajo se expone el problema de **encontrar el mejor aproximante** (polinomio algebraico o generalizado) en norma uniforme a una función continua f definida sobre un subconjunto compacto de R^k para $k = 1, 2, 3$, y se describe una metodología para su solución. Existen diferentes métodos para resolver tal problema como la *aproximación por polinomios de Taylor* [2], sin embargo estos polinomios brindan una buena aproximación a una función sólo en un punto específico y en puntos muy cercanos a éste. Otro método para aproximar funciones es el de *interpolación polinomial*, pero tiene el problema de que puede presentar oscilaciones muy grandes que en la función original no existen [2]. El método de mínimos cuadrados es también utilizado para la aproximación de funciones definidas sobre un conjunto discreto, este método involucra la resolución de un sistema de ecuaciones que resulta complicado, por ejemplo cuando la función a aproximar tiene forma exponencial [2].

Nosotros utilizamos la optimización lineal para resolver el problema de encontrar el mejor aproximante, el planteamiento de este problema lo hacemos en

dos formas, una mediante un programa lineal con un número finito de restricciones (Programación lineal ordinaria) y otra con un número infinito de restricciones (Programación lineal semi-infinita). Ambos programas son resueltos utilizando una implementación adecuada del Método Simplex.

En el primer capítulo revisamos la teoría de aproximación uniforme de funciones por polinomios algebraicos y generalizados tanto para el caso continuo como para el discreto, estudiamos la forma en la que se caracterizan estos polinomios y se tratan los temas de existencia y unicidad.

El segundo capítulo de este trabajo está enfocado a la optimización lineal, aquí desarrollamos los dos métodos para la resolución de programas lineales (número finito e infinito de restricciones), describimos la teoría del método Simplex y su algoritmo numérico.

En el tercer capítulo hacemos el planteamiento del problema de aproximación uniforme como un programa lineal para aplicar los métodos del capítulo II, demostramos particularmente que el algoritmo de solución de un programa lineal semi-infinito para este caso tiene una terminación finita. Se aproximan varias funciones obteniendo buenos resultados.

El cuarto capítulo es una aplicación a la Óptica, utilizamos el método de aproximación mediante programación lineal ordinaria para ajustar una superficie cónica a un conjunto de datos experimentales y conocer los parámetros geométricos de ésta, para realizar esa tarea fue necesario la búsqueda de una combinación de dos funciones aproximadoras adecuadas con las cuales el problema de aproximación fue resuelto satisfactoriamente.

Finalmente presentamos las conclusiones derivadas de los cuatro capítulos del trabajo realizado.

Capítulo 1

Aproximación Uniforme

Este capítulo está enfocado al estudio del mejor aproximante en la norma uniforme y a caracterizar el llamado mejor aproximante polinomial y el generalizado. Así mismo se estudia la aproximación sobre un conjunto finito de puntos y su relación con la aproximación sobre un intervalo. Esto permite abordar los métodos computacionales para obtener el mejor aproximante uniforme, numéricamente.

1.1 Desarrollo histórico de la Teoría de Aproximación

El hombre ha establecido leyes cuantitativas las cuales se expresan matemáticamente por medio de funciones, no con absoluta precisión, sino aproximadamente. A veces también es necesario aproximar funciones por medio de otras funciones.

El problema de determinar una función por medio de otras tiene gran importancia teórica; así por ejemplo, el desarrollo del análisis matemático ha impulsado el descubrimiento y estudio de clases muy importantes de funciones que bajo condiciones conocidas han demostrado ser los medios naturales de aproximar otras más o menos arbitrarias. Estas clases resultaron ser, sobre todo, los polinomios algebraicos y trigonométricos, y también sus distintas generalizaciones. Se ha demostrado que a partir de las propiedades de la función que se desea aproximar es posible estimar, bajo ciertas condiciones, el carácter de su desviación respecto de una sucesión de funciones aproximadoras.

Un método importante de aproximación es la fórmula de Taylor. Aquí una función que satisface ciertas condiciones se aproxima con otra de la forma $P(x) = a_0 + a_1x + \dots + a_nx^n$, llamada polinomio algebraico. Además de esta fórmula hay otras de gran importancia práctica en la aproximación de funciones. Entre ellas están las distintas fórmulas de interpolación, el método de aproximación en el sentido del cuadrado medio, etc.

Cada uno de los métodos de cálculo surgió en su momento y tiene su teoría e historia características. Newton utilizaba ya las fórmulas de interpolación, que expresó de un modo muy conveniente para el cálculo práctico con los llamados cocientes de diferencias. Pero durante mucho tiempo estos métodos no dieron lugar a una teoría coherente.

La teoría actual de la aproximación de funciones emanó del trabajo de Chebichev, quien introdujo el importante concepto de aproximación óptima.

El ulterior desarrollo de la teoría de aproximación de funciones se vió influido por un importante descubrimiento matemático, hecho a finales del siglo XVIII por el matemático alemán Weierstrass, quien demostró con absoluto rigor la posibilidad de aproximar una función continua arbitraria por un polinomio algebraico con cualquier grado de exactitud.

Las ideas de Chebichev y el teorema de Weierstrass sirvieron de base, a principios de este siglo para el desarrollo actual de la teoría de la aproximación.

Hasta Chebichev, los problemas consistían generalmente en la aproximación de funciones concretas, mientras que el problema característico del momento actual es la aproximación, por polinomios o de otro modo, de clases enteras de funciones, analíticas, diferenciables, etc.

Además de los polinomios algebraicos, otro medio muy importante de aproximación consiste en los polinomios trigonométricos. Existen distintos métodos particulares de aproximación mediante estos polinomios, generalmente relacionados de una forma más bien simple con los métodos correspondientes de polinomios algebraicos.

1.2 Aproximación Polinomial Uniforme

Definición 1. Sea V un espacio lineal. Llamamos *norma* a una función definida en V hacia los números reales no negativos. Esta función se escribe $\|\cdot\|$ y debe satisfacer las siguientes tres propiedades:

Dados $w, v \in V, \lambda \in R$.

- (i) $\|v\| \geq 0$ con igualdad si y sólo si $v = 0$.
- (ii) $\|\lambda v\| = |\lambda| \|v\|$ para cualquier escalar λ . (1.2.1)
- (iii) $\|v + w\| \leq \|v\| + \|w\|$ (la desigualdad del triángulo).

La norma dá una noción de distancia en V . Si $w, v \in V$, entonces la distancia de w a v es $\|v - w\|$.

Teorema 1.1 (*Existencia del mejor aproximante*) Si V es un espacio lineal normado y W un subespacio de V de dimensión finita, entonces, dado $v \in V$, existe $w^* \in W$ tal que

$$\|v - w^*\| \leq \|v - w\|$$

para toda $w \in W$.

Demostración. Supongamos que W tiene dimensión k y w_1, \dots, w_k es una base para W ; entonces probaremos que la función $f : R^k \mapsto R$ definida por

$$f(\lambda_1, \dots, \lambda_k) = \|v - (\lambda_1 w_1 + \lambda_2 w_2 + \dots + \lambda_k w_k)\| \quad (1.2.2)$$

tiene un mínimo sobre R^k .

Veamos primero que f es continua.

Por equivalencia de normas, dados $\lambda, \lambda' \in R^k$

$$\|\lambda - \lambda'\|_{máx} < c \|\lambda - \lambda'\|,$$

donde $c > 0$, y $\|\cdot\|$ es la norma euclidiana y $\|\cdot\|_{máx}$ es la norma del máximo.

Sea $\varepsilon > 0, \lambda \in R^k$. Haciendo $\delta = \frac{\varepsilon}{c \sum_{i=1}^k \|w_i\|}$ tenemos que para toda $\lambda' \in R^k$

$$\text{si } \|\lambda' - \lambda\| < \delta, \quad \text{entonces } \|\lambda' - \lambda\|_{máx} < \frac{\varepsilon}{\sum_{i=1}^k \|w_i\|},$$

es decir

$$|\lambda'_i - \lambda_i| < \frac{\varepsilon}{\sum_{i=1}^k \|w_i\|} \quad i = 1, \dots, k;$$

luego

$$|\lambda'_1 - \lambda_1| \|w_1\| + \cdots + |\lambda'_k - \lambda_k| \|w_k\| < \frac{\varepsilon}{\sum_{i=1}^k \|w_i\|} \left(\sum_{i=1}^k \|w_i\| \right) = \varepsilon,$$

y como

$$\begin{aligned} |f(\lambda') - f(\lambda)| &= \left| \|v - (\lambda'_1 w_1 + \cdots + \lambda'_k w_k)\| - \|v - (\lambda_1 w_1 + \cdots + \lambda_k w_k)\| \right| \\ &\leq |\lambda'_1 - \lambda_1| \|w_1\| + \cdots + |\lambda'_k - \lambda_k| \|w_k\|, \end{aligned}$$

entonces

$$|f(\lambda') - f(\lambda)| < \varepsilon.$$

Con esto queda demostrada la continuidad de f .

De igual forma la función $h : R^k \mapsto R$ definida por

$$h(\lambda_1, \dots, \lambda_k) = \|(\lambda_1 w_1 + \lambda_2 w_2 + \cdots + \lambda_k w_k)\|$$

es continua.

Considérese la superficie esférica:

$$S = \{(\lambda_1, \dots, \lambda_k) \in R^k : \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_k^2 = 1\}. \quad (1.2.3)$$

Sabemos que S es cerrada y acotada, por tanto h alcanza su mínimo m aquí. Este valor mínimo, m , es positivo. Para ver esto, observemos que si $h(\lambda_1^*, \dots, \lambda_k^*) = m$, entonces $m \geq 0$. Si $m = 0$ entonces $\|(\lambda_1^* w_1 + \lambda_2^* w_2 + \cdots + \lambda_k^* w_k)\| = 0$, lo cual por (1.2.1) implicaría que $\lambda_1^* w_1 + \lambda_2^* w_2 + \cdots + \lambda_k^* w_k = 0$. Como w_1, \dots, w_k son linealmente independientes, concluimos que $\lambda_1^* = \lambda_2^* = \cdots = \lambda_k^* = 0$, contradiciendo la igualdad en (1.2.3).

Escribiendo $r = (\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_k^2)^{\frac{1}{2}}$, se tiene

$$h(\lambda_1, \dots, \lambda_k) = r \left\| \left(\frac{\lambda_1}{r} w_1 + \frac{\lambda_2}{r} w_2 + \cdots + \frac{\lambda_k}{r} w_k \right) \right\|,$$

así que

$$h(\lambda_1, \dots, \lambda_k) \geq mr.$$

Más aún,

$$\begin{aligned} f(\lambda_1, \dots, \lambda_k) &= \|v - (\lambda_1 w_1 + \cdots + \lambda_k w_k)\| \geq \|(\lambda_1 w_1 + \cdots + \lambda_k w_k)\| - \|v\| \\ &\geq mr - \|v\|. \end{aligned}$$

Pongamos

$$\rho = \inf_{\lambda \in R^k} f(\lambda) \quad \text{y} \quad d = \frac{1 + \rho + \|v\|}{m}.$$

Si $\lambda_1^2 + \dots + \lambda_k^2 > d^2$, entonces $f(\lambda_1, \dots, \lambda_k) > md - \|v\| = 1 + \rho > \rho$, por tanto el mínimo de f sobre R^k es igual al mínimo de f sobre $A = \{\lambda_1^2 + \lambda_2^2 + \dots + \lambda_k^2 \leq d^2\}$. Como f es una función continua de $(\lambda_1, \dots, \lambda_k)$, entonces f alcanza su mínimo en A . ■

Definición 2. Un elemento w que cumple las condiciones del teorema 1.1 es llamado *mejor aproximante*.

Consideremos el espacio vectorial de las funciones continuas sobre un intervalo cerrado $[a, b]$ denotado por $C[a, b]$. Si P_n denota al espacio de polinomios de grado menor o igual a n , entonces el teorema 1.1 asegura que, dada $f \in C[a, b]$, existe un polinomio $p_n^* \in P_n$ tal que

$$\|f - p_n^*\| \leq \|f - p\|, \quad \text{para todo } p \in P_n,$$

donde $\|\cdot\|$ es la norma uniforme sobre el intervalo $[a, b]$, esto es,

$$\|g\| = \max_{a \leq x \leq b} |g(x)| \tag{1.2.4}$$

para cualquier $g \in C[a, b]$.

Denotaremos $E_n(f; [a, b]) = E_n(f) = \|f - p_n^*\|$ al error cometido al aproximar f .

1.3 Caracterización de los Mejores Aproximantes

Entramos ahora al estudio de las propiedades de los polinomios $p_n^* \in P_n$ de mejor aproximación a una función continua dada f sobre el intervalo $[a, b]$.

Sea $e = f - p_n^*$; tenemos $\|e(x)\| = E_n(f; [a, b])$.

Definición 3. Un conjunto de $k + 1$ puntos distintos x_0, \dots, x_k que satisfacen $a \leq x_0 < x_1 < \dots < x_{k-1} < x_k \leq b$ es llamado un *conjunto alternante* para la función error $f - p_n$ si

$$|f(x_j) - p_n(x_j)| = \|f - p_n\|, \quad j = 0, \dots, k$$

y

$$[f(x_j) - p_n(x_j)] = -[f(x_{j+1}) - p_n(x_{j+1})], \quad j = 0, \dots, k - 1 \quad (1.3.1)$$

Teorema 1.2 Sea $f \in C[a, b]$; $p_n^* \in P_n$ es un mejor aproximante para f sobre $[a, b]$ si y sólo si existe un conjunto alternante para $e = f - p_n^*$ de $n + 2$ puntos.

Demostración. Supongamos que x_0, \dots, x_{n+1} forman un conjunto alternante para $f - p_n^*$. Se mostrará que p_n^* es un mejor aproximante. Si no lo es, entonces existe $q_n \in P_n$ tal que

$$\|f - q_n\| < \|f - p_n^*\| \quad (1.3.2)$$

En particular, como x_0, \dots, x_{n+1} forman un conjunto alternante, entonces

$$|f(x_j) - q_n(x_j)| < \|f - p_n^*\| = |f(x_j) - p_n^*(x_j)|, \quad j = 0, \dots, n + 1 \quad (1.3.3)$$

(1.3.3) y (1.3.1) implican que la diferencia

$$[f(x_j) - p_n^*(x_j)] - [f(x_j) - q_n(x_j)] \quad (1.3.4)$$

alterna de signo cuando j va de 0 a $n + 1$. Veamos como sucede esto, supongamos que para $j = 0$, $f(x_j) - p_n^*(x_j) > 0$, entonces debido a (1.3.3) la diferencia (1.3.4) es mayor que cero, después para $j = 1$, $f(x_j) - p_n^*(x_j) < 0$ por (1.3.1), y otra vez por (1.3.3) la diferencia (1.3.4) es menor que cero, esta alternancia se va presentando hasta $j = n + 1$. La alternancia de signos se presenta de manera similar cuando $f(x_j) - p_n^*(x_j) < 0$ para $j = 0$. Ahora bien,

$$[f(x_j) - p_n^*(x_j)] - [f(x_j) - q_n(x_j)] = q_n(x_j) - p_n^*(x_j).$$

Así que $q_n - p_n^* \in P_n$ tiene un cero en cada intervalo (x_j, x_{j+1}) , $j = 0, \dots, n$, para un total de $n + 1$ ceros, luego $q_n - p_n^*$ debe ser el polinomio cero, lo cual implica que $q_n = p_n^*$. Esto contradice (1.3.2), por lo tanto p_n^* es un mejor aproximante.

Recíprocamente supongamos que $p_n^* \in P_n$ es un mejor aproximante para f . Hagamos

$$\rho = \|f - p_n^*\|,$$

y seleccionemos $\varepsilon > 0$ suficientemente pequeño para que $|x_1 - x_2| < \varepsilon$ implique

$$|e(x_1) - e(x_2)| \leq \frac{1}{2}\rho$$

para $x_1, x_2 \in [a, b]$. Esto es posible por la continuidad uniforme de la función de error e (e es una función continua sobre un intervalo cerrado).

Dividamos $[a, b]$ en intervalos cerrados consecutivos de longitud menor o igual a ε . Denotemos por I_1, I_2, \dots, I_m a los intervalos en los cuales $|e(x)|$ alcanza su máximo valor. Como e puede variar a lo más $\frac{1}{2}\rho$ en cualquiera de estos intervalos, entonces

$$e(x) > \frac{1}{2}\rho \quad \text{ó} \quad e(x) < -\frac{1}{2}\rho,$$

en tales intervalos.

Consideremos la sucesión u_1, u_2, \dots, u_m formada por valores 1 ó -1 de acuerdo al signo que presente e sobre los intervalos I_1, I_2, \dots, I_m , por ejemplo, si e toma el valor $-\rho$ sobre I_j entonces $u_j = -1$.

Debemos demostrar que en ésta sucesión se presentan al menos $n+1$ cambios de signo, con lo cual podemos formar un conjunto alternante para $e = f - p_n^*$ de al menos $n+2$ puntos. Haremos esto mostrando que si hubieran menos de $n+1$ cambios, podríamos encontrar un polinomio para el cual el error es menor que ρ .

Si $u_1 = u_2 = \dots = u_m$ sumando una constante apropiada a p_n^* obtendríamos una mejor aproximación, para ver esto supongamos que $u_i = 1, i = 1, \dots, m$, es decir, $e(x) > -\rho$ sobre todo $[a, b]$; entonces

$$\min_{a \leq x \leq b} e(x) = M > -\rho,$$

hagamos

$$c = \frac{\rho + M}{2} > 0.$$

Tomando $q_n = p_n^* + c$, $f(x) - q_n(x) = f(x) - p_n^*(x) - c = e(x) - c$ se tiene

$$-(\rho - c) = c - \rho = M - c \leq e(x) - c \leq \rho - c,$$

por tanto $\|f - q_n\| = \rho - c$. Lo que significa que q_n es mejor que p_n^* al aproximar f . El caso $u_i = -1$ se establece similarmente.

Agrupemos entonces los intervalos consecutivos I_1, I_2, \dots, I_m en grupos donde el signo de las $u_i, i = 1, \dots, m$ sea el mismo:

$$\begin{array}{ll} \text{Grupo 1} & \{I_1, I_2, \dots, I_{j_1}\} \quad 1^{er} \text{ signo} \\ \text{Grupo 2} & \{I_{j_1+1}, I_{j_1+2}, \dots, I_{j_2}\} \quad 2^{do} \text{ signo} \\ \text{Grupo 3} & \{I_{j_2+1}, I_{j_2+2}, \dots, I_{j_3}\} \quad 1^{er} \text{ signo} \\ & \vdots \\ \text{Grupo } k & \{I_{j_{k-1}+1}, I_{j_{k-1}+2}, \dots, I_{j_k} = I_m\} \end{array} \quad (1)$$

Este esquema muestra $k - 1$ cambios de signo, supongamos entonces que se presentan menos de $n + 1$ cambios, es decir

$$k - 1 < n + 1.$$

Consideremos los intervalos I_{j_1}, I_{j_1+1} . Estos intervalos son disjuntos pues para alguno $e(x) > \frac{1}{2}\rho$ y para otro $e(x) < -\frac{1}{2}\rho$. Por tanto podemos encontrar un x_1 que cumpla:

$$s < x_1 < t \text{ para todo } s \in I_{j_1} \text{ y } t \in I_{j_1+1} \quad (1.3.6)$$

De igual manera existen x_2, \dots, x_{k-1} tales que

$$\begin{aligned} s < x_2 < t & \text{ para todo } s \in I_{j_2} \text{ y } t \in I_{j_2+1} \quad 1.3.7 \quad (2) \\ & \vdots \\ s < x_{k-1} < t & \text{ para todo } s \in I_{j_{k-1}} \text{ y } t \in I_{j_{k-1}+1}. \end{aligned}$$

Pongamos

$$q(x) = (x_1 - x)(x_2 - x) \cdots (x_{k-1} - x),$$

Como $k - 1 < n + 1$, $k - 1 \leq n$ luego $q \in P_n$. Las raíces del polinomio q son los valores x_i , los cuales no pertenecen a ninguno de los intervalos I_1, \dots, I_m .

Supongamos que la función e toma el valor ρ sobre I_1 , por construcción e tomará éste valor sobre el grupo 1 de intervalos en (1.3.5), sobre el grupo 2 tomará entonces el valor $-\rho$ y así alternadamente. Ahora bien, sobre el primer grupo de intervalos en (1.3.5) el valor de $q(x) = (x_1 - x)(x_2 - x) \cdots (x_{k-1} - x)$ es positivo, pues en vista de (1.3.6) el valor de todos sus factores es positivo. Sobre el segundo grupo el valor de q es negativo pues únicamente su primer factor es negativo (debido a (1.3.6)) y los restantes positivos (por (1.3.7)), sobre el tercero el valor de q es positivo ya que el valor de sus primeros dos factores es negativo y del resto positivo, así sucesivamente se presenta la alternancia de signo en q .

Observamos entonces que las evaluaciones de e y q coinciden en signo sobre los intervalos I_1, \dots, I_m , cuando e toma el valor $-\rho$ sobre I_1 , multiplicando q por -1 se tiene el mismo resultado.

Sea $T = [a, b] - I_1 - I_2 - \cdots - I_m$;

$$\max_{x \in T} |e(x)| = \rho' < \rho.$$

Supongamos que $\|q\| = M$ y que $\lambda > 0$ se escoge tan pequeño como se requiera para hacer válida la desigualdad

$$\lambda M < \min(\rho - \rho', \frac{\rho}{2}).$$

Mostraremos que $p = \lambda q + p_n^* \in P_n$ aparece como mejor aproximante para f , en lugar de p_n^* .

Si $x \in T$ entonces

$$|f(x) - p(x)| = |e(x) - \lambda q(x)| \leq |e(x)| + \lambda |q(x)| \leq \rho' + \lambda M < \rho. \quad (1.3.8)$$

Por otro lado, si $x \notin T$ entonces $x \in I_j$ para algún j entre 1 y m , aquí $e(x)$ y $\lambda q(x)$ tienen el mismo signo, y $|e(x)| \geq \frac{\rho}{2} > \lambda M \geq |\lambda q(x)|$, así tenemos

$$|f(x) - p(x)| = |e(x) - \lambda q(x)| = |e(x)| - \lambda |q(x)| \leq \rho - |\lambda q(x)| < \rho$$

que junto con (1.3.8) implican

$$\|f - p\| < \rho = \|f - p_n^*\|,$$

lo cual es una contradicción, por lo tanto $k - 1 \geq n + 1$, y el teorema está demostrado. ■

Teorema 1.3 *Si p_n^* es una mejor aproximación a $f \in C[a, b]$ en norma uniforme, $f \notin P_n$, entonces p_n^* es único; esto es, si $p \in P_n$ y $p \neq p_n^*$, $\|f - p\| > \|f - p_n^*\|$.*

Demostración. Supongamos que $p \in P_n$ es también un mejor aproximante para f , entonces $\|f - p\| = \|f - p_n^*\| = E_n(f)$ y

$$q = \frac{p + p_n^*}{2}$$

es también un mejor aproximante para f , pues

$$\begin{aligned} E_n(f) &\leq \|f - q\| = \left\| f - \frac{p + p_n^*}{2} \right\| = \left\| \frac{1}{2}(f - p) + \frac{1}{2}(f - p_n^*) \right\| \\ &\leq \frac{1}{2}(\|f - p\| + \|f - p_n^*\|) = E_n(f). \end{aligned}$$

Sea x_0, \dots, x_{n+1} un conjunto alternante para $f - q$, entonces para algún entero l ,

$$\begin{aligned} f(x_j) - q(x_j) &= \frac{f(x_j) - p(x_j)}{2} + \frac{f(x_j) - p_n^*(x_j)}{2} \\ &= (-1)^{l+j} E_n(f), \quad j = 0, \dots, n+1, \end{aligned}$$

o bien,

$$f(x_j) - p(x_j) + f(x_j) - p_n^*(x_j) = 2(-1)^{l+j} E_n(f), \quad j = 0, \dots, n+1$$

lo cual implica

$$f(x_j) - p(x_j) = f(x_j) - p_n^*(x_j) = (-1)^{l+j} E_n(f) \quad j = 0, \dots, n+1,$$

pues

$$|f(x_j) - p(x_j)| \leq E_n(f) \quad y \quad |f(x_j) - p_n^*(x_j)| \leq E_n(f),$$

así obtenemos

$$p(x_j) = p_n^*(x_j) \quad j = 0, \dots, n + 1,$$

luego

$$p = p_n^*.$$

Por lo tanto p_n^* es único. ■

1.4 Aproximación sobre un conjunto finito de puntos.

No siempre se puede encontrar el polinomio de mejor aproximación a una función dada, pero algo que siempre es posible es aproximar al mejor aproximante. El procedimiento a considerar es reemplazar el intervalo I sobre el cual se aproxima por un conjunto finito de puntos y resolver el problema de aproximación sobre ese conjunto.

Así tenemos dos problemas: El primero es encontrar el mejor aproximante sobre el conjunto finito de puntos y el segundo que éste aproxime al mejor aproximante sobre I cuando el número de puntos crece apropiadamente.

Consideremos un conjunto X_m de puntos $x_1 < x_2 < \dots < x_m$, todos ellos distintos, sobre el intervalo I . Llamaremos E_m al conjunto de funciones definidas sobre x_1, x_2, \dots, x_m , esto es, el espacio m -dimensional de vectores $f : (f_1, \dots, f_m)$; f_j es el valor de $f(x)$ en $x = x_j$, $j = 1, \dots, m$. La norma uniforme sobre E_m está definida por

$$\|f\| = \max_{i=1, \dots, m} |f_i|$$

Tomando $V = E_m$ y W el subespacio de E_m consistente de todos los vectores $p : (p(x_1), \dots, p(x_m))$, donde $p \in P_n$, el Teorema 1.1 nos dice que existe $p^* \in P_n$ tal que

$$\max_{i=1, \dots, m} |f_i - p^*(x_i)| \leq \max_{i=1, \dots, m} |f_i - p(x_i)|$$

para todo $p \in P_n$.

Al igual que $E_n(f; I) = \|f - p_n^*\| = \max_{x \in I} |f(x) - p_n^*(x)|$, denotamos

$$E_n(f, X_m) = \max_{i=1, \dots, m} |f(x_i) - p^*(x_i)|.$$

Para cualquier f sobre X_m , si $m \leq n + 1$, siempre existe $p \in P_n$ tal que

$$p(x_j) = f(x_j), \quad j = 1, \dots, m. \quad (1.4.1)$$

Este p es un polinomio de interpolación y es un mejor aproximante. Por lo tanto, supondremos que $m \geq n + 2$. El mejor aproximante uniforme en P_n para f sobre X_m es caracterizado por un teorema análogo al teorema 1.2. Enunciamos el siguiente teorema, su demostración es la misma que la del teorema 1.2 con X_m en lugar de $[a, b]$.

Teorema 1.4 $p_n^*(X_m) \in P_n$ es un mejor aproximante sobre X_m para f si y sólo si existe un conjunto alternante (de puntos de X_m) para $f - p_n^*$ de $n + 2$ puntos.

Igualmente se tiene el análogo al teorema 1.3.

Teorema 1.5 El mejor aproximante sobre X_m para f es único.

Estudiaremos ahora la relación entre el mejor aproximante sobre un intervalo y el mejor aproximante sobre un conjunto de puntos del intervalo.

Consideremos el intervalo $I : [-1, 1]$ y supongamos $X_m \subset I$ es un conjunto de m puntos distintos, incluyendo los puntos -1 y 1 . Si por ejemplo, X_m consiste de puntos equitativamente espaciados en I , es decir, los puntos

$$-1 + \frac{2j}{m-1}, \quad j = 0, \dots, m-1, \quad (1.4.2)$$

la distancia entre dos puntos consecutivos de (1.4.2), denotada por δ_m es

$$\delta_m = \left| \left(-1 + \frac{2j}{m-1} \right) - \left(-1 + \frac{2(j+1)}{m-1} \right) \right| = \frac{2}{m-1} > 0. \quad (1.4.3)$$

Lema 1.1 *Consideremos el intervalo I , $X_m \subset I$ y δ_m antes mencionados.*

Si $p \in P_n$ satisface

$$|p(x)| \leq K, \quad x \in X_m \quad (1.4.4)$$

y

$$\delta_m < \frac{\sqrt{6}}{n\sqrt{n^2-1}}, \quad (1.4.5)$$

entonces

$$|p(x)| \leq \frac{K}{1-\tau}, \quad x \in I, \quad (1.4.6)$$

donde

$$\tau = \frac{n^2(n^2-1)}{6} \delta_m^2. \quad (1.4.7)$$

Demostración. Como $\delta_m < \frac{\sqrt{6}}{n\sqrt{n^2-1}}$, entonces $\delta_m^2 < \frac{6}{n^2(n^2-1)}$, de aquí

$$0 \leq \tau = \frac{n^2(n^2-1)}{6} \delta_m^2 < 1.$$

Supongamos que $\|p\| = |p(\xi)|$ para algún $\xi \in I$ ($\|p\| = \max_{x \in I} |p(x)|$). Si $\xi = \pm 1$, entonces $\xi \in X_m$, de (1.4.4) dado $x \in I$,

$$|p(x)| \leq |p(\xi)| \leq K,$$

y como $0 \leq \tau < 1$, $K \leq \frac{K}{1-\tau}$, con lo cual se tiene (1.4.6)

$$|p(x)| \leq \frac{K}{1-\tau}, \quad x \in I.$$

Si $-1 < \xi < 1$, sea x_i el punto de X_m más cercano a ξ . Entonces

$$p(x_i) = p(\xi) + (x_i - \xi)p'(\xi) + \frac{(x_i - \xi)^2}{2!} p''(\eta)$$

para algún $\eta \in I$. Pero $p'(\xi) = 0$ pues $p(x)$ tiene un extremo relativo en ξ ; por tanto

$$-p(\xi) = -p(x_i) + \frac{(x_i - \xi)^2}{2!} p''(\eta)$$

luego

$$\begin{aligned} \|p\| &= |p(\xi)| = \left| -p(x_i) + \frac{(x_i - \xi)^2}{2!} p''(\eta) \right| \\ &\leq |p(x_i)| + \frac{(x_i - \xi)^2}{2!} |p''(\eta)| \leq K + \frac{\delta_m^2}{2} |p''(\eta)|. \end{aligned} \quad (3)$$

Necesitamos ahora una cota para $|p''(\eta)|$, la cual se obtiene de un resultado de V. Markov¹, con $k = 2$ aplicado al polinomio $q = \frac{1}{\|p\|} p$,

$$|p''(\eta)| \leq \frac{n^2(n^2 - 1)}{3} \|p\|. \quad (1.4.9)$$

Substituyendo (1.4.9) en (1.4.8) tenemos

$$\|p\| \leq K + \frac{\delta_m^2}{6} n^2(n^2 - 1) \|p\| = K + \tau \|p\|,$$

así

$$|p(x)| \leq \frac{K}{1 - \tau}, \quad x \in I \blacksquare$$

Definición 4. Sea f definida sobre $[a, b]$, el *módulo de continuidad* de f sobre $[a, b]$, $\omega(f; [a, b]; \delta)$, es definido para $\delta > 0$ por

$$\omega(f, [a, b], \delta) = \sup_{\substack{x_1, x_2 \in [a, b] \\ |x_1 - x_2| \leq \delta}} |f(x_1) - f(x_2)|.$$

Lema 1.2 Para $p \in P_n$

$$\omega(p; I; \delta) \leq \delta n^2 \|p\| \quad (1.4.10)$$

Demostración. Si $x_1, x_2 \in I$, por el teorema del valor medio,

$$p(x_2) - p(x_1) = (x_2 - x_1) p'(\eta), \text{ para algún } \eta \in (x_1, x_2)$$

Por el resultado de Markov¹ con $k = 1$ aplicado al polinomio $\frac{1}{\|p\|} p$ tenemos que

$|p'(\eta)| \leq n^2 \|p\|$, y entonces,

$$|p(x_2) - p(x_1)| \leq \delta n^2 \|p\|,$$

¹ $\|p^{(k)}\| \leq \frac{n^2(n^2-1)\dots(n^2-(k-1)^2)}{1 \cdot 3 \cdot 5 \dots (2k-1)}$, cuando $p \in P_n$ y $\|p\| \leq 1$.

lo cual implica (1.4.10). ■

Teorema 1.6 *Si f es una función continua sobre I , entonces*

$$E_n(f; I) - \omega(f; I; \delta_m) - \delta_m n^2 \left[\frac{\|f\| + E_n(f; I)}{1 - \tau} \right] \leq E_n(f; X_m) \leq E_n(f; I), \quad (1.4.11)$$

suponiendo que δ_m satisface (1.4.5).

Demostración. Sea p_n^* el mejor aproximante para f sobre I y q_n^* el mejor aproximante para f sobre X_m . Entonces

$$E_n(f; X_m) \leq \max_{x \in X_m} |f(x) - p_n^*(x)| \leq \max_{x \in I} |f(x) - p_n^*(x)| = E_n(f; I). \quad (1.4.12)$$

Dado $x \in I$, sea $x_i \in X_m$ un punto de X_m más cercano a x ; tenemos

$$f(x) - q_n^*(x) = [f(x) - f(x_i)] + [f(x_i) - q_n^*(x_i)] + [q_n^*(x_i) - q_n^*(x)],$$

por tanto

$$|f(x) - q_n^*(x)| \leq \omega(f; I; \delta_m) + E_n(f; X_m) + \omega(q_n^*; I; \delta_m). \quad (1.4.13)$$

Como $E_n(f; I) = \|f - p_n^*\| \leq \|f - q_n^*\|$, de (1.4.13) obtenemos

$$E_n(f; X_m) \geq E_n(f; I) - \omega(f; I; \delta_m) - \omega(q_n^*; I; \delta_m). \quad (1.4.14)$$

Sólo falta ahora una cota superior para $\omega(q_n^*; I; \delta_m)$. Combinando los resultados de los Lemas 1.1 y 1.2 obtenemos

$$\omega(q_n^*; I; \delta_m) \leq \delta_m n^2 \|q_n^*\| \leq \delta_m n^2 (1 - \tau)^{-1} \max_{x \in X_m} |q_n^*(x)|. \quad (1.4.15)$$

Pero como $|f(x) - q_n^*(x)| \leq E_n(f; X_m)$ para $x \in X_m$,

$$\begin{aligned} \max_{x \in X_m} |q_n^*(x)| &= \max_{x \in X_m} |q_n^*(x) - f(x) + f(x)| \\ &\leq \max_{x \in X_m} |f(x) - q_n^*(x)| + \max_{x \in X_m} |f(x)| \\ &= E_n(f; X_m) + \|f\| \leq E_n(f; I) + \|f\|. \end{aligned} \quad (1.4.16)$$

Combinando (1.4.15) y (1.4.16):

$$\omega(q_n^*; I; \delta_m) \leq \delta_m n^2 (1 - \tau)^{-1} (E_n(f; I) + \|f\|). \quad (1.4.17)$$

Y de (1.4.14) y (1.4.17):

$$E_n(f; X_m) \geq E_n(f; I) - \omega(f; I; \delta_m) - \delta_m n^2 \left[\frac{\|f\| + E_n(f; I)}{1 - \tau} \right]$$

■

Corolario 1.6.1 Si $\delta_m \rightarrow 0$ cuando $m \rightarrow \infty$, entonces $E_n(f; X_m) \rightarrow E_n(f; I)$ cuando $m \rightarrow \infty$.

Demostración. Cuando $m \rightarrow \infty$, $\omega(f; I; \delta_m) \rightarrow 0$ por la continuidad de f . Como $\tau \rightarrow 0$, de (1.4.11) se obtiene el resultado. ■

1.5 Aproximación Uniforme por polinomios generalizados

La aproximación de una función por polinomios algebraicos de grado menor o igual a n puede generalizarse de la siguiente forma; dado $p \in P_n$, p tiene la forma

$$p(x) = a_0 + a_1x + \cdots + a_nx^n, \quad (1.5.1)$$

es decir p es una combinación lineal de las funciones $\{1, x, x^2, \dots, x^n\}$. Suponiendo ahora que tenemos n funciones continuas g_1, g_2, \dots, g_n , en algún intervalo, podemos formar entonces una combinación lineal de éstas como en (1.5.1)

$$\sum_{i=1}^n c_i g_i, \quad (1.5.2)$$

La combinación lineal (1.5.2) es llamada *polinomio generalizado*.

La existencia del mejor aproximante para una función $f \in V = C[a, b]$ está garantizada por el Teorema 1.1 tomando como subespacio W al subespacio de V generado por las funciones g_1, g_2, \dots, g_n continuas en $[a, b]$. En lo sucesivo nos enfocaremos a mostrar la unicidad del mejor aproximante en este caso.

Definición 5. Sea A un espacio lineal, el conjunto

$$H(A) = \left\{ g : g = \sum_{i=1}^m \theta_i f_i, f_i \in A, \theta_i \geq 0, \sum_{i=1}^m \theta_i = 1 \right\},$$

es llamado *cápsula convexa*.

Teorema 1.7 (Teorema de Carathéodory) *Sea V un espacio lineal de dimensión n . Sea $A \subset V$. Todo punto de la cápsula convexa de A se puede expresar como una combinación lineal de $n + 1$ elementos (ó menos) de A .*

Demostración. Sea $g \in H(A)$, entonces

$$g = \sum_{i=0}^k \theta_i f_i, \quad f_i \in A, \theta_i \geq 0, \sum_{i=0}^k \theta_i = 1. \quad (1.5.3)$$

Haciendo que k sea el menor número para el cual se cumple (1.5.3), se tiene que $\theta_i > 0$. Ahora,

$$\sum_{i=0}^k \theta_i (f_i - g) = \sum_{i=0}^k \theta_i f_i - \sum_{i=0}^k \theta_i g = g - g = 0,$$

de modo que el conjunto de vectores $g_i = f_i - g$, $i = 1, \dots, k$, es linealmente dependiente.

Supongamos que $k > n$, entonces $\{g_1, \dots, g_k\}$ es un conjunto linealmente dependiente. Supongamos $\sum_{i=0}^k \alpha_i g_i = 0$, con $\sum_{i=0}^k |\alpha_i| \neq 0$. Haciendo $\alpha_0 = 0$ tenemos que para toda λ ,

$$\sum_{i=0}^k (\theta_i + \lambda \alpha_i) g_i = 0. \quad (1.5.4)$$

Si

$$\lambda = \max \left\{ -\frac{\theta_i}{\alpha_i} : \alpha_i \neq 0 \right\},$$

entonces $\theta_j + \lambda \alpha_j = 0$ para alguna j entre 1 y k . Para los demás índices se tiene $\theta_i + \lambda \alpha_i \geq 0$, donde $\theta_0 + \lambda \alpha_0 = \theta_0 > 0$. Sustituyendo g_i por $f_i - g$ en (1.5.4) obtenemos

$$\sum_{i=0}^k (\theta_i + \lambda \alpha_i) (f_i - g) = 0,$$

es decir,

$$\sum_{i=0}^k (\theta_i + \lambda \alpha_i) f_i = g \sum_{i=0}^k (\theta_i + \lambda \alpha_i),$$

dividiendo por $\sum_{i=0}^k (\theta_i + \lambda \alpha_i)$ se tiene una expresión para g con menos de k términos, contradiciendo que k era el menor número que cumplía (1.5.3), por lo tanto debemos tener que $k \leq n$.

Si $k = n$, la demostración está completa, si $k < n$, agregando coeficientes $\theta_j = 0$ tantas veces como sea necesario en la representación (1.5.3) de g se tendrá el mismo resultado. ■

Corolario 1.7.1 *La cápsula convexa de un subconjunto compacto de un espacio de dimensión n es compacto.*

Demostración. Sea X un subconjunto compacto de un espacio de dimensión n . Sea $\{v_i\}$ una sucesión en $H(X)$. Por el teorema 1.7, para todo k :

$$v_k = \sum_{i=0}^n \theta_{k_i} x_{k_i}, \quad \theta_{k_i} \geq 0, \quad \sum_{i=0}^n \theta_{k_i} = 1, \quad x_{k_i} \in X.$$

Ahora bien, el conjunto $\{(\theta_0, \dots, \theta_n) : \theta_i \geq 0, \sum_{i=0}^n \theta_i = 1\}$ es compacto, al igual que X . Entonces existe una sucesión $\{k_n\}$ tal que los límites $\lim_{j \rightarrow \infty} \theta_{k_j i} = \theta_i$ y $\lim_{j \rightarrow \infty} x_{k_j i} = x_i$ existen para $i = 0, \dots, n$, con $x_i \in X$. Se tiene entonces que la subsucesión $\{v_{k_j}\}$ converge a $\sum_{i=0}^n \theta_i x_i \in H(X)$, con lo cual demostramos que $H(X)$ es compacto. ■

Definición 6. El *producto interior* de dos vectores x y y es un número real denotado por $\langle x, y \rangle$ el cual cumple lo siguiente:

- (i) $\langle x, x \rangle > 0$, con igualdad si y sólo si $x = 0$

- (ii) $\langle x, y \rangle = \langle y, x \rangle$
 (iii) $\langle x, \lambda y + \mu z \rangle = \lambda \langle x, y \rangle + \mu \langle x, z \rangle$

En los teoremas 1.8 y 1.9 utilizaremos la norma definida por el producto interior, $\|x\| = \sqrt{\langle x, x \rangle}$.

Teorema 1.8 *Todo subconjunto cerrado y convexo de un espacio euclideo de dimensión finita posee un punto de norma mínima.*

Demostración. Sea K un subconjunto cerrado y convexo de un espacio euclideo de dimensión finita. Sea $\{x_i\} \subset K$ tal que $\lim_{i \rightarrow \infty} \|x_i\| = \inf_{x \in K} \|x\| = d$. Por la ley del paralelogramo,

$$\|x_i + x_j\|^2 + \|x_i - x_j\|^2 = 2\|x_i\|^2 + 2\|x_j\|^2,$$

o bien,

$$\|x_i - x_j\|^2 = 2\|x_i\|^2 + 2\|x_j\|^2 - 4\left\|\frac{1}{2}(x_i + x_j)\right\|^2$$

Como K es convexo, $\frac{1}{2}(x_i + x_j) \in K$, así que $\left\|\frac{1}{2}(x_i + x_j)\right\| \geq d$ luego

$$\|x_i - x_j\|^2 \leq 2\|x_i\|^2 + 2\|x_j\|^2 - 4d^2. \quad (1.5.5)$$

Cuando i, j tienden a infinito el lado derecho de (1.5.5) tiende a 0, por lo que $\{x_i\}$ es de Cauchy y converge a un punto x . Como K es cerrado, $x \in K$. Por la continuidad de la norma, $d = \lim_{i \rightarrow \infty} \|x_i\| = \left\|\lim_{i \rightarrow \infty} x_i\right\| = \|x\|$. ■

Teorema 1.9 (Teorema sobre desigualdades lineales) *Sea U un subconjunto compacto de R^n . El sistema de desigualdades lineales*

$$u_1^i z_1 + \cdots + u_n^i z_n > 0, \quad i = 1, \dots, m, \quad u^i = (u_1^i, \dots, u_n^i) \in U, \quad (1.5.6)$$

es inconsistente si y sólo si $0 \in H(U)$.

Demostración. Si $0 \in H(U)$, $0 = \sum_{i=1}^m \lambda_i u^i$, $u^i \in U$, $\lambda_i \geq 0$, $\sum_{i=1}^m \lambda_i = 1$, entonces para toda $z = (z_1, \dots, z_n) \in R^n$:

$$\begin{aligned} \sum_{i=1}^m \lambda_i (u_1^i z_1 + \cdots + u_n^i z_n) &= \sum_{i=1}^m (\lambda_i u_1^i z_1 + \cdots + \lambda_i u_n^i z_n) \\ &= \left(\sum_{i=1}^m \lambda_i u_1^i \right) z_1 + \cdots + \left(\sum_{i=1}^m \lambda_i u_n^i \right) z_n = 0. \end{aligned}$$

Esta ecuación no se cumple si $u_1^i z_1 + \cdots + u_n^i z_n > 0$ para $i = 1, \dots, m$. Luego el sistema (1.5.6) es inconsistente.

Supongamos ahora que $0 \notin H(U)$. Por el teorema anterior, existe un punto $z \in H(U)$ de norma mínima. Sea $u \in U \subset H(U)$, por convexidad $\theta u + (1-\theta)z \in H(U)$ para $0 \leq \theta \leq 1$; se tiene entonces que

$$\begin{aligned} 0 &\leq \|\theta u + (1-\theta)z\|^2 - \|z\|^2 = \|\theta(u-z) + z\|^2 - \|z\|^2 \\ &= \theta^2 \|u-z\|^2 + 2\theta \langle u-z, z \rangle + \|z\|^2 - \|z\|^2 \\ &= \theta^2 \|u-z\|^2 + 2\theta \langle u-z, z \rangle. \end{aligned} \quad (5)$$

Para valores pequeños de θ , ésta desigualdad implica que $\langle u-z, z \rangle \geq 0$, o bien, $\langle u, z \rangle \geq \langle z, z \rangle > 0$ y por lo tanto, z es la solución del sistema. ■

Teorema 1.10 *Dadas las funciones $f, g_1, \dots, g_n \in C[X]$ (X es un espacio métrico compacto) y sea $r = \sum_{i=1}^n c_i g_i - f$, para que los coeficientes c_1, \dots, c_n satisfagan que $\|r\|$ es mínima, es necesario y suficiente que $0 \in H(\bar{R})$, donde $\bar{R} = \{r(x)\hat{x} : |r(x)| = \|r\|\}$, con $\hat{x} = (g_1(x), \dots, g_n(x))$ ($\|\cdot\|$ es la norma uniforme definida en (1.2.4)).*

Demostración. Supongamos que $\|r\|$ no es mínima. Entonces para algún n -vector d tenemos que

$$\left\| \sum_{i=1}^n (c_i - d_i) g_i - f \right\| < \left\| \sum_{i=1}^n c_i g_i - f \right\|. \quad (1.5.8)$$

Sea $X_0 = \{x \in X : |r(x)| = \|r\|\}$. Para $x \in X_0$ debido a (1.5.8)

$$\left(r(x) - \sum_{i=1}^n d_i g_i \right)^2 < r(x)^2.$$

Desarrollando el primer término de ésta desigualdad

$$r(x)^2 - 2r(x) \sum_{i=1}^n d_i g_i + \left(\sum_{i=1}^n d_i g_i \right)^2 < r(x)^2,$$

y reacomodando los términos,

$$0 < \left(\sum_{i=1}^n d_i g_i \right)^2 < 2r(x) \sum_{i=1}^n d_i g_i$$

de donde obtenemos el sistema de desigualdades lineales que d debe satisfacer:

$$\sum_{i=1}^n (r(x)g_i(x)) d_i > 0, \quad x \in X_0. \quad (1.5.9)$$

Veamos que \bar{R} es compacto; primeramente X_0 es compacto, pues dada una sucesión convergente $s_n \in X_0$, $s_n \rightarrow s$, se tiene que

$$\|r\| = \lim_{n \rightarrow \infty} \|r\| = \lim_{n \rightarrow \infty} |r(s_n)| = \left| r \left(\lim_{n \rightarrow \infty} s_n \right) \right| = |r(s)|,$$

así $s \in X_0$, luego X_0 es un subconjunto cerrado de un conjunto compacto, por lo tanto X_0 es compacto.

Ahora bien, $h(x) = r(x)\hat{x} = (r(x)g_1(x), \dots, r(x)g_n(x))$ es una función continua que va de X a R^n , como $X_0 \subset X$ es compacto, entonces $h(X_0) = \bar{R}$ es compacto. Por (1.5.9) y por el teorema 1.9, $0 \notin H(\bar{R})$.

Para el recíproco, supongamos $0 \notin H(\bar{R})$. Por el teorema 1.9, existe un n -vector d tal que la desigualdad (1.5.9) se cumple para $x \in X_0$. Hagamos

$$\varepsilon = \min_{x \in X_0} r(x) \sum_{i=1}^n d_i g_i(x), \quad (1.5.10)$$

este mínimo es alcanzado en X_0 ya que éste es compacto, además $\varepsilon > 0$ por (1.5.9). Definamos $X_1 = \{x \in X : r(x) \sum_{i=1}^n d_i g_i(x) \leq \frac{\varepsilon}{2}\}$, X_1 es compacto y no contiene puntos de X_0 , observando que cualquier punto de X_1 arroja un valor menor que ε para la función en (1.5.10). La función $|r(x)|$ alcanza su máximo valor E sobre X_1 y $E < \|r\|$ pues $|r(x)| = \|r\|$ sólo para puntos de X_0 .

Sea $x \in X_1$ y supongamos que

$$0 < \lambda < \frac{\|r\| - E}{\|\sum_{i=1}^n d_i g_i(x)\|}.$$

Entonces

$$\begin{aligned} \left| r(x) - \lambda \sum_{i=1}^n d_i g_i(x) \right| &\leq |r(x)| + \lambda \left| \sum_{i=1}^n d_i g_i(x) \right| \\ &\leq E + \lambda \left\| \sum_{i=1}^n d_i g_i(x) \right\| \\ &< E + \frac{\|r\| - E}{\|\sum_{i=1}^n d_i g_i(x)\|} \left\| \sum_{i=1}^n d_i g_i(x) \right\| = \|r\|. \end{aligned}$$

Sea ahora $x \notin X_1$ y supongamos que

$$0 < \lambda < \frac{\varepsilon}{\|\sum_{i=1}^n d_i g_i(x)\|^2}.$$

Como $x \notin X_1$, x cumple

$$r(x) \sum_{i=1}^n d_i g_i(x) > \frac{\varepsilon}{2},$$

luego multiplicando esta desigualdad por 2λ ,

$$\lambda \varepsilon < 2\lambda r(x) \sum_{i=1}^n d_i g_i(x).$$

Entonces

$$\begin{aligned}
\left(r(x) - \lambda \sum_{i=1}^n d_i g_i(x)\right)^2 &= r(x)^2 - 2\lambda r(x) \sum_{i=1}^n d_i g_i(x) + \lambda^2 \left(\sum_{i=1}^n d_i g_i(x)\right)^2 \\
&< r(x)^2 - \lambda \varepsilon + \lambda^2 \left(\sum_{i=1}^n d_i g_i(x)\right)^2 \\
&\leq \|r\|^2 + \lambda \left(-\varepsilon + \lambda \left\|\sum_{i=1}^n d_i g_i(x)\right\|^2\right) \\
&< \|r\|^2 + \lambda \left(-\varepsilon + \frac{\varepsilon}{\left\|\sum_{i=1}^n d_i g_i(x)\right\|^2} \left\|\sum_{i=1}^n d_i g_i(x)\right\|^2\right) \\
&= \|r\|^2.
\end{aligned}$$

Así, hemos demostrado que para algún $\lambda > 0$, $\|r - \sum_{i=1}^n \lambda d_i g_i\| < \|r\|$, luego $\|r\|$ no es mínima. ■

Definición 7. Se dice que el conjunto de funciones $\{g_1, \dots, g_n\}$ satisface la *condición de Haar* sobre X si cada $g_i \in C[X]$ y todo conjunto de n vectores de la forma $\hat{x} = (g_1(x), \dots, g_n(x))$ es linealmente independiente. Dicho en otra forma, si para cualquier selección de puntos distintos $x_i \in X$, $i = 1, \dots, n$, el determinante

$$D[x_1, \dots, x_n] = \begin{vmatrix} g_1(x_1) & \cdots & g_n(x_1) \\ \vdots & \ddots & \vdots \\ g_1(x_n) & \cdots & g_n(x_n) \end{vmatrix}$$

es diferente de cero.

Lema 1.3 Consideremos $\{g_1, \dots, g_n\}$ un conjunto de funciones que satisfacen la condición de Haar sobre $[a, b]$. Sean los puntos $a \leq x_0 < x_1 < \dots < x_n \leq b$, y sean los números $\lambda_0, \dots, \lambda_n$ diferentes de cero. Entonces $\lambda_i \lambda_{i-1} < 0$ para $i = 1, \dots, n$, si y sólo si $0 \in H(A)$, donde

$$A = \{\lambda_i \hat{x}_i : \hat{x}_i = (g_1(x_i), \dots, g_n(x_i)), i = 1, \dots, n\}.$$

Demostración. Todos los determinantes

$$D[x_1, \dots, x_n] = \begin{vmatrix} g_1(x_1) & \cdots & g_n(x_1) \\ \vdots & \ddots & \vdots \\ g_1(x_n) & \cdots & g_n(x_n) \end{vmatrix}$$

con $x_1 < \dots < x_n$ tienen el mismo signo. Supongamos que existen $x = (x_1, \dots, x_n)$ con $x_1 < \dots < x_n$ y $y = (y_1, \dots, y_n)$ con $y_1 < \dots < y_n$, tales que $D[x_1, \dots, x_n] < 0 < D[y_1, \dots, y_n]$, entonces para algún $\lambda \in (0, 1)$,

$$D[\lambda x_1 + (1 - \lambda) y_1, \dots, \lambda x_n + (1 - \lambda) y_n] = 0,$$

por la condición de Haar esto implica que para algún i y para algún j , $j \neq i$,

$$\lambda x_i + (1 - \lambda) y_i = \lambda x_j + (1 - \lambda) y_j,$$

es decir,

$$\lambda (x_i - x_j) = (\lambda - 1) (y_i - y_j).$$

Entonces $(x_i - x_j)$ y $(y_i - y_j)$ tienen signos opuestos, lo cual es una contradicción.

Supongamos que $0 \in H(A)$, entonces $\sum_{i=0}^n \theta_i \lambda_i \hat{x}_i = 0$ con $\theta_i > 0$; despejando \hat{x}_0 ,

$$\hat{x}_0 = \sum_{i=1}^n \frac{-\theta_i \lambda_i}{\theta_0 \lambda_0} \hat{x}_i,$$

y resolviendo este sistema por la regla de Cramer,

$$\frac{-\theta_i \lambda_i}{\theta_0 \lambda_0} = \frac{D[x_1, \dots, x_{i-1}, x_0, x_{i+1}, \dots, x_n]}{D[x_1, \dots, x_n]}. \quad (1.5.11)$$

Ahora bien,

$$\operatorname{sgn} \left(\frac{-\theta_i \lambda_i}{\theta_0 \lambda_0} \right) = (-1)^{i-1} \operatorname{sgn} \left(\frac{D[x_0, x_2, \dots, x_n]}{D[x_1, \dots, x_n]} \right)$$

(haciendo $i-1$ intercambios de columna en el determinante, con lo cual se tienen $i-1$ cambios de signo de éste). Pero

$$\operatorname{sgn}(D[x_0, x_2, \dots, x_n]) = \operatorname{sgn}(D[x_1, x_2, \dots, x_n]),$$

luego el signo de $\left(\frac{D[x_0, x_2, \dots, x_n]}{D[x_1, \dots, x_n]} \right)$ es positivo, de donde $\operatorname{sgn} \left(\frac{-\theta_i \lambda_i}{\theta_0 \lambda_0} \right) = (-1)^{i-1}$, es decir, $-\theta_i \operatorname{sgn} \lambda_i = (-1)^{i-1} \operatorname{sgn} \lambda_0$, o bien, $\operatorname{sgn} \lambda_i = (-1)^i \operatorname{sgn} \lambda_0$.

Recíprocamente, supongamos que se cumple esta última igualdad, tomemos un $\theta_0 > 0$, y definamos θ_i a partir de (1.5.11). Tendremos entonces que

$$\sum_{i=0}^n \theta_i \lambda_i \hat{x}_i = 0,$$

dividiendo entre $\theta_1 + \dots + \theta_n$ se tiene el resultado deseado. ■

Teorema 1.11 (Teorema de alternancia) Sean $\{g_1, \dots, g_n\}$ un conjunto de funciones que satisfacen la condición de Haar sobre $[a, b]$, y sea X un subconjunto cerrado de $[a, b]$. El polinomio generalizado $P = \sum_{i=1}^n c_i g_i$ es un mejor aproximante para $f \in C[X]$ si y sólo si la función de error $r = f - P$ tiene al menos $n + 1$ alternancias sobre X ; esto es,

$$r(x_i) = -r(x_{i-1}) = \pm \|r\|, \quad i = 1, \dots, n, \quad \text{con } x_0 < x_1 < \dots < x_n, \quad x_i \in X$$

($\|\cdot\|$ es la norma uniforme definida en (1.2.4)).

Demostración. Por el teorema 1.10 $\|r\|$ es mínima si y sólo si $0 \in R^n$ está en la cápsula convexa del conjunto

$$\{r(x)\hat{x} : |r(x)| = \|r\|\}, \quad \text{con } \hat{x} = (g_1(x), \dots, g_n(x)).$$

Por el teorema 1.7, 0 se puede expresar como una combinación lineal de $n + 1$ de estos elementos,

$$0 = \sum_{i=0}^n \theta_i r(x_i) \hat{x}_i. \quad (1.5.12)$$

Ordenando los puntos x_i , tenemos $a \leq x_0 < \dots < x_n \leq b$, luego por el lema 1.3, (1.5.12) se cumple si y sólo si los números $\theta_i r(x_i)$ alternan de signo, es decir, $r(x_i) = -r(x_{i-1}) = \pm \|r\|$, $i = 1, \dots, n$. ■

Teorema 1.12 (Teorema de unicidad) Supongamos que $\{g_1, \dots, g_n\}$ es un conjunto de funciones que satisfacen la condición de Haar sobre $[a, b]$, entonces el polinomio generalizado $P = \sum_{i=1}^n c_i g_i$ que mejor aproxima a una función continua $f \in C[a, b]$ es único.

Demostración. Supongamos que el polinomio de mejor aproximación no es único, entonces existen $P = \sum_{i=1}^n c_i g_i$ y $Q = \sum_{i=1}^n d_i g_i$ polinomios generalizados de mejor aproximación a f , luego $\|f - P\| = \|f - Q\|$; veamos que el polinomio generalizado

$$\frac{P + Q}{2}$$

también es de mejor aproximación,

$$\begin{aligned} \|f - P\| &\leq \left\| f - \frac{P + Q}{2} \right\| = \left\| \frac{1}{2}(f - P) + \frac{1}{2}(f - Q) \right\| \\ &\leq \frac{1}{2}(\|f - P\| + \|f - Q\|) = \frac{1}{2}(\|f - P\| + \|f - P\|) = \|f - P\|. \end{aligned}$$

Por el teorema 1.11, existen puntos $x_0 < x_1 < \dots < x_n$ en $[a, b]$ tales que

$$f(x_i) - \frac{1}{2}(P + Q)(x_i) = (-1)^i \epsilon, \quad \text{donde } |\epsilon| = \|f - P\|,$$

luego

$$\frac{1}{2}(f(x_i) - P(x_i)) + \frac{1}{2}(f(x_i) - Q(x_i)) = (-1)^i \epsilon,$$

como ningún sumando en esta igualdad es mayor que $\frac{1}{2} |\epsilon|$, entonces

$$f(x_i) - P(x_i) = f(x_i) - Q(x_i) = (-1)^i \epsilon,$$

de aquí,

$$P(x_i) = Q(x_i), \quad i = 0, \dots, n,$$

es decir,

$$\sum_{j=1}^n c_j g_j(x_i) = \sum_{j=1}^n d_j g_j(x_i), \quad i = 0, \dots, n,$$

o bien,

$$\sum_{j=1}^n (c_j - d_j) g_j(x_i) = 0, \quad i = 0, \dots, n, \text{ donde } c_k - d_k \neq 0 \text{ para alguna } 1 \leq k \leq n,$$

lo que significa que tenemos un conjunto de n vectores:

$$\left\{ \begin{bmatrix} g_1(x_1) \\ \vdots \\ g_n(x_1) \end{bmatrix}, \dots, \begin{bmatrix} g_1(x_n) \\ \vdots \\ g_n(x_n) \end{bmatrix} \right\}$$

linealmente dependientes, contradiciendo la hipótesis de la condición de Haar sobre $\{g_1, \dots, g_n\}$. ■

1.6 Aproximación por polinomios generalizados sobre un conjunto finito de puntos.

Cuando deseamos encontrar el mejor aproximante a una función $f \in C[a, b]$ podemos discretizar el intervalo tomando un subconjunto finito de puntos Y en $[a, b]$. A continuación, establecemos algunos resultados que muestran que la aproximación obtenida sobre Y (bajo cierta condición) aproxima al mejor aproximante sobre el intervalo original.

Definición 8. Sea $Y \subset [a, b]$, denotamos la *densidad* de Y por $|Y|$ y la definimos por la ecuación

$$|Y| = \max_{x \in [a, b]} \inf_{y \in Y} |x - y|.$$

Lema 1.4 Sean $\{g_1, \dots, g_n\}$ un conjunto de funciones continuas sobre $[a, b]$, entonces existe $\alpha(\delta)$ tal que

$$\|P\| < \alpha(\delta) \|P\|_Y, \quad \alpha(\delta) > 1,$$

para todo polinomio generalizado $P = \sum_{i=1}^n c_i g_i$, y para todo $Y \subset [a, b]$ tal que

$$|Y| < \delta, \delta > 0. \quad (1.6.1)$$

$$(\|P\| = \max_{x \in [a, b]} |P(x)|, \|P\|_Y = \max_{y \in Y} |P(y)|).$$

Demostración. La independencia de $\{g_1, \dots, g_n\}$ puede ser supuesta, pues si no es linealmente independiente, podemos reemplazar $\{g_1, \dots, g_n\}$ por $\{g_1, \dots, g_k\}$ linealmente independiente con $k < n$. Dado un polinomio generalizado cualquiera $P = \sum_{i=1}^n c_i g_i$.

Sea θ el mínimo de $\|\sum_{i=1}^n c'_i g_i\|$ sobre el conjunto compacto en R^n definido por la ecuación $\sum_{i=1}^n |c'_i| = 1$. $\theta > 0$, ya que si $\theta = 0$, entonces $\sum_{i=1}^n c'_i g_i = 0$ contradiciendo la independencia de $\{g_1, \dots, g_n\}$. Como

$$\theta \leq \left\| \sum_{i=1}^n \frac{c_i}{\sum_{i=1}^n |c_i|} g_i \right\|,$$

entonces

$$\theta \sum_{i=1}^n |c_i| \leq \left\| \sum_{i=1}^n c_i g_i \right\| = \|P\|.$$

Definamos una función Ω sobre los números reales positivos:

$$\Omega(\delta) = \max_{1 \leq i \leq n} \max_{|x-y| \leq \delta} |g_i(x) - g_i(y)|.$$

Cuando $\delta \rightarrow 0$, $g_i(x) - g_i(y) \rightarrow 0$ porque las funciones g_i son uniformemente continuas, luego $\Omega(\delta) \rightarrow 0$; podemos tomar entonces un δ suficientemente pequeño para que $\Omega(\delta) < \theta$.

Sea $P = \sum_{i=1}^n c_i g_i$ un polinomio generalizado cualquiera, sea $x \in [a, b]$ tal que $|P(x)|$ es máximo, si para todo $y \in Y$, $|x - y| > \delta$ entonces para este $x \in [a, b]$ se tiene que $\inf_{y \in Y} |x - y| \geq \delta$, contradiciendo (1.6.1), luego podemos tomar un $y \in Y$ tal que $|x - y| \leq \delta$. Entonces,

$$\begin{aligned}
\theta \sum_{i=1}^n |c_i| &\leq \|P\| = |P(x)| \\
&\leq |P(x) - P(y)| + |P(y)| \\
&= \left| \sum_{i=1}^n c_i (g_i(x) - g_i(y)) \right| + |P(y)| \\
&\leq \sum_{i=1}^n |c_i| |g_i(x) - g_i(y)| + \|P\|_Y \\
&\leq \Omega(\delta) \sum_{i=1}^n |c_i| + \|P\|_Y,
\end{aligned}$$

de aquí,

$$\sum_{i=1}^n |c_i| \leq \frac{\|P\|_Y}{\theta - \Omega(\delta)},$$

y

$$\begin{aligned}
\|P\| &\leq \Omega(\delta) \frac{\|P\|_Y}{\theta - \Omega(\delta)} + \|P\|_Y \\
&= \left[1 + \frac{\Omega(\delta)}{\theta - \Omega(\delta)} \right] \|P\|_Y.
\end{aligned}$$

Haciendo $\alpha(\delta) = 1 + \frac{\Omega(\delta)}{\theta - \Omega(\delta)}$ se tiene el resultado, pues $\alpha(\delta) \rightarrow 1^+$ cuando $\delta \rightarrow 0$. ■

Lema 1.5 Consideremos las funciones $f, g_1, \dots, g_n \in C[a, b]$. Sea $Y \subset [a, b]$ tal que $|Y| < \delta$, $\delta > 0$. Para cualquier polinomio generalizado $P = \sum_{i=1}^n c_i g_i$,

$$\|f - P\| \leq \|f - P\|_Y + \omega(f; [a, b]; \delta) + \beta \|P\| \Omega(\delta),$$

donde β es independiente de f y P y $\omega(f; [a, b]; \delta)$ es el módulo de continuidad de f .

Demostración. Como en el lema anterior, podemos suponer la independencia lineal de $\{g_1, \dots, g_n\}$. De igual manera definamos a θ como en ese lema, entonces $\theta \sum_{i=1}^n |c_i| \leq \|\sum_{i=1}^n c_i g_i\| = \|P\|$ para cualesquiera c_i . Tomemos cualquier polinomio generalizado $P = \sum_{i=1}^n c_i g_i$. Sea $x \in [a, b]$ tal que $|f(x) - P(x)|$ es máximo, de igual forma que en el lema anterior podemos tomar un $y \in Y$ tal

que $|x - y| \leq \delta$. Entonces,

$$\begin{aligned}
\|f - P\| &= |f(x) - P(x)| \leq |f(x) - f(y)| + |f(y) - P(y)| + |P(y) - P(x)| \\
&\leq \omega(f; [a, b]; \delta) + \|f - P\|_Y + \sum_{i=1}^n |c_i| |g_i(y) - g_i(x)| \\
&\leq \omega(f; [a, b]; \delta) + \|f - P\|_Y + \left(\sum_{i=1}^n |c_i| \right) \max_{1 \leq i \leq n} |g_i(y) - g_i(x)| \\
&\leq \|f - P\|_Y + \omega(f; [a, b]; \delta) + \frac{1}{\theta} \|P\| \Omega(\delta). \quad \blacksquare
\end{aligned}$$

Teorema 1.13 Consideremos las funciones $f, g_1, \dots, g_n \in C[X]$, con $X = [a, b]$. Para cada $Y \subseteq X$ sea P_Y el polinomio generalizado de mejor aproximación a f sobre Y , supongamos que $|Y| < \delta, \delta > 0$. Entonces $\|f - P_Y\| \rightarrow \|f - P_X\|$ cuando $\delta \rightarrow 0$.

Demostración. Primero tenemos que $\|f - P_Y\|_Y \leq \|f - P_X\|_Y$ pues P_Y es el mejor aproximante sobre Y . Luego

$$\|f - P_Y\|_Y \leq \|f - P_X\|_Y \leq \|f - P_X\|.$$

Por el lema 1.5,

$$\begin{aligned}
\|f - P_Y\| - \|f - P_X\| &\leq \|f - P_Y\| - \|f - P_Y\|_Y \\
&\leq \omega(f; [a, b]; \delta) + \beta \|P_Y\| \Omega(\delta). \quad (6)
\end{aligned}$$

Ahora bien,

$$\|P_Y\|_Y \leq \|P_Y - f\|_Y + \|f\|_Y,$$

pero $\|P_Y - f\|_Y \leq \|0 - f\|_Y$, ya que P_Y es el mejor aproximante sobre Y , luego

$$\|P_Y\|_Y \leq \|P_Y - f\|_Y + \|f\|_Y \leq \|0 - f\|_Y + \|f\|_Y \leq 2\|f\|.$$

Por el lema 1.4,

$$\|P_Y\| \leq \alpha \|P_Y\|_Y \leq 2\alpha \|f\|. \quad (1.6.3)$$

Combinando (1.6.2) y (1.6.3) tenemos

$$\|f - P_Y\| - \|f - P_X\| \leq \omega(f; [a, b]; \delta) + 2\beta\alpha \|f\| \Omega(\delta),$$

cuando $\delta \rightarrow 0$ el lado derecho de esta desigualdad tiende a cero, con lo cual se tiene el resultado. \blacksquare

Teorema 1.14 *Supongamos que $\{g_1, \dots, g_n\}$ es un conjunto de funciones que satisfacen la condición de Haar sobre $[a, b]$. Sea P_0 el polinomio generalizado de mejor aproximación a una función continua f . Entonces existe una constante $\gamma > 0$ dependiente de f tal que para cualquier polinomio generalizado P ,*

$$\|f - P\| \geq \|f - P_0\| + \gamma \|P_0 - P\|. \quad (1.6.4)$$

Demostración. Si $\|f - P_0\| = 0$, entonces $\gamma = 1$ pues $\|P_0 - P\| \leq \|f - P_0\| + \|P - f\| = \|f - P\|$. Supongamos $\|f - P_0\| > 0$. Por el teorema 1.11 existen puntos x_0, \dots, x_k y signos $\sigma_0, \dots, \sigma_k$ tales que

$$f(x_i) - P_0(x_i) = \sigma_i \|f - P_0\|;$$

por el lema 1.3 se tiene también que 0 está en la cápsula convexa $H(A)$ donde $A = \{\sigma_i \hat{x}_i : \hat{x}_i = (g_1(x_i), \dots, g_n(x_i))\}$, luego $\sum_{i=0}^n \theta_i \sigma_i \hat{x}_i = 0$ con $\theta_i > 0$, por la condición de Haar, $k \geq n$. Sea Q un polinomio generalizado cualquiera, tenemos que

$$\sum_{i=0}^n \theta_i \sigma_i Q(x_i) = \sum_{i=0}^n \left(\theta_i \sigma_i \sum_{j=1}^n c_j g_j(x_i) \right) = \sum_{j=1}^n \left(c_j \sum_{i=0}^n \theta_i \sigma_i g_j(x_i) \right) = 0.$$

Por la condición de Haar no todos los números $\sigma_i Q(x_i)$ son cero. Como todos los $\theta_i > 0$ debemos tener que $\sigma_i Q(x_i)$ es positivo para alguna i , por tanto $\max_{0 \leq i \leq n} \sigma_i Q(x_i)$ es una función positiva de Q . Luego el número

$$\gamma = \min_{\|Q\|=1} \max_{0 \leq i \leq n} \sigma_i Q(x_i)$$

es positivo, γ es alcanzado pues es el mínimo de una función continua sobre un conjunto compacto.

Sea P un polinomio generalizado cualquiera, si $P = P_0$, entonces (1.6.4) es obvia. En otro caso hagamos

$$Q = \frac{(P_0 - P)}{\|P_0 - P\|},$$

como $\|Q\| = 1$ existe un índice i tal que $\sigma_i Q(x_i) \geq \gamma$, es decir $\sigma_i (P_0 - P) \geq \gamma \|P_0 - P\|$. Luego

$$\begin{aligned} \|f - P\| &= \max_{x \in [a, b]} |f(x) - P(x)| \geq \sigma_i (f - P)(x_i) \\ &= \sigma_i (f - P_0)(x_i) + \sigma_i (P_0 - P)(x_i) \\ &= \sigma_i^2 \|f - P_0\| + \sigma_i (P_0 - P)(x_i) \\ &\geq \|f - P_0\| + \gamma \|P_0 - P\|. \end{aligned}$$

■

Teorema 1.15 *Supongamos que $\{g_1, \dots, g_n\}$ es un conjunto de funciones que satisfacen la condición de Haar sobre $X = [a, b]$, sea $f \in C[X]$. Si f tiene un único polinomio generalizado P_X de mejor aproximación sobre X , entonces sus mejores aproximantes P_Y sobre subconjuntos Y convergen a P_X cuando $|Y| \rightarrow 0$.*

Demostración. Por el teorema anterior tomando $P_0 = P_X$ y $P = P_Y$

$$\|f - P_Y\| \geq \|f - P_X\| + \gamma \|P_X - P_Y\|,$$

es decir,

$$\|P_Y - P_X\| \leq \gamma^{-1} (\|f - P_Y\| - \|f - P_X\|),$$

por el teorema 1.13 el lado derecho de esta desigualdad tiende a cero cuando $|Y| \rightarrow 0$, con lo cual se tiene el resultado. ■

En los teoremas anteriores a la discretización Y se le pide la condición

$$\max_{x \in [a, b]} \min_{y \in Y} |x - y| < \delta, \delta > 0.$$

Tomando Y como el conjunto de m puntos

$$\left\{ a + \frac{(b-a)j}{m-1}, \quad j = 0, \dots, m-1 \right\},$$

tenemos que $\max_{x \in [a, b]} \min_{y \in Y} |x - y| < \delta = \frac{b-a}{m-1}$, así, δ tiende a cero cuando m tiende a infinito.

Capítulo 2

Optimización lineal

En este capítulo presentamos un resumen de algunos resultados básicos de la optimización lineal con restricciones, desarrollamos dos métodos computacionales para resolver programas lineales, uno cuando la formulación es como un programa lineal con un número finito de restricciones y otro cuando se trata el problema con un número infinito de éstas. Estos métodos incorporan una implementación del método Simplex en la etapa de optimización, así que también estudiamos la teoría de este método y describimos la forma en la que se implementó.

2.1 El Problema de Optimización Lineal

Definición 9. Sea $M \subset R$ un conjunto fijo y f una función real definida sobre M . Buscamos un elemento \bar{x} en M tal que

$$f(\bar{x}) \leq f(x) \quad \text{para todo } x \in M.$$

M es llamado *conjunto factible*, f *función objetivo* y \bar{x} es denominado *punto óptimo*.

Definición 10. Al número v dado por

$$v = \{\inf f(x) : x \in M\}$$

se le denomina *valor* del correspondiente problema de optimización.

Definición 11. Un elemento $x \in M$ es llamado *punto factible*.

Si M es el conjunto vacío, es decir no hay puntos factibles, se dice que el problema es *inconsistente* y $v = \infty$. Si existen puntos factibles, el problema es *factible* o *consistente*. Si $v = -\infty$, se dice que el problema es *no acotado inferiormente*.

Entonces todo problema de minimización debe estar en uno y sólo uno de los siguientes tres estados:

- IC = *Inconsistente*; el conjunto factible es vacío y el valor del problema es ∞ .
- B = *Acotado*; existen puntos factibles y el valor es finito.
- UB = *No acotado*; hay puntos factibles, la función objetivo no es acotada inferiormente, y el valor es $-\infty$

El valor de un problema de maximización es $-\infty$ en el estado IC , finito en B , y $+\infty$ en el estado UB .

Un problema de optimización será llamado *problema de optimización lineal (POL)* cuando la función objetivo sea lineal y el conjunto factible esté definido mediante restricciones lineales.

En este caso la función objetivo tiene la forma

$$c^T y = \sum_{r=1}^n c_r y_r,$$

donde c es un vector fijo en R^n . El conjunto de vectores factibles de un (POL) es definido como una intersección de semiespacios; sea T un conjunto de índices dado *finito* o *infinito*; con cada $t \in T$ asociamos un vector $a(t) \in R^n$ y un número real $b(t)$. El conjunto de vectores factibles de un problema de optimización lineal consiste de todos los vectores $y \in R^n$ pertenecientes a todos los semiespacios

$$\{y : a(t)^T y \geq b(t)\}, \quad t \in T$$

Definición 12. Un *problema de optimización lineal* es definido de la siguiente forma:

Dados: Un vector $c = (c_1, \dots, c_n)^T \in R^n$, un conjunto de índices (finito o infinito) no vacío T , una función $a(t) = (a_1(t), \dots, a_n(t))^T$ que mapea T en R^n y una función escalar $b(t)$.

Buscamos un vector $y \in R^n$ que resuelva el problema:

$$(P) \text{ Minimizar } c^T y \text{ sujeto a las restricciones } a(t)^T y \geq b(t), \quad t \in T. \quad (2.1.1)$$

Cuando el conjunto T es infinito se tiene un número infinito de restricciones y el problema (P) es llamado *Programa Lineal Semi-infinito (PLSI)*. En el caso de que T sea finito se tiene un *Programa Lineal Ordinario (PLO)*.

2.2 Dos métodos computacionales para la resolución de Programas Lineales.

2.2.1 Programación Lineal Ordinaria

Cuando en el Programa Lineal (2.1.1) el conjunto de índices $T \subset R^k$ es un conjunto finito X_m de puntos se tiene un Programa Lineal Ordinario (PLO) pues el número de restricciones lineales es finito.

El programa lineal a resolver queda de la siguiente forma:

$$\begin{aligned} & \text{Minimizar } c^\top y \\ \text{s.a. } & a(t_j)^\top y \geq b(t_j), \quad t_j \in X_m, \quad j = 1, \dots, m \end{aligned} \quad (2.2.1.1)$$

Donde m es el número de elementos de X_m .

El programa (2.2.1.1) puede resolverse mediante el método Simplex, el cual implementamos en Matlab. En la sección 2.3 se estudia la teoría del método Simplex; se describe la forma en la que se implementó y los requerimientos especiales para poder ser aplicado a la resolución de programas lineales en la forma (2.2.1.1).

2.2.2 Programación Lineal Semi-infinita

Cuando en el Programa Lineal (2.1.1) el conjunto de índices T es infinito (por ejemplo, en el caso de una variable, un intervalo cerrado tiene infinitos puntos), se tiene un programa lineal semi-infinito; para resolver este programa utilizamos un Algoritmo que trabaja discretizando el conjunto T y resolviendo el subproblema resultante de la discretización, a continuación verifica si la solución óptima del problema discretizado es también solución del programa original, cuando esto no ocurre se construye una discretización más fina de T .

A continuación describimos la forma en que hacemos la discretización del conjunto T .

Definición. Dado $T = [x_1, x_2]$, una *red* del intervalo T es un subconjunto finito de puntos pertenecientes al intervalo.

Denotamos por T_r a una red de T que consiste de r puntos igualmente espaciados en T , esto es, los puntos

$$x_1 + \frac{(x_2 - x_1)j}{r - 1}, \quad j = 0, \dots, r - 1; \quad r > 1.$$

ó bien,

$$T_r = \left\{ x_1, x_1 + \frac{(x_2 - x_1)}{r - 1}, x_1 + \frac{(x_2 - x_1)2}{r - 1}, \dots, x_1 + \frac{(x_2 - x_1)(r - 2)}{r - 1}, x_2 \right\}.$$

Una red con $k(r-1) + 1$ puntos está dada por

$$T_{k(r-1)+1} = \left\{ x_1, x_1 + \frac{(x_2 - x_1)}{k(r-1)}, \dots, x_1 + \frac{(x_2 - x_1)(kr - k - 1)}{k(r-1)}, x_2 \right\}.$$

Definición 14. Una *sucesión expansiva* de conjuntos es una sucesión de conjuntos $\{X_j\} = X_0, X_1, \dots, X_j, X_{j+1}, \dots$, que cumplen $X_0 \subset X_1 \subset \dots \subset X_j \subset X_{j+1} \subset \dots$

Teorema 2.1 La sucesión de redes $\{T_{r_i}\}_{i=1}^{\infty}$ donde $r_i = k(r_{i-1} - 1) + 1$, $i = 2, \dots$, con $k \geq 2$, $r_1 > 1$ es una sucesión expansiva de conjuntos.

Demostración. Sea n un entero positivo, supongamos $x \in T_{r_n}$, entonces

$$x = x_1 + \frac{(x_2 - x_1)j}{r_n - 1}, \quad \text{para alguna } j, 0 \leq j \leq r_n - 1, \quad (2.2.2.1)$$

multiplicando numerador y denominador por k en el segundo sumando de la igualdad anterior

$$x = x_1 + \frac{(x_2 - x_1)kj}{k(r_n - 1)}, \quad \text{para alguna } j, 0 \leq j \leq r_n - 1, \quad (2.2.2.2)$$

sustituyendo kj por j' en (2.2.2.2) tenemos

$$x = x_1 + \frac{(x_2 - x_1)j'}{k(r_n - 1)}, \quad \text{para alguna } j', 0 \leq j' \leq k(r_n - 1),$$

de aquí $x \in T_{k(r_n-1)+1}$. Hemos demostrado que

$$T_{r_n} \subset T_{k(r_n-1)+1} = T_{r_{n+1}},$$

donde n es un entero positivo cualquiera, luego la sucesión de redes $\{T_r\}$ es una sucesión expansiva de conjuntos. ■

Llamamos a

$$\{T_{r_i}\}_{i=1}^{\infty} \quad \text{donde } r_i = k(r_{i-1} - 1) + 1, \quad i = 2, \dots, \quad (2.2.2.3)$$

sucesión expansiva de redes.

Describimos aquí el algoritmo para resolver (2.1.1) cuando T es un intervalo cerrado $[x_1, x_2]$:

Algoritmo 2.2.2. Dado el conjunto infinito T y dada una tolerancia $\varepsilon > 0$, inicializamos un índice $i = n$, donde n es el número de elementos de la primera discretización de T . Construimos la primera discretización T_i .

Primera Etapa:

Resolver el subproblema:

$$(P_i) \quad \text{Minimizar } c^\top y$$

$$\text{s.a. } a(t)^\top y \geq b(t), \quad t \in T_i.$$

Si (P_i) es inconsistente parar ((P) es inconsistente).
Si no, calcular una solución óptima de (P_i) , y^i .

Segunda etapa: Verificar la factibilidad de y^i para todo $t \in T$.

Para esto, encontramos s_i , el menor de los mínimos de la función:

$$\varphi(t) = a(t)^\top y^i - b(t), \quad t \in T.$$

Si $s_i \geq -\varepsilon$ parar (y^i es aceptada como solución óptima).

Si no, incrementar el índice i y regresar a la primera etapa.

Cuando en la segunda etapa se satisface la desigualdad $s_i \geq -\varepsilon$ la solución y^i es aceptada como solución del programa (2.1.1), cabe aclarar que la solución y^i arrojada por el algoritmo cumple las restricciones:

$$a(t)^\top y^i \geq b(t) - \varepsilon, \quad t \in T,$$

que para tolerancias ε muy pequeñas nos brindan una solución casi factible.

En la primera etapa cuando (P_i) es inconsistente (No existen puntos factibles) se tiene que para todo $y \in R^n$, existe $t' \in T_i$ tal que

$$a(t')^\top y < b(t') \tag{2.2.2.4}$$

luego como $T_i \subset T$, (P) es inconsistente.

Definición 18. Un *hiperplano* H en R^n es un conjunto $\{x \in R^n : a^\top x = c\}$, donde $c \in R$ y a es un vector distinto de cero en R^n . Un hiperplano es el conjunto de soluciones de una ecuación lineal.

La solución de un programa lineal, puede presentarse en alguna de las siguientes formas:

1. *Solución Óptima Única.* Veremos que ocurre siempre en un punto extremo de la región factible. Para este tipo de solución la región factible puede ser acotada o no acotada.

2. *Soluciones Óptimas Alternativas.* No existe un punto único para la solución, sino que un conjunto de puntos forman la solución del problema.

3. *Solución Óptima no acotada.* Se presenta cuando la región factible no está acotada y el hiperplano $cx = k$ (función objetivo) se puede mover indefinidamente en la dirección $-c$, intersecando siempre a la región factible.

4. *Región Factible Vacía.* En este caso, no existe solución para el sistema de ecuaciones o desigualdades que definen la región factible y se dice que el problema es *no factible*.

Definición 19. Dado un hiperplano H , llamaremos *semiespacios cerrados* a los conjuntos $H_+ = \{x \in R^n : a^\top x \geq c\}$ y $H_- = \{x \in R^n : a^\top x \leq c\}$.

Definición 20. Un politopo es un conjunto formado por la intersección de un número finito de semiespacios cerrados.

El conjunto $K = \{x \in R^n : Ax = b, x \geq 0\}$ de soluciones factibles de un programa lineal es un politopo, pues la ecuación

$$a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n = b_i,$$

es equivalente al sistema de desigualdades

$$\begin{aligned} a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n &\geq b_i \\ a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n &\leq b_i. \end{aligned}$$

Puntos extremos y soluciones básicas factibles.

Definición 21. Consideremos el sistema

$$\begin{aligned} Ax &= b \\ \text{y } x &\geq 0, \end{aligned}$$

en donde A es una matriz de $m \times n$ y b es un m -vector. Supongamos que $\text{rango}(A, b) = \text{rango}(A) = m$. Después de un posible reordenamiento de las columnas de A , sea $A = [B, N]$, en donde B es una matriz invertible de $m \times m$ y N es una matriz de $m \times (n - m)$. La solución

$$x = \begin{bmatrix} x_B \\ x_N \end{bmatrix}$$

de las ecuaciones $Ax = b$, en donde $x_B = B^{-1}b$ y $x_N = 0$ se denomina *solución básica* del sistema, B se denomina *matriz básica o base* y N se llama *matriz no básica*. Las componentes de x_B son llamadas *variables básicas*, y las de x_N *variables no básicas*.

Definición 22. Si una o más de las variables básicas de una solución básica de

$$\begin{aligned} Ax &= b \\ x &\geq 0, \end{aligned} \tag{7}$$

es cero, la solución es llamada *solución básica degenerada*.

Definición 23. Una solución básica de (2.3.1) en la que todos sus componentes son no negativos se denomina *solución básica factible*; si algún componente es cero, la solución básica factible se dice *solución básica factible degenerada*.

Teorema 2.2 (*Equivalencia entre puntos extremos y soluciones básicas*) Sean $A \in R^{m \times n}$ una matriz de rango m y $b \in R^m$. Sea P el polítopo convexo

$$P = \{x \in R^n : Ax = b, x \geq 0\}.$$

Un vector $x \in P$ es un punto extremo de P si y sólo si x es una solución básica de

$$\begin{aligned} Ax &= b \\ x &\geq 0. \end{aligned}$$

Es decir, si y sólo si los vectores columna de la matriz A asociados a las componentes positivas de x son linealmente independientes.

Demostración. Supongamos, sin pérdida de generalidad, que los m primeros componentes del vector x son positivos y los $n - m$ últimos cero. Si $x = [\bar{x}^T, 0^T]$, $\bar{x} > 0$, y nombramos \bar{A} las m primeras columnas de A , tenemos que $Ax = \bar{A}\bar{x} = b$.

Para demostrar que x es una solución básica debemos demostrar que los vectores columna de \bar{A} son linealmente independientes. Haremos esto por contradicción. Supongamos que las columnas de \bar{A} no son linealmente independientes. Entonces existe un vector $\bar{w} \neq 0$ tal que $\bar{A}\bar{w} = 0$. Luego $\bar{A}(\bar{x} \pm \varepsilon\bar{w}) = \bar{A}\bar{x} \pm \varepsilon\bar{A}\bar{w} = \bar{A}\bar{x} = b$ y, para un ε suficientemente pequeño tal que $(\bar{x} \pm \varepsilon\bar{w}) \geq 0$. Los puntos

$$y' = \begin{bmatrix} \bar{x} + \varepsilon\bar{w} \\ 0 \end{bmatrix} \quad y \quad y'' = \begin{bmatrix} \bar{x} - \varepsilon\bar{w} \\ 0 \end{bmatrix}$$

están en P . Tenemos entonces que $x = \frac{1}{2}(y' + y'')$, que describe a x como una combinación convexa de dos vectores distintos en P , por tanto x no puede ser un punto extremo de P . Hemos probado que si x es un punto extremo entonces las columnas de la matriz \bar{A} son linealmente independientes y x es una solución básica.

Supongamos ahora que x no es un punto extremo de P . Esto significa que x puede formularse como una combinación convexa de otros dos puntos en P , por ejemplo $x = \lambda y' + (1-\lambda)y''$, $y' \neq y''$ y $0 < \lambda < 1$. Como todas las componentes de x , y' , y'' son no negativas y $0 < \lambda < 1$, el resultado inmediato es que las últimas $n - m$ componentes de y' y y'' son cero pues lo son las de x , lo mismo sucede para $x - y'$. Ahora bien $A(x - y') = Ax - Ay' = b - b = 0$. Luego las columnas de la matriz \bar{A} son linealmente dependientes. Demostramos entonces que si las columnas de \bar{A} son linealmente independientes x es un punto extremo. ■

Corolario 2.2.1 *El conjunto $P = \{x \in R^n : Ax = b, x \geq 0\}$ tiene un número finito de puntos extremos.*

Demostración. El número máximo de formas de seleccionar m vectores básicos de las n columnas de A es

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}.$$

Luego el número de puntos extremos es a lo más $\binom{n}{m}$. ■

Consideremos un programa lineal en forma estándar

$$\begin{aligned} & \text{Min } cx \\ & \text{s.a } Ax = b \\ & \quad x \geq 0. \end{aligned} \tag{2.3.2}$$

Una solución factible que alcanza el valor mínimo de la función objetivo se denomina *solución factible óptima*. Si esta solución es básica, es una *solución factible básica óptima*.

Teorema 2.3 *Dado el programa lineal (2.3.2), donde A es una matriz de $m \times n$ de rango m :*

- i) Si hay una solución factible, hay una solución factible básica.*
- ii) Si hay una solución factible óptima, hay una solución factible básica óptima.*

Este teorema muestra la importancia de las soluciones básicas (puntos extremos) para resolver un problema de programación lineal, pues establece que al buscar una solución óptima sólo es necesario tener en cuenta las soluciones básicas, ya que en ese tipo de solución siempre se consigue el valor óptimo.

Basándose en estos resultados teóricos podemos abordar la descripción del *Método Simplex*. Para resolver un programa lineal en su forma estándar, basta con estudiar los puntos extremos del conjunto factible (que por el corolario 2.2.1 el número de éstos es finito) y buscar aquel en el que la función objetivo se hace mínima.

Descripción del Método Simplex.

Con el método simplex se puede deducir la optimalidad de un punto extremo solución dado basándonos en consideraciones locales sin tener que revisar globalmente todos los puntos extremos.

Consideremos el problema de programación lineal

$$(1) \begin{aligned} & \text{Min } cx \\ & \text{s.a } Ax = b \\ & \quad x \geq 0. \end{aligned}$$

en donde $A \in R^{m \times n}$ y $\text{rango}(A) = m$. Supongamos que se tiene una solución básica factible $\begin{pmatrix} B^{-1}b \\ 0 \end{pmatrix}$ con valor objetivo

$$z_0 = c \begin{pmatrix} B^{-1}b \\ 0 \end{pmatrix} = (c_B, c_N) \begin{pmatrix} B^{-1}b \\ 0 \end{pmatrix} = c_B B^{-1}b. \quad (2.3.3)$$

Denotemos por x_B y x_N a los vectores de variables básicas y de variables no básicas respectivamente. La factibilidad requiere que $x_B \geq 0$, $x_N \geq 0$ y $b = Ax = Bx_B + Nx_N$. Multiplicando por B^{-1} ésta ecuación tenemos

$$\begin{aligned} x_B &= B^{-1}b - B^{-1}Nx_N \\ &= B^{-1}b - \sum_{j \in I} B^{-1}a_j x_j \\ &= \bar{b} - \sum_{j \in I} y_j x_j, \end{aligned} \quad (8) \quad 2.3.4$$

donde I es el conjunto actual de índices de las variables no básicas, a_j es la j -ésima columna de A y $y_j = B^{-1}a_j$. De (2.3.3) y (2.3.4) tenemos

$$\begin{aligned} z &= cx \\ &= c_B x_B + c_N x_N \\ &= c_B \left(B^{-1}b - \sum_{j \in I} B^{-1}a_j x_j \right) + \sum_{j \in I} c_j x_j \\ &= c_B B^{-1}b - \sum_{j \in I} c_B B^{-1}a_j x_j + \sum_{j \in I} c_j x_j \\ &= z_0 - \sum_{j \in I} (c_B B^{-1}a_j - c_j) x_j \\ &= z_0 - \sum_{j \in I} (z_j - c_j) x_j, \end{aligned} \quad (9) \quad 2.3.5$$

donde $z_j = c_B B^{-1}a_j$ para todo $j \in I$ y $z_0 = c_B B^{-1}b$.

Reformulando el problema (1) con estas igualdades obtenemos

$$\begin{aligned}
(2) \quad & \text{Min } z = z_0 - \sum_{j \in I} (z_j - c_j) x_j \\
& \text{s.a. } \sum_{j \in I} y_j x_j + x_B = \bar{b} \\
& x_j \geq 0, j \in I \text{ y } x_B \geq 0.
\end{aligned}$$

Sin perder generalidad, suponemos que ningún renglón de la segunda ecuación de (2) $\sum_{j \in I} y_j x_j + x_B = \bar{b}$ tiene sólo ceros en las columnas de las variables no básicas $x_j, j \in I$. En caso contrario, podemos conocer el valor de la variable básica en tal renglón, y este renglón puede eliminarse del problema. Las variables x_B pueden verse como variables de holgura en ésta ecuación, luego, es posible reformular (2) en términos de las variables no básicas de la siguiente manera

$$\begin{aligned}
(3) \quad & \text{Min } z = z_0 - \sum_{j \in I} (z_j - c_j) x_j \\
& \text{s.a. } \sum_{j \in I} y_j x_j \leq \bar{b} \\
& x_j \geq 0, j \in I.
\end{aligned}$$

El número de variables no básicas es $p = n - m$. La representación (3) en la que la función objetivo z y las variables básicas han sido resueltas en términos de las variables no básicas se denomina *representación de la solución básica en forma canónica*. El resultado esencial se plantea de la siguiente manera:

Criterio de optimalidad

Si $(z_j - c_j) \leq 0$ para todo $j \in I$, la solución básica factible es óptima.
(2.3.6)

Esto es debido a que si $z_j - c_j \leq 0$ para todo $j \in I$, entonces $z \geq z_0$ para cualquier solución factible, y para la solución básica factible actual, $z = z_0$, pues $x_j = 0$ para todo $j \in I$.

Criterio para seleccionar variable de entrada

Consideremos la representación (2). Si $(z_j - c_j) \leq 0$ para todo $j \in I$, entonces $x_j = 0, j \in I$ y $x_B = \bar{b}$ es óptimo para (2) como se vió en (2.3.6). Si no, $p - 1$ variables no básicas se mantendrán fijas en cero, pues se escoge como

variable no básica a incrementar una de aquellas donde $z_k - c_k$ sea más grande. Entonces hagamos $x_j = 0$ para $j \in I - \{k\}$, de (2) obtenemos

$$z = z_0 - (z_k - c_k)x_k \quad (2.3.7)$$

y

$$\begin{bmatrix} x_{B_1} \\ x_{B_2} \\ \vdots \\ x_{B_r} \\ \vdots \\ x_{B_m} \end{bmatrix} = \begin{bmatrix} \bar{b}_1 \\ \bar{b}_2 \\ \vdots \\ \bar{b}_r \\ \vdots \\ \bar{b}_m \end{bmatrix} - \begin{bmatrix} y_{1k} \\ y_{2k} \\ \vdots \\ y_{rk} \\ \vdots \\ y_{mk} \end{bmatrix} x_k. \quad (2.3.8)$$

Criterio para seleccionar variable de salida

Si $y_{ik} \leq 0$, entonces x_{B_i} crece cuando x_k crece, así, x_{B_i} continúa siendo no negativo.

Si $y_{ik} > 0$, entonces x_{B_i} decrece cuando x_k crece. Como se debe satisfacer la no negatividad, x_k se incrementa hasta el primer punto donde una variable básica x_{B_r} es cero. De (2.3.8) $x_{B_i} = 0$ cuando $x_k = \frac{\bar{b}_i}{y_{ik}}$, luego, es posible incrementar x_k hasta

$$x_k = \frac{\bar{b}_r}{y_{rk}} = \text{Mínimo} \left\{ \frac{\bar{b}_i}{y_{ik}} : y_{ik} > 0 \right\}.$$

Cuando no hay degeneración $\bar{b}_r > 0$ y entonces $x_k = \frac{\bar{b}_r}{y_{rk}} > 0$. Por (2.3.7) $z < z_0$ pues $z_k - c_k > 0$ y la función objetivo mejora estrictamente. Sustituyendo x_k en (2.3.8) obtenemos

$$\begin{aligned} x_{B_i} &= \bar{b}_i - \frac{y_{ik}\bar{b}_r}{y_{rk}} \quad i = 1, \dots, m \\ x_k &= \frac{\bar{b}_r}{y_{rk}}, \quad 2.3.9 \\ x_j &= 0 \quad j \in I - \{k\}. \end{aligned} \quad (10)$$

De (2.3.9) $x_{B_r} = 0$ y se tienen a lo más m variables positivas. Las columnas linealmente independientes correspondientes en A son $a_{B_1}, a_{B_2}, \dots, a_{B_{r-1}}, a_k, a_{B_{r+1}}, \dots, a_{B_m}$. Así, el punto dado por (2.3.9) es una solución básica factible.

Hemos completado una iteración del método simplex: el proceso de transformación de una base a una base adyacente. Lo que se hizo fue incrementar el valor de una variable no básica x_k con $z_k - c_k$ positivo y ajustar las variables básicas actuales. La variable x_{B_r} se hace cero, por tanto la variable x_k entra a la base y x_{B_r} sale de la base. El valor de la función objetivo decrece estrictamente y así las soluciones factibles generadas son distintas; como existe un número finito de éstas, se termina en un número finito de pasos.

Resumiendo, dadas una solución básica factible y una base correspondiente, entonces es posible mejorar la solución si $z_k - c_k > 0$ para alguna variable no básica x_k , o bien, el proceso se detiene en un punto óptimo si $z_j - c_j \leq 0$ para todas las variables no básicas. Si $z_k - c_k > 0$ y el vector y_k contiene por lo menos una componente positiva, entonces el incremento en x_k será bloqueado por una de las variables básicas presentes, que se vuelve cero y sale de la base (x_k entra). Ahora bien, si $z_k - c_k > 0$ y $y_k \leq 0$ de (2.3.8) deducimos que x_k se puede incrementar indefinidamente sin que ninguna de las variables básicas se haga negativa. Por tanto, las soluciones x (en donde $x_B = B^{-1}b - y_k x_k$, $x_k > 0$ y todas las componentes no básicas son cero) son factibles y su valor objetivo es $z = z_0 - (z_k - c_k)x_k$ el cual tiende a $-\infty$ cuando x_k tiende a ∞ .

Lo anterior lo presentamos ahora como un algoritmo que resuelve el siguiente problema:

$$\begin{aligned} & \text{Min } cx \\ & \text{s.a } Ax = b \\ & \quad x \geq 0. \end{aligned}$$

Algoritmo Simplex

Primera etapa

Encontrar una solución básica factible inicial x_B con base B .

Segunda etapa

1. Resolver el sistema $Bx_B = b$, $x_B = B^{-1}b = \bar{b}$. Hacer $x_N = 0$ y $z = c_B x_B$.
2. Resolver $wB = c_B$, $w = c_B B^{-1}$. Para todas las variables no básicas calcular $z_j - c_j = w a_j - c_j$. Sea

$$z_k - c_k = \max_{j \in I} z_j - c_j$$

donde I es el conjunto actual de índices asociados con las variables no básicas. Si $z_k - c_k \leq 0$, entonces el proceso se detiene, la solución básica factible es óptima. En otro caso, continuamos con el paso 3, con x_k como variable de entrada.

3. Resolver el sistema $B y_k = a_k$. Si $y_k \leq 0$, parar, la solución es no acotada. Si $y_k \not\leq 0$ continuar con 4.

4. La variable x_k entra a la base. El índice r de la variable x_{B_r} , que sale de la base, se determina de la siguiente forma

$$\frac{\bar{b}_r}{y_{rk}} = \text{Mínimo} \left\{ \frac{\bar{b}_i}{y_{ik}} : y_{ik} > 0 \right\}$$

Actualizamos B reemplazando a_{B_r} por a_k , actualizamos el conjunto de índices I , y se repite el paso 1.

En cada iteración de la segunda etapa, se realiza una de las siguientes acciones:

1. El proceso se detiene con un punto extremo óptimo si $z_k - c_k > 0$.
2. Si $z_k - c_k > 0$ y $y_k \leq 0$ el proceso termina con una solución no acotada.
3. Se genera una nueva solución básica factible si $z_k - c_k > 0$ y $y_k \not\leq 0$.

Si no hay degeneración, $\bar{b}_r > 0$ y entonces $x_k = \frac{\bar{b}_r}{y_{rk}} > 0$, luego la diferencia entre los valores objetivos de la iteración anterior y los de la iteración presente es $x_k(z_k - c_k) > 0$, es decir, la función objetivo decrece estrictamente en cada iteración y por tanto las soluciones básicas generadas son distintas, como hay un número finito de soluciones básicas factibles entonces el método se detendrá en un número finito de pasos con una solución óptima finita o con una solución óptima no acotada.

El Método de la gran M y la solución básica inicial

Consideremos el siguiente problema

$$\begin{aligned} & \text{Min } cx \\ & \text{s.a } Ax \geq b \\ & \quad x \geq 0. \end{aligned} \tag{2.3.10}$$

Donde A es una matriz de $m \times n$, y b es un m -vector.

Para poder aplicar el método simplex a este problema, podemos abstraer un vector de holgura x_s a la restricción $Ax \geq b$, obteniendo la restricción en forma estándar: $Ax - x_s = b$, $x \geq 0$, $x_s \geq 0$. El nuevo problema tiene la forma

$$\begin{aligned} & \text{Min } c'x' \\ & \text{s.a } A'x' = b \\ & \quad x' \geq 0. \end{aligned} \tag{2.3.11}$$

Donde A' es igual a la matriz $(A, -I)$ de $m \times (n + m)$, $x' = \begin{pmatrix} x \\ x_s \end{pmatrix}$ es un vector con $n + m$ entradas y $c' = (c, \bar{0})$ donde $\bar{0}$ es un m -vector renglón de ceros. Es claro que encontrar una solución óptima x'^* para (2.3.11) significa encontrar una solución óptima x^* para (2.3.10) tomando las primeras n entradas de x'^* como entradas de x^* .

Enfoquémonos ahora a la resolución de (2.3.11); el método simplex puede ser aplicado aquí, para esto, necesitamos tener una solución básica factible inicial.

Podemos suponer sin pérdida de generalidad que $b \geq 0$ (Si alguna entrada b_i de b es menor que cero, entonces el i -ésimo renglón del sistema $A'x' = b$ se puede multiplicar por -1).

Si A' contiene una submatriz identidad, entonces de inmediato se obtiene una solución básica factible tomando simplemente $B = I$, y como $b \geq 0$, se

tiene que $x_B = B^{-1}b = b \geq 0$. En caso contrario recurrimos a las variables artificiales para obtener una solución básica factible inicial.

Introduciendo el vector artificial x_a a las restricciones de (2.3.11) tenemos

$$\begin{aligned} A'x' + x_a &= b \\ x', x_a &\geq 0 \end{aligned}$$

Así, la solución básica factible inicial está dada por $x_a = b$, $x' = 0$ y el método simplex puede ser aplicado.

Sin embargo, debido a que $A'x' = b$ si y sólo si $A'x' + x_a = b$, con $x_a = 0$, es necesario que el vector artificial x_a valga cero al finalizar el método simplex, para lograr esto aplicamos el método de la gran M, que consiste en asignar coeficientes muy grandes a éstas variables en la función objetivo original, a fin de hacer que su presencia en la base sea muy poco atractiva desde el punto de vista de querer minimizar la función objetivo. Entonces tenemos

$$\begin{aligned} \text{Min } c'x' + Mx_a \\ \text{s.a } A'x' + x_a &= b \\ x', x_a &\geq 0. \end{aligned}$$

en donde M es un número positivo muy grande.

La estrategia anterior puede interpretarse como aquella que minimiza x_a con prioridad uno, y entre todas las soluciones óptimas alternativas para este objetivo, la que minimiza el objetivo secundario $c'x'$.

Aunque la solución inicial $x_a = b$, $x' = 0$ es factible para las nuevas restricciones, tiene un valor objetivo poco atractivo Mb , por lo tanto el método saca de la base a las variables artificiales, y después continúa hasta encontrar la solución óptima del problema original.

Capítulo 3

Optimización lineal aplicada al problema de aproximación uniforme.

En este capítulo hacemos el planteamiento del problema de aproximación uniforme como un programa lineal para así aplicar los métodos de solución desarrollados en el capítulo anterior. También estudiamos el comportamiento de los algoritmos mediante unos ejemplos donde aproximamos varias funciones, comprobando la eficacia de los métodos propuestos.

3.1 Optimización Lineal y Aproximación Uniforme.

En el problema de Aproximación de funciones, lo que se desea es aproximar una función f por medio de una combinación lineal de otras funciones v_1, v_2, \dots, v_n fácilmente manejables (v_1, v_2, \dots, v_n son llamadas funciones aproximadoras).

Este problema podemos plantearlo en forma de un programa lineal de la siguiente manera:

Sea T un subconjunto compacto de R^k y sean $v_r : T \rightarrow R$, $r = 1, \dots, n$ y $f : T \rightarrow R$ funciones continuas definidas sobre T .

El problema de aproximación uniforme consiste en determinar una combinación lineal

$$\sum_{r=1}^n y_r v_r$$

que mejor aproxime a f en el sentido de minimizar:

$$\max_{t \in T} \left| \sum_{r=1}^n y_r v_r(t) - f(t) \right|$$

Tenemos entonces el problema de aproximación:

$$\text{Minimizar } \max_{t \in T} \left| \sum_{r=1}^n y_r v_r(t) - f(t) \right| \quad y = (y_1, \dots, y_n)^\top \in R^n.$$

Que también se formula así:

$$\begin{aligned} (PA) \quad & \text{Minimizar } y_{n+1} \\ & \text{sujeto a } \left| \sum_{r=1}^n y_r v_r(t) - f(t) \right| \leq y_{n+1}, \\ & y = (y_1, \dots, y_n)^\top \in R^n, y_{n+1} \in R, t \in T. \end{aligned}$$

La desigualdad que hay en la restricción de (PA) es equivalente a otras dos desigualdades. Por tanto el problema de aproximación (PA) puede reescribirse como un programa lineal (P) de la siguiente manera:

$$\begin{aligned}
(P) \quad & \text{Minimizar } y_{n+1} \\
\text{s.a.} \quad & \sum_{r=1}^n v_r(t)y_r + y_{n+1} \geq f(t), \quad 3.1.1 \\
& -\sum_{r=1}^n v_r(t)y_r + y_{n+1} \geq -f(t), \\
& y = (y_1, \dots, y_n)^\top \in R^n, y_{n+1} \in R, t \in T.
\end{aligned} \tag{11}$$

Definición 23. Un *PLSI* de la forma (3.1.1) se dice que es *continuo* cuando las funciones v_1, \dots, v_n y f son continuas.

Hagamos $\rho = \max_{t \in T} |f(t)|$ (ρ existe porque la función f es continua y T es compacto).

Nota. El *PLSI* (3.1.1) es *consistente* (El conjunto factible tiene al menos un elemento, $(y', y'_{n+1}), y' = 0 \in R^n, y'_{n+1} = \rho = \max_{t \in T} |f(t)|$) y *continuo*.

Antes de abordar los métodos para la resolución del programa (3.1.1), establecemos un teorema que afirma la existencia de una solución de éste; en realidad este teorema se cumple debido al teorema 1.1 (sección 1.2), sin embargo presentamos una demostración la cual utiliza conceptos de la programación lineal como el de punto factible.

Teorema 3.1 Sea T subconjunto compacto no vacío de R^k y sean $v_r : T \rightarrow R, r = 1, \dots, n$ y $f : T \rightarrow R$ funciones continuas definidas sobre T , supongamos también que las funciones $v_r, r = 1, \dots, n$ son linealmente independientes. Entonces existe una solución óptima del problema de aproximación (3.1.1).

Demostración. Definamos una norma en R^n por

$$\|y\|_v = \max_{t \in T} \left| \sum_{r=1}^n v_r(t)y_r \right|.$$

Veamos que en efecto $\|\cdot\|_v$ verifica las propiedades (1.2.1) de la definición de norma:

Dados $x, y \in R^n, \lambda \in R^+$.

$$(i) \|y\|_v = \max_{t \in T} \left| \sum_{r=1}^n v_r(t)y_r \right| \geq 0. \text{ Si } \|y\|_v = 0, \text{ entonces } \max_{t \in T} \left| \sum_{r=1}^n v_r(t)y_r \right| = 0,$$

esto se cumple sólo si $\sum_{r=1}^n v_r(t)y_r = 0$, para toda $t \in T$, pero por ser las funciones $v_r, r = 1, \dots, n$ linealmente independientes debemos tener que $y = 0$.

$$(ii) \|\lambda y\|_v = \max_{t \in T} \left| \sum_{r=1}^n v_r(t)\lambda y_r \right| = |\lambda| \max_{t \in T} \left| \sum_{r=1}^n v_r(t)y_r \right| = |\lambda| \|y\|_v.$$

$$\begin{aligned}
\text{(iii) } \|x + y\|_v &= \max_{t \in T} \left| \sum_{r=1}^n v_r(t)(x_r + y_r) \right| = \max_{t \in T} \left| \sum_{r=1}^n v_r(t)x_r + \sum_{r=1}^n v_r(t)y_r \right| \leq \\
&\leq \max_{t \in T} \left(\left| \sum_{r=1}^n v_r(t)x_r \right| + \left| \sum_{r=1}^n v_r(t)y_r \right| \right) = \max_{t \in T} \left| \sum_{r=1}^n v_r(t)x_r \right| + \\
&\quad + \max_{t \in T} \left| \sum_{r=1}^n v_r(t)y_r \right| = \|x\|_v + \|y\|_v.
\end{aligned}$$

Por tanto, $\|\cdot\|_v$ es una norma.

Tomemos $\rho = \max_{t \in T} |f(t)|$. Cualquier punto factible (y, y_{n+1}) de (3.1.1), $y \in R^n$, $y_{n+1} \in R$ donde $y_{n+1} > \rho$ puede ser mejorado por el punto factible $(0, y'_{n+1})$, $0 \in R^n$, $y'_{n+1} = \rho$ por lo tanto podemos agregar la restricción $0 \leq y_{n+1} \leq \rho$ en (3.1.1) y obtener como resultado un problema equivalente. Ahora, sea (y, y_{n+1}) un punto factible de este nuevo problema, tenemos para algún $t' \in T$

$$\begin{aligned}
\|y\|_v &= \max_{t \in T} \left| \sum_{r=1}^n v_r(t)y_r \right| = \left| \sum_{r=1}^n v_r(t')y_r \right| = \\
&= \left| f(t') - \sum_{r=1}^n v_r(t')y_r + f(t') \right| \leq \left| f(t') - \sum_{r=1}^n v_r(t')y_r \right| + |f(t')| \leq \\
&\leq y_{n+1} + |f(t')| \leq 2\rho.
\end{aligned}$$

Esto significa que el conjunto factible sobre el que se necesita minimizar es un subconjunto compacto de R^{n+1} y por lo tanto el mínimo del problema es alcanzado. ■

Cuando en el Programa Lineal (3.1.1) el conjunto de índices $T \subset R^k$ es un conjunto finito X_m de puntos se tiene un Programa Lineal con un número finito de restricciones.

El programa lineal a resolver queda de la siguiente forma:

$$\begin{aligned}
&\text{Minimizar } y_{n+1} \\
&s.a \sum_{r=1}^n v_r(x_j)y_r + y_{n+1} \geq f(x_j), \quad j = 1, \dots, m \quad 3.1.2 \\
&\quad - \sum_{r=1}^n v_r(x_j)y_r + y_{n+1} \geq -f(x_j), \quad j = 1, \dots, m. \\
&\quad y = (y_1, \dots, y_n)^\top \in R^n, \quad y_{n+1} \in R, \quad x_i \in X_m, \quad i = 1, \dots, m.
\end{aligned}$$

Donde m es el número de elementos de X_m .

Aplicamos entonces el método simplex desarrollado en el capítulo II para resolver el programa lineal (3.1.2).

Consideremos la representación matricial de este programa lineal.

$$\begin{aligned}
&\text{Min } cy \\
&s.a Ay \geq b
\end{aligned} \quad (3.1.3)$$

Donde

$$A = \begin{bmatrix} v_1(x_1) & \cdots & v_n(x_1) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ v_1(x_m) & \cdots & v_n(x_m) & 1 \\ -v_1(x_1) & \cdots & -v_n(x_1) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ -v_1(x_m) & \cdots & -v_n(x_m) & 1 \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n+1} \end{bmatrix}, b = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_m) \\ -f(x_1) \\ \vdots \\ -f(x_m) \end{bmatrix},$$

$$c = (0, 0, \dots, 1).$$

Observemos que en (3.1.3) el vector y no está restringido en su signo, para eliminar este inconveniente hacemos la transformación $y = y' - y''$ en donde $y' \geq 0$ y $y'' \geq 0$. A continuación en el figura 3.1 mostramos un diagrama de flujo que describe todo el proceso de la resolución de (3.1.2) mediante el método simplex.

itbpFU371.875pt383.1875pt0ptFigura 3.1. Diagrama de flujo que describe el proceso para resolver (3.1.2).Figure

Cuando en el Programa Lineal (P) el conjunto de índices T es infinito, se tiene un programa lineal semi-infinito; para resolver éste programa utilizamos el Algoritmo 2.2.2.

Si T es un intervalo cerrado $[x_1, x_2]$ el algoritmo toma la forma siguiente:

Algoritmo 3.1. Dado el intervalo $T = [x_1, x_2]$ y dada una tolerancia $\varepsilon > 0$, iniciamos un índice $i = n$, donde n es el número de elementos de la primera discretización de T . Construimos la primera discretización T_i .

Primera Etapa Dado el subproblema:

$$(P_i) \quad \text{Minimizar } y_{n+1}$$

$$\text{s.a. } \sum_{r=1}^n v_r(t)y_r + y_{n+1} \geq f(t), \quad t \in T_i$$

$$-\sum_{r=1}^n v_r(t)y_r + y_{n+1} \geq -f(t), \quad t \in T_i.$$

Calcular una solución óptima de (P_i) , y^i .

Segunda etapa Verificar la factibilidad de y^i para (P) .

Para esto, encontramos s_i , el menor de los mínimos de las funciones:

$$\sum_{r=1}^n v_r(t)y_r^i + y_{n+1}^i - f(t), \quad t \in T$$

$$-\sum_{r=1}^n v_r(t)y_r^i + y_{n+1}^i + f(t), \quad t \in T.$$

Si $s_i \geq -\varepsilon$ parar (y^i es aceptada como solución óptima factible de (P)).

Si no, incrementar el índice i y regresar a la primera etapa.

La terminación de este algoritmo en un número finito de iteraciones está garantizada por el teorema 3.2.

Sabemos de la existencia de la solución para el problema (P) , es decir del mejor aproximante $p^* = \sum_{r=1}^n v_r(t)y_r^*$ para f sobre T con error y_{n+1}^* . En la primera etapa del algoritmo 3.1 lo que obtenemos explícitamente es el mejor aproximante $p^i = \sum_{r=1}^n v_r(t)y_r^i$ para f sobre el subconjunto discretizado T_i . El teorema 1.15 garantiza que p^i puede estar tan cerca como queramos de p^* haciendo $|T_i|$ suficientemente pequeño; luego, como p^* está contenido en una banda para f de ancho $2y_{n+1}^*$, si seleccionamos una tolerancia ε pequeña y

hacemos tender $|T_i|$ a cero el aproximante p^i puede estar contenido en una banda para f de ancho $2(y_{n+1}^i + \varepsilon)$; esto es lo que se verifica en la segunda etapa, dependiendo de la tolerancia ε escogida el máximo error y_{n+1}^i cometido por el aproximante p^i no debe exceder a $y_{n+1}^i + \varepsilon$ sobre todo el intervalo T . La figura 3.2 muestra geoméricamente lo anterior.

itbpFU335.8125pt157.125pt0ptFigura 3.2Figure

Teorema 3.2 Sea $T = [x_1, x_2]$, sean f y v_r , $r = 1, \dots, n$ como en el teorema 3.1, si $\{T_i\}$ es una sucesión expansiva de redes en T , el algoritmo 3.1 tiene una terminación finita.

Demostración. Por el teorema 3.1 sabemos que existe una solución óptima para el problema (P) , denotémosla por $y^* = (y_1^*, \dots, y_n^*)^\top \in R^n$, $y_{n+1}^* \in R$. El algoritmo iterativamente calcula una sucesión $\{y^i\}$ de soluciones óptimas para (P_i) .

Veamos como es la densidad de $T_i = \left\{x_1 + \frac{(x_2 - x_1)j}{i-1}, j = 0, \dots, i-1\right\}$, $|T_i| = \max_{x \in [x_1, x_2]} \inf_{y \in T_i} |x - y| < \frac{x_2 - x_1}{i-1}$, luego cuando $i \rightarrow \infty$, $|T_i| \rightarrow 0$. Por el teorema 1.13,

$$\lim_{i \rightarrow \infty} y_{n+1}^i = y_{n+1}^*, \quad (3.1.4)$$

pues $\|f - P_Y\| \rightarrow \|f - P_X\|$ (aquí $Y = T_i$, $X = [x_1, x_2]$ y $P = \sum_{r=1}^n y_r^i v_r(t)$). Por el teorema 1.15 se cumple también que

$$\lim_{i \rightarrow \infty} \|P_Y - P_X\| = 0. \quad (3.1.5)$$

Tomemos un t arbitrario en T , si $g_1(y) = \sum_{r=1}^n v_r(t)y_r + y_{n+1} - f(t)$ entonces

$$\begin{aligned} |g_1(y^i) - g_1(y^*)| &= \left| \sum_{r=1}^n v_r(t)y_r^i + y_{n+1}^i - f(t) - \sum_{r=1}^n v_r(t)y_r^* - y_{n+1}^* + f(t) \right| \\ &= \left| \sum_{r=1}^n v_r(t)(y_r^i - y_r^*) + y_{n+1}^i - y_{n+1}^* \right| \\ &\leq \left| \sum_{r=1}^n v_r(t)(y_r^i - y_r^*) \right| + |y_{n+1}^i - y_{n+1}^*| \\ &\leq \|P_Y - P_X\| + |y_{n+1}^i - y_{n+1}^*|. \end{aligned}$$

Por (3.1.4) y (3.1.5) para todo $\varepsilon > 0$, existe un número natural i suficientemente grande tal que $\|P_Y - P_X\| + |y_{n+1}^i - y_{n+1}^*| \leq \varepsilon$. Por lo tanto dado $\varepsilon > 0$, existe un número natural i tal que

$$|g_1(y^i) - g_1(y^*)| \leq \varepsilon,$$

es decir,

$$-\varepsilon \leq g_1(y^i) - g_1(y^*) \leq \varepsilon,$$

como $g_1(y^*) \geq 0$,

$$-\varepsilon \leq -\varepsilon + g_1(y^*) \leq g_1(y^i) \leq \varepsilon + g_1(y^*),$$

tenemos entonces que

$$g_1(y^i) \geq -\varepsilon \text{ para todo } t \in T;$$

esto se cumple también tomando $g_2(y) = -\sum_{r=1}^n v_r(t)y_r + y_{n+1} + f(t)$, lo cual significa que la condición de paro $s_i \geq -\varepsilon$ del algoritmo se cumple para algún i suficientemente grande, por tanto el algoritmo tiene una terminación finita. ■

3.2 Resultados Numéricos

Para ilustrar el funcionamiento de los algoritmos desarrollados, presentamos varios ejemplos donde se aproxima una función dada, señalemos previamente que el conjunto de funciones aproximadoras que utilizamos en los ejemplos 1 a 3 y 7 es el de los polinomios de grado menor o igual que n , en todos los ejemplos utilizamos el método de la programación lineal ordinaria y en el ejemplo 7 aplicamos también el de la programación lineal semi-infinita. Indicamos el error máximo cometido, el conjunto discretizado sobre el que se aproxima y mostramos la gráfica de la función y del mejor aproximante obtenido.

Ejemplo 1.

En este ejemplo aproximamos la función $f : [-1, 1] \rightarrow R$, definida por $f(x) = \text{sen}(10x)$ (medido en radianes), hacemos $n = 5$, tomamos al conjunto $\left\{-1 + \frac{2j}{20-1} : j = 0, \dots, 20 - 1\right\}$ como discretización de intervalo $[-1, 1]$. El mejor aproximante de grado menor o igual a $n = 5$, obtenido es: $p_5(x) = 0.1198x - 0.9632x^3 + 1.2840x^5$. El error máximo es de 0.9847 sobre el conjunto discretizado y sobre el intervalo es 0.9848.

itbpFU327.25pt129.25pt0.75ptFigura 3.3. Función f y su mejor aproximante p ($n=5$).Figure

La aproximación no resulta muy buena, sin embargo podemos mejorarla si tomamos $n = 15$, es decir, hacemos crecer el espacio de funciones aproximadoras a los polinomios de grado menor o igual a 15.

itbpFU326.5pt129.25pt0ptFigura 3.4. Función f y su mejor aproximante p ($n=15$).Figure

Así obtenemos $p_{15}(x) = 10x - 166.5x^3 + 829.3x^5 - 1949.5x^7 + 2596.9x^9 - 2090.1x^{11} + 971x^{13} - 201.8x^{15}$ (Redondeando a décimos). El error máximo es de 0.0001 sobre el conjunto discretizado y sobre el intervalo es 0.015.

Ejemplo 2.

Aproximamos la función $f : [-1, 1] \rightarrow R$, definida por $f(x) = \frac{\cos(3\pi(x+4))}{\pi(x+4)}$, hacemos $n = 5$, tomamos al conjunto $\left\{-1 + \frac{2j}{30-1} : j = 0, \dots, 30 - 1\right\}$ como discretización del intervalo $[-1, 1]$. Obtenemos un aproximante de grado menor o igual a 5, con *error máximo* = 0.09 sobre el intervalo $[-1, 1]$.

itbpFU332.75pt158.75pt0ptFigura 3.5. Función f y su mejor aproximante p ($n=5$).Figure

Tomando $n = 7$ el error resulta 0.065. Con $n = 15$ obtenemos un aproximante con error máximo igual a 0.00006 sobre el conjunto discretizado y sobre el intervalo 0.00065. Éste último es el que graficamos a continuación. Notamos con este ejemplo que las aproximaciones mejoran cuando tomamos n mayor al número de mínimos y máximos locales de la función a aproximar.

itbpFU332.5625pt158.5pt0ptFigura 3.6. Función f y su mejor aproximante p ($n=15$).Figure

Ejemplo 3.

La función a aproximar es $f : [-1, 1] \rightarrow [\frac{1}{2}, 1]$, definida por $f(x) = \frac{1}{1+x^2}$, con $n = 4$, y el conjunto $\{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$ como discretización del intervalo $[-1, 1]$. El aproximante obtenido es $p(x) = 0.9936 - 0.8367x^2 + 0.349x^4$. El error máximo sobre la discretización es 0.0063 y sobre $[-1, 1]$ es 0.009.

itbpFU340.6875pt164.625pt0ptFigura 3.7. Función f y su mejor aproximante p ($n=4$).Figure

Ejemplo 4.

En este ejemplo aproximamos una función de dos variables $f : U \subset R^2 \rightarrow R$ definida por $f(x, y) = e^{(-x^2-y^2)}$ donde $U = \{(x, y) : -1 \leq x, y \leq 1\}$. Aquí el conjunto de funciones aproximadoras lo definimos como $\{1, x, y, xy, x^2, y^2\}$. Para la discretización tomamos el conjunto de puntos $\{(a, a') : a, a' \in A\}$, donde $A = \{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$. El aproximante obtenido es el polinomio en dos variables $p(x, y) = 0.8975 - 0.4323x^2 - 0.4323y^2$. *Error máximo* = 0.102 sobre U .

itbpFU324.9375pt175.125pt0ptFigura 3.8. Función f y su mejor aproximante p .Figure

Ejemplo 5.

Aproximamos la función de dos variables $f : U \subset R^2 \rightarrow R$ definida por $f(x, y) = \frac{\frac{1}{2}(x^2+y^2)}{1 + \sqrt[2]{1 - \frac{1}{4}(x^2+y^2)}}$ donde $U = \{(x, y) : -1 \leq x, y \leq 1\}$. Aquí el conjunto de funciones aproximadoras lo definimos también como $\{1, x, y, xy, x^2, y^2\}$. Para la discretización tomamos el conjunto $A \times A = \{(a, a') : a, a' \in A\}$, donde $A = \{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$. El aproximante obtenido es el polinomio en dos variables $p(x, y) = -0.0125 + 0.2928x^2 + 0.2928y^2$. *Error máximo* = 0.0125 sobre U .

itbpFU313.6875pt144.3125pt0.75ptFigura 3.9. Función f y su mejor aproximante p .Figure

Ejemplo 6.

Aproximamos la función de dos variables $f : U \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ definida por $f(x, y) = \text{sen}(((x - 2.5)^2 + (y - 2.5)^2)/2)$ donde $U = \{(x, y) : 0 \leq x, y \leq 5\}$. Aquí el conjunto de funciones aproximadoras es el de los polinomios en dos variables de grado menor o igual a 4. Para la discretización tomamos el conjunto $A \times A = \{(a, a') : a, a' \in A\}$, donde $A = \{0, 1, 2, 3, 4, 5\}$.

itbpFU335.8125pt233.625pt0ptFigura 3.10. Función f y su mejor aproximante p .Figure

Para una aproximación del error tomamos el conjunto $C = \{(b, b') : b, b' \in B\}$, donde $B = \left\{ \frac{5j}{20-1} : j = 0, \dots, 20 - 1 \right\}$. C consta de 400 puntos distribuidos uniformemente en U . La media del error sobre esos puntos resultó 0.07505439713114.

Nota: En el artículo *Interpolación Polinómica en dos variables con Mathematica* publicado en internet [15] ésta función es aproximada interpolando con polinomios de Lagrange en dos variables de grado menor o igual a 5, utilizan los mismos puntos del conjunto $A \times A$, el error que obtienen coincide con el nuestro hasta las millonésimas.

Ejemplo 7.

En los ejemplos anteriores dada una función $f : [a, b] \rightarrow R$, la búsqueda de su mejor aproximante fué realizada sobre una discretización de $[a, b]$, razón por la cual fueron resueltos mediante el método de la programación lineal ordinaria. Si deseamos realizar la búsqueda sobre el intervalo original $[a, b]$ debemos aplicar el método de la programación lineal semi-infinita.

Es claro que realizar la aproximación sobre $[a, b]$ debe arrojar un mejor resultado que sobre una discretización T de $[a, b]$, pues dado un mejor aproximante P para f sobre T , puede existir un punto $x \in [a, b]$ no contenido en T tal que $\|f - P\|_T = \max_{t \in T} |f(t) - P(t)| < f(x) - P(x)$. El siguiente ejemplo nos muestra cómo la búsqueda del mejor aproximante sobre el intervalo original mejora la aproximación a la función.

a) Sobre una discretización.

Deseamos aproximar la función $f : [-1, 1] \rightarrow R$, definida por

$$f(x) = \frac{10}{1 + 50(x - 0.47)^2},$$

hacemos $n = 25$, tomamos al conjunto $\left\{-1 + \frac{2j}{51-1} : j = 0, \dots, 51 - 1\right\}$ como discretización del intervalo $[-1, 1]$. Obtenemos un aproximante de grado menor o igual a $n = 25$, con *error máximo* = 2.6.

itbpFU343.0625pt228.625pt0ptFigura 3.11. Función f y su mejor aproximante p
(Programación Lineal Ordinaria).Figure

Una aproximación no muy buena para puntos cercanos a -1 y 1.

b) Sobre el intervalo $[-1, 1]$.

Hacemos ahora la aproximación utilizando una sucesión expansiva de redes de acuerdo a (2.2.2.3) con constante $k = 2$, $r_1 = 51$, $\varepsilon = 0.001$, y aplicamos el algoritmo 3.1.

itbpFU343.0625pt172.125pt0ptFig. 3.12. Función f y su mejor aproximante p
(Programación Lineal Semi-infinita).Figure

Obtenemos así, después de 3 iteraciones un aproximante de grado menor o igual a $n = 25$, con *error máximo* = 0.18.

Capítulo 4

Una aplicación a la Óptica

En este capítulo mostramos una aplicación de la Aproximación de Funciones resuelta con programación lineal; lo que aproximamos son los parámetros geométricos de superficies cónicas (de gran importancia en los sistemas ópticos). Primeramente mencionamos cual es la representación matemática de estas superficies, después describimos los requerimientos especiales para poder aplicar los métodos desarrollados en el capítulo anterior y finalmente se presentan los resultados obtenidos al evaluar algunas superficies.

4.1 Introducción

En Óptica es de gran importancia desarrollar métodos de prueba para verificar el buen funcionamiento de los instrumentos, o bien para verificar por ejemplo, que las superficies fabricadas en serie mantienen un patrón de parámetros preestablecidos.

Existen métodos ópticos que utilizan instrumentos de medición especiales los cuales obtienen experimentalmente las coordenadas de cierto número de puntos en la superficie bajo prueba. Lo que se desea obtener es la forma analítica de la superficie que mejor se ajuste a ellos en el sentido de la norma uniforme.

Los datos obtenidos experimentalmente pueden ser vistos como evaluaciones de una función, la cual se puede aproximar por medio de una combinación lineal de otras funciones, éstas últimas deben ser elegidas apropiadamente de tal manera que la función resulte aproximada por una superficie cónica.

4.2 Superficies cónicas

Las secciones cónicas surgen de la intersección de un plano con un cono circular recto, variando la posición del plano se pueden obtener las diversas cónicas. Al rotar una sección cónica alrededor de un eje de simetría, se obtiene una superficie cónica. Dichas superficies tienen diversas representaciones matemáticas, nosotros tomaremos la representación más empleada en óptica [10]:

$$Z = \frac{cS^2}{1 + \sqrt{1 - (K + 1)c^2S^2}}, \quad (3.2.1)$$

donde Z es llamada la sagita, $S^2 = x^2 + y^2$, $c = \frac{1}{r}$, donde r es el radio de curvatura y K es la constante de conicidad definida como $K = -e^2$, siendo e la excentricidad.

4.3 Cálculo de parámetros

Para resolver el problema de la obtención de la forma analítica de una superficie cónica que mejor se ajusta a un conjunto de puntos aplicamos la aproximación de funciones de la siguiente forma:

Dado el conjunto de datos experimentales (x_i, y_i, z_i) , $i = 1, \dots, m$, consideramos al conjunto de valores z_i como evaluaciones de la función $f : R^3 \rightarrow R$, definida por $f(x, y, z) = z$. Esta función es aproximada por medio de una combinación lineal de dos funciones $v_1 : R^3 \rightarrow R$ y $v_2 : R^3 \rightarrow R$, definidas por

$$v_1(x, y, z) = x^2 + y^2, \quad v_2(x, y, z) = z^2.$$

Tenemos entonces el problema de aproximación:

$$\text{Minimizar } \max_{t \in T} \left| \sum_{r=1}^2 y_r v_r(t) - f(t) \right| \quad y = (y_1, y_2) \in R^2.$$

Formulando este problema como el programa lineal (2.2.1)

$$\begin{aligned} & \text{Minimizar} \quad y_{n+1} \\ & \text{s.a.} \sum_{r=1}^2 v_r(t) y_r + y_{n+1} \geq f(t), \\ & \quad - \sum_{r=1}^2 v_r(t) y_r + y_{n+1} \geq -f(t), \\ & \quad y \in R^2, \quad y_{n+1} \in R, \quad t \in T. \end{aligned} \tag{3.3.1}$$

Donde T es el conjunto de datos $(x_i, y_i, z_i) \in R^3$ y $f(x, y, z) = z$.

Al resolver (3.3.1) obtenemos la combinación lineal de v_1 y v_2 que mejor ajusta a f , es decir

$$z = f(x, y, z) = y_1(x^2 + y^2) + y_2 z^2. \tag{3.3.2}$$

Ahora, despejando z de (3.3.2) tenemos

$$z = \frac{1 \pm \sqrt[2]{1 - 4y_2 y_1 (x^2 + y^2)}}{2y_2}, \tag{3.3.3}$$

o bien,

$$z = \frac{2y_1(x^2 + y^2)}{1 + \sqrt[2]{1 - 4y_2 y_1 (x^2 + y^2)}}, \tag{3.3.4}$$

lo anterior se obtiene tomando la rama negativa de la solución (3.3.3) y multiplicando por su conjugado $\frac{1 + \sqrt{1 - 4y_2y_1(x^2 + y^2)}}{1 + \sqrt{1 - 4y_2y_1(x^2 + y^2)}}$.

La ecuación (3.3.4) tiene la forma de la superficie cónica (3.2.1) donde z es la sagita, $S^2 = x^2 + y^2$, $c = 2y_1$, y $K = \frac{y_2}{y_1} - 1$.

Un diagrama de flujo del algoritmo utilizado es mostrado en la figura 4.1.

itbpFU318.375pt410.25pt0ptFigura 4.1Figure

4.4 Ejemplos

Para ilustrar el funcionamiento del algoritmo desarrollado, presentamos algunos ejemplos: en los primeros cuatro se obtuvieron las coordenadas de algunos puntos de manera analítica, tomando superficies cónicas conocidas (es decir, con K y c conocidos), pero en los ejemplos 3 y 4 se introducen errores en las coordenadas de manera aleatoria. Las superficies que se consideraron tienen un radio de curvatura $r = 200$ cm., solamente se cambió la constante de conicidad y el número de puntos. Finalmente en el ejemplo 5, se midieron las coordenadas de 49 puntos en una superficie real, el espejo secundario del Gran Telescopio Milimétrico (GTM) construido en el INAOE [14].

Ejemplo 1.

Se generaron las coordenadas de 100 puntos distribuidos homogéneamente sobre la superficie de un elipsoide, con constante de conicidad $K = 1$, y radio de curvatura $r = 200$. El programa implementado obtuvo los siguientes parámetros geométricos para la superficie que mejor se ajustó en el proceso de optimización: $K = 1.000007$ y $r = 200.0003$ cm.; con una desviación (error máximo de ajuste) de 0.00003188 cm. La figura 4.2 muestra un corte de la superficie de ajuste junto con los puntos originales.

itbpFU272.3125pt165.5pt0ptFigura 4.2. Ajuste obtenido para puntos medidos en un elipsoide $K = 1$.Figure

Ejemplo 2.

Se generaron las coordenadas de 50 puntos distribuidos homogéneamente sobre la superficie de un paraboloide, con constante de conicidad $K = -1$, y radio de curvatura $r = 200$. El programa implementado obtuvo los siguientes parámetros geométricos para la superficie que mejor se ajustó en el proceso de optimización: $K = -0.9999990$ y $r = 200.00004$, con una desviación (error máximo de ajuste) de 0.00000498. La figura 4.3 muestra un corte de la superficie de ajuste junto con los puntos originales.

itbpFU273.8125pt165.5pt0ptFigura 4.3. Ajuste obtenido para puntos medidos en un paraboloide $K = -1$.Figure

Ejemplo 3.

Se generaron las coordenadas de 50 puntos distribuidos homogéneamente sobre la superficie de un paraboloide, con constante de conicidad $K = -1$, y radio de curvatura $r = 200$, pero se introdujo ruido aleatorio del 10% como máximo en todas las coordenadas (X , Y). El programa implementado obtuvo los siguientes parámetros geométricos para la superficie que mejor se ajustó en el proceso de optimización: $K = -1.09312179$ y $r = 199.848025$, con una desviación (error máximo de ajuste) de 2.8182250. La figura 4.4 muestra un corte de la superficie de ajuste junto con los puntos originales.

itbpFU270.3125pt166.125pt0ptFigura 4.4. Ajuste obtenido para puntos medidos en un paraboloides $K = -1$.Figure

Ejemplo 4.

Se generaron las coordenadas de 100 puntos distribuidos homogéneamente sobre la superficie de un elipsoide, con constante de conicidad $K = 3$, y radio de curvatura $r = 200$, pero se introdujo ruido aleatorio del 10% como máximo en todas las coordenadas (X, Y) . El programa obtuvo como parámetros geométricos $K = -2.8019045626$ y $r = 203.69327196$, con una desviación (error máximo de ajuste) de 0.42730835. La figura 4.5 muestra un corte de la superficie de ajuste junto con los puntos originales.

itbpFU279.375pt165.8125pt0ptFigura 4.5. Ajuste obtenido con $K = 3$.Figure

Ejemplo 5.

En la construcción del Gran Telescopio Milimétrico (GTM) en el INAOE [14], surge la necesidad de probar su espejo secundario, que es una superficie cónica convexa. Para probar este tipo de superficies, el INAOE ha construido una máquina XYZ [6], la cual, cuenta con un sistema de referencia guiado por láser con el se midieron experimentalmente las coordenadas de 49 puntos distribuidos homogéneamente sobre esta superficie real. El programa implementado obtuvo los siguientes parámetros geométricos: $K = 3.247771$ y $r = 144.801855$ con un error máximo de 0.623220. La figura 4.6 muestra la superficie de ajuste junto con los puntos originales y la figura 4.7 muestra un corte de éstos.

itbpFU346.375pt162.75pt0ptFigura 4.6. Ajuste obtenido para puntos medidos en una superficie real.Figure

itbpFU350.875pt171.25pt0ptFigura 4.7. Ajuste obtenido para puntos medidos en una superficie real.Figure

CONCLUSIONES

El objetivo primordial de la tesis fue alcanzado, se realizó la aproximación de funciones de una y dos variables mediante métodos sustentados teóricamente tanto en la programación lineal como en la teoría de aproximación. Conocimos la teoría de esta última y algunos tópicos de la Programación Semi-infinita.

Dependiendo de la función a aproximar, cuando la aproximación se hace con polinomios algebraicos a veces es necesario hacer varias pruebas con diferentes valores del grado de los polinomios para encontrar un buen aproximante. En el caso de que se utilice una combinación lineal de otras funciones aproximadoras la aproximación depende en gran parte de la elección de estas funciones, los conocimientos, la experiencia y los ensayos ayudan a elegir funciones apropiadas para la combinación lineal.

En el desarrollo de los capítulos se completaron los detalles de las demostraciones que dan los textos; cabe señalar que existe una demostración para el teorema 3.2 en [8] pero aquí se dio una demostración original y contextualizada para el caso específico del algoritmo 3.1.

Cuando deseamos aproximar un conjunto de datos por una combinación de funciones aproximadoras o por polinomios de cierto grado, el método a aplicar es el de la programación lineal ordinaria pues en este caso el número de restricciones es finito, sin embargo, este método puede arrojar malos resultados cuando queremos aproximar una función definida sobre un conjunto compacto T con infinitos elementos (por ejemplo un intervalo cerrado). Como se vió en el ejemplo 7 de la sección 3.2, el algoritmo de programación semi-infinita resuelve mejor este problema pues en su segunda etapa verifica si no hay puntos en T para los cuales la diferencia entre la función y su aproximante es muy grande todavía respecto al mejor valor posible, en caso contrario realiza una nueva aproximación sobre un conjunto discretizado T_i con $|T_i|$ más pequeña hasta que la diferencia sea mínima para todo elemento de T .

En la aplicación a una superficie Óptica realizada en el capítulo 4, el algoritmo implementado permitió conocer la forma analítica de la cónica que mejor se ajusta a los datos proporcionados. Esto se logró resolviendo por el método de la programación lineal ordinaria un problema de aproximación polinomial que modela adecuadamente el problema de ajuste.

El programa implementado se diseñó especialmente para ajustar superficies cónicas a puntos medidos en superficies ópticas con estas formas. Si la distribución espacial de los datos no se acerca a un patrón de este tipo, deberán probarse otras funciones v_i en (3.1.1), diferentes a las empleadas, para encontrar una representación propicia.

Esta metodología de encontrar la forma de superficies a partir de conocer las coordenadas de algunos puntos medidos sobre ella, se puede generalizar a cualquier tipo de superficies siempre y cuando se pueda modelar como un problema de aproximación polinomial, aplicando la norma uniforme.

Escribimos un artículo conjuntamente con el Dr. Agustín Santiago y el M.C. Cuauhtémoc H. Castañeda con los resultados obtenidos en la aplicación del capítulo 4. Este artículo fue enviado a la Revista Mexicana de Física y se encuentra sujeto a arbitraje para su publicación.

Como trabajo futuro se pretende generalizar el método propuesto a superficies esféricas de revolución [10].

Bibliografía

- [1] Bazaraa M. S., Jarvis J., y Sherali H. D., *Programación Lineal y Flujo en Redes*, Edit. Limusa, 1999.
- [2] Burden R. L., y Douglas F., *Análisis Numérico*, International Thomson Editores, 1998.
- [3] Cheney E. W., *Introduction to Approximation Theory*, Mc Graw-Hill Book Company, 1966.
- [4] Davis P. J., *Interpolation and Approximation*, Dover Publications Inc., 1975.
- [5] Friedberg H., Insel A., y Spence E., *Algebra Lineal*, Prentice Hall, 1998.
- [6] Gale D. M., *Marco de Metrología para una Máquina de Medición por Coordenadas*, Academia Colombiana de Ciencias Exactas, Física y Naturales, OPTILAS'98.
- [7] Glashoff K., and Gustafson S. A., *Linear Optimization and Approximation: An Introduction to the Theoretical Analysis and Numerical Treatment of Semi-infinite Programs*, Berlin: Springer-Verlag, 1983.
- [8] Goberna M. A., and López M. A., *Linear Semi-infinite Optimization*, John Wiley & Sons, 1998.
- [9] Luenberger D. G., *Programación Lineal y no Lineal*, Addison-Wesley Iberoamericana, 1989.
- [10] Santiago A. A., *Obtención de la Forma de Superficies convexas esféricas con razón focal menor a 1*, Tesis doctoral, INAOE-México, 2001.
- [11] Mortenson Michael E., *Geometric Modeling*, John Wiley & Sons, Inc., 1998.
- [12] Nakamura S., *Análisis Numérico y Visualización Gráfica con Matlab*, Prentice Hall, 1997.
- [13] Rivlin T. J., *An Introduction to the Approximation of Functions*, Dover Publications Inc., 1989.
- [14] <http://www.lmtgtm.org/>.
- [15] <http://csimbolico.rediris.es/math97/resum/a63.htm>.