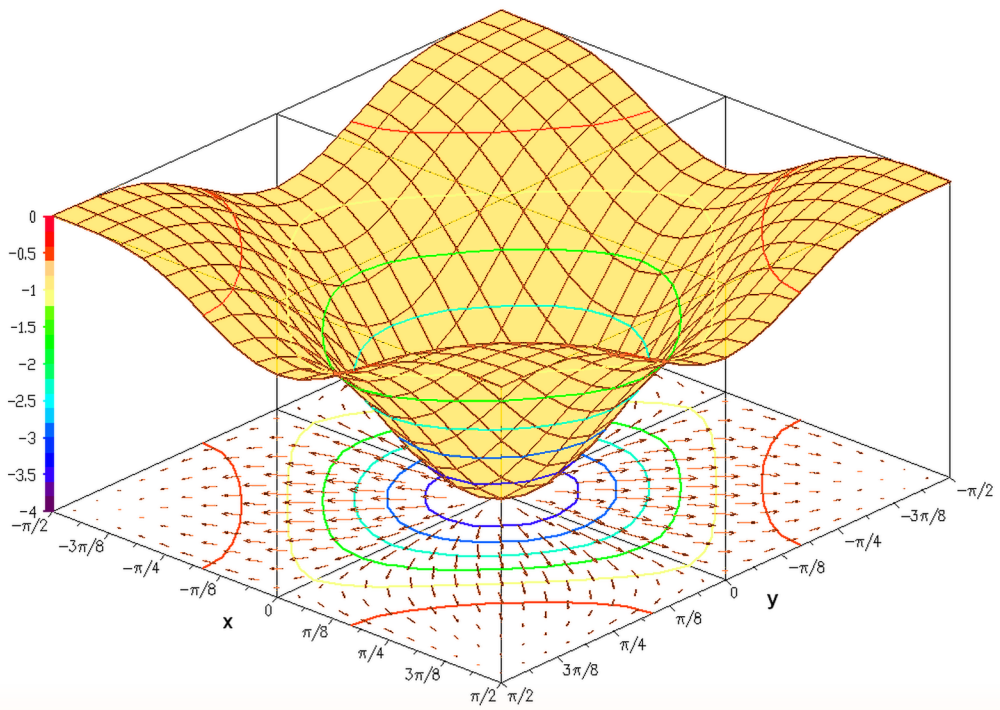
Aplicaciones de Programación no Lineal

Gilberto Espinosa-Paredes y Alejandro Vázquez Rodríguez



OmniaScience
Scholar

Aplicaciones de Programación no Lineal

Gilberto Espinosa-Paredes
Alejandro Vázquez Rodríguez

OmniaScience

Aplicaciones de programación no lineal

Autores:
Gilberto Espinosa-Paredes, Alejandro Vázquez Rodríguez
Profesor-Investigador. Área de Ingeniería en Recursos Energéticos. División de Ciencias Básicas e Ingeniería. Universidad Autónoma Metropolitana, México

ISBN: 978-84-942118-5-0
DOI: http://dx.doi.org/10.3926/oss.21

Contenido

Prólogo

Se pretende que este libro se pueda utilizar en cursos de optimización no lineal de pregrado en especialidades de las ciencias básicas e ingeniería, y otras áreas afines. Se espera que este libro también sea útil como material de consulta a profesionales, ya que cubre aspectos teóricos y computacionales, por ello presenta una variedad de ejemplos y algoritmos para su programación y, al final de cada capítulo se proponen problemas sencillos con aplicación a las ciencias de la ingeniería.

El contenido del libro conjunta el material fundamental de un curso introductorio de optimización no lineal utilizado por los autores, en un período de más de veinte años, en la enseñanza de cursos de optimización matemática de nivel licenciatura para los estudiantes de ingeniería de la Universidad Autónoma Metropolitana Unidad Iztapalapa en México. La motivación principal para escribir esta obra no ha sido la enseñanza de las matemáticas en sí, sino para equipar a los estudiantes en los fundamentos de la optimización matemática y sus algoritmos de manera integral: conceptos, desarrollo de habilidades (para la implementación de programas de cómputo) y metodología en la solución de problemas que se pueden abordar con los métodos de la optimización no lineal en particular. El enfoque particular adoptado aquí por los autores se deriva de experiencias personales en practicar la docencia en las ingenierías en energía y química, donde fue aplicada constantemente para resolver problemas que directamente y de manera fácil pueden abordarse a través del uso cuidadoso de técnicas de optimización matemática.

Buscando darle un orden más claro y didáctico al material, su estructura didáctica incluye los siguientes aspectos:

a) Explicar los temas con toda claridad a partir de su concepto o definición matemática. La formación académica de un ingeniero demanda que se conozca la metodología desde su fundamento, lo cual no significa que se hagan extensas deducciones matemáticas; por ello, se ha intentado emplear conceptos de álgebra lineal, de álgebra de matrices y de algunos temas introductorios donde se requiere el cálculo de varias variables apoyado en ejemplos expuestos de una manera sencilla.

b) Exponer en forma ordenada las metodologías y algoritmos de cálculo.

c) Presentar ejemplos explicados.

d) Poner a disposición del lector problemas propuestos.

e) Emplear una nomenclatura y una simbología ordenadas y lógicas.

Por supuesto se han escrito en el pasado y pasado reciente, muchos textos excelentes y más completos sobre optimización matemática en general, y estamos en deuda con muchos de estos autores por la influencia directa e indirecta que su trabajo ha tenido en la redacción de este texto. Con tantos excelentes textos sobre el tema de la optimización matemática disponibles, se puede plantear la pregunta con justificada razón: ¿Por qué otro libro de optimización y que es lo diferente aquí?

Conscientes de que la optimización es una herramienta importante y necesaria para la formación de los estudiantes de las ciencias e ingeniería, el texto se organizó considerando dos principios básicos; primero, el texto se escribió de manera que aliente el estudio personal. Se ha tratado de conservar la belleza intrínseca de las matemáticas en la optimización; sin embargo, a fin de lograr una mayor flexibilidad y claridad para el estudiante, se simplificó el lenguaje. El rigor matemático se desarrolla con un enfoque computacional, sin ahondar en la demostración de teoremas, lemas y postulados entre otros, ni en el diseño de programas y códigos profesionales; en su lugar se enfoca la idea central y la lógica de cada método. Para la selección de una técnica apropiada y su aplicación, se explican las limitaciones y se proporcionan advertencias acerca de la aplicación ciega y la falta de comprensión de las relaciones matemáticas en la optimización. El segundo principio consiste en dar énfasis e ilustrar los conceptos con ejemplos en el texto para motivar al estudiante e ilustrar la importancia de la optimización en la ingeniería.

El texto tiene una presentación gradual, de lo simple a lo complejo, y se inicia con la introducción de los conceptos elementales de optimización considerando la clasificación de los problemas, las propiedades de las funciones de una y varias variables, las condiciones necesarias y suficientes de optimalidad sin restricciones y de optimización cuadrática, y su interpretación geométrica. Se estudian los métodos unidimensionales con derivadas, donde se revisan los métodos de bisección y Newton-Raphson. El estudio de técnicas de búsqueda unidimensional sin derivadas se realiza empleando la sección aúrea, y el método de interpolación cuadrática. Para lograr la resolución de problemas cada vez más complejos se presentan técnicas de búsqueda multidimensional sin derivadas y con derivadas, como son: simplex irregular, direcciones conjugadas, descenso acelerado, Newton, gradientes conjugados y métodos de métrica variable como DFP y BFGS. Posteriormente, se presentan los métodos de mínimos cuadrados lineales (regresión lineal simple, polinomial y múltiple), métodos de mínimos cuadrados no lineales (métodos de Newton, Gauss-Newton y Levenberg-Marquardt) y métodos de solución de sistemas de ecuaciones no lineales (métodos de Newton y Broyden) como aplicaciones particulares y de gran amplitud para la optimización sin restricciones. Por último, se abordan temas de la optimización con restricciones, como las técnicas de reducción de variable y de multiplicadores de Lagrange, éstas se usan para convertir un problema restringido en uno no restringido, y resolver una función objetivo de menor o mayor número de variables independientes respectivamente.

El texto termina con un apéndice; este puede utilizarse para el repaso de las matemáticas preliminares necesarias para el desarrollo de la optimización de una a varias variables por su contenido básico de conceptos de álgebra lineal y de álgebra de matrices.

CAPÍTULO 1

Conceptos de optimización

1.1 Introducción

La teoría de optimización matemática está constituida por un conjunto de resultados y métodos numéricos enfocados a encontrar e identificar al mejor candidato entre diversas alternativas, sin tener que enumerar y evaluar de manera explícita todas ellas. El proceso de optimización está en la base de la ingeniería, puesto que la función clásica del ingeniero es diseñar sistemas nuevos, mejores, más eficientes y menos costosos.

La potencia de los métodos de optimización para determinar el mejor caso sin comprobar todos los posibles se basa en la utilización de un nivel relativamente modesto de matemáticas y en el desarrollo de cálculos numéricos iterativos empleando procedimientos lógicos claramente definidos o algoritmos pensados para computadoras. El desarrollo de esta metodología de optimización requiere conocimientos básicos de matrices, de álgebra lineal (Apéndice A) y algunos de los resultados del cálculo de varias variables más elementales sobre derivación que se introducirá en este capítulo.

Un problema de optimización es, en general, un problema de decisión. A partir del valor de una función, que se llama la función objetivo y que se diseña para cuantificar el rendimiento y medir la calidad de la decisión, se obtendrán valores para cierto número de variables, relacionadas entre sí mediante expresiones matemáticas, de manera que minimicen o maximicen esa función objetivo y, por lo general, teniendo en cuenta una serie de restricciones que limitan la elección de esos valores.

El planteamiento general para resolver problemas de este tipo es, por tanto, el siguiente:

Se desea optimizar $f(\mathbf{x})$

Sujeta a: Restricciones

El empleo del término "optimizar" en la definición incluye los objetivos de minimización o maximización de la función $f(\mathbf{x})$ cuando el punto $\mathbf{x}$ cumple el conjunto de restricciones. En cualquier caso, siempre se puede transformar un problema con miras a la maximización en otro equivalente para su minimización y viceversa, simplemente multiplicando la función objetivo por -1. En adelante y sin pérdida de generalidad, se considerará, salvo indicación en contrario, que todos los desarrollos, deducciones y problemas presentarán la función objetivo en forma de minimización.

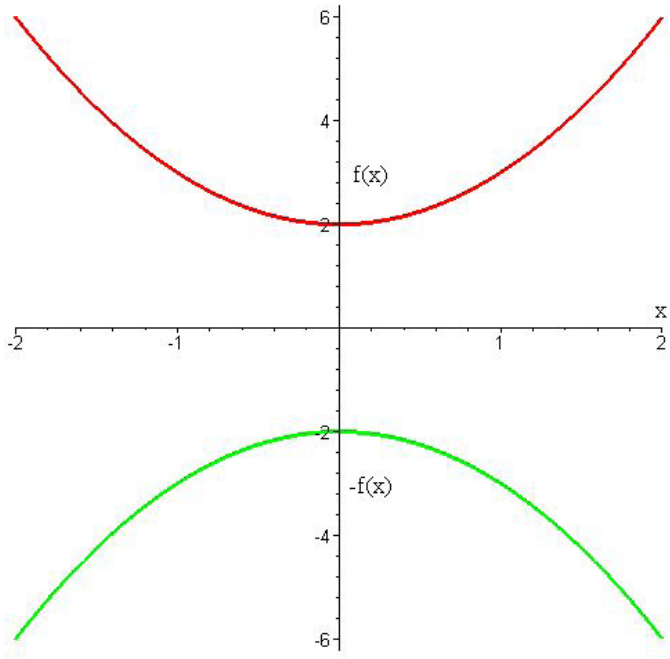


Figura 1.1 *Equivalencia del problema de* $\underset{x\in[a,b]}{Min}\ f(\mathbf{x}) = -\underset{x\in[a,b]}{Max}\left(-f(\mathbf{x})\right)$

Ejemplo 1.1 Verificar que

$$\underset{x\in[-2,2]}{Minimizar}\ f(\mathbf{x}) = -\underset{x\in[-2,2]}{Maximizar}\left(-f(\mathbf{x})\right)$$

considerando las gráficas de las funciones $f(x) = x^2 + 2$ y $-f(x) = -x^2 - 2$ mostradas en la figura 1.1.

Solución:

La función $f(\mathbf{x})$ que se extiende hacia arriba toma su mínimo valor de 2 en el punto $x = 0$. En la misma figura se muestra la función $f(\mathbf{x})$ extendida hacia abajo, es claro que $f(\mathbf{x})$ es una reflexión de $-f(\mathbf{x})$ con respecto al eje x. También es claro de la gráfica que en el mismo punto $x = 0$, ocurre el máximo valor de -2 de $-f(\mathbf{x})$, por lo tanto, la minimización $f(\mathbf{x})$ es equivalente a menos la maximización de $-f(\mathbf{x})$ en el mismo intervalo de interés.

1.2 Elementos de la Optimización

Para aplicar los conceptos matemáticos y técnicas numéricas necesarias de la teoría de optimización en problemas concretos de Ingeniería, es necesario definir previamente lo que se pretende optimizar. El enunciado general de un problema de programación matemática con restricciones podría ser:

Optimizar: $f(\mathbf{x})$

sujeta a: $h_i(\mathbf{x}) = 0 \quad i = 1,2,\ldots,m$

$g_j(\mathbf{x}) \le 0 \quad j = 1,2,\ldots,l$

$\mathbf{x} \in \Omega$

donde $\mathbf{x}^T = (x_1, x_2, \ldots, x_n)$ es el vector de las variables independientes, $f(\mathbf{x})$ es la función objetivo, $\Omega \subseteq R^n$ (aunque puede ser cualquier espacio vectorial), $h_i(\mathbf{x})$ son funciones que representan las restricciones de igualdad mientras que las $g_j(\mathbf{x})$ representan el conjunto de las restricciones de desigualdad. El hecho de que solamente aparezcan restricciones del tipo $g_j(\mathbf{x}) \leq 0$ y no aparezcan restricciones del tipo $g_j(\mathbf{x}) \geq 0$ se debe a que estas últimas pueden transformarse en las primeras multiplicando la desigualdad por -1. En principio, las funciones implícitas en el problema no necesariamente tienen alguna propiedad particular, pero en nuestro caso se van a introducir hipótesis adicionales que nos ayuden a simplificar el problema. Por ejemplo, se supondrá de forma general que las funciones $f(\mathbf{x})$, $h_i(\mathbf{x})$, $g_j(\mathbf{x})$ son continuas y que en la mayoría de los casos tienen derivadas primeras y segundas, también continuas. Además, el conjunto $\boldsymbol{\Omega}$ será en la mayoría de los casos un conjunto convexo, aunque generalmente $\Omega = R^n$.

1.3 Definiciones Básicas

Como se comentó en la sección anterior, el planteamiento general de un problema de programación matemática no lineal es el siguiente:

$$\begin{aligned} &\text{Optimizar: } f(\mathbf{x}) \\ &\text{sujeta a: } \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ &\qquad\quad \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \end{aligned} \qquad 1.1$$

Donde:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{h}(\mathbf{x}) = \begin{pmatrix} h_1(\mathbf{x}) \\ h_2(\mathbf{x}) \\ \vdots \\ h_m(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0} \quad y \quad \mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ g_2(\mathbf{x}) \\ \vdots \\ g_l(\mathbf{x}) \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0}$$

Todas estas $(m+l)$ relaciones deben ser independientes.

Definición 1.1 Una solución factible es un vector $\mathbf{x}$ que satisface todas las restricciones con el grado de precisión requerido.

Definición 1.2 Una región factible es la región del espacio n dimensional constituida por todas las soluciones factibles.

Ejemplo 1.2 Dado el siguiente problema de optimización:

Minimizar: $f(\mathbf{x}) = x_1^3 - 3x_1x_2 + 4$

sujeta a: $h_1(\mathbf{x}) = -2x_1 + x_2 - 5 = 0$

$g_1(\mathbf{x}) = 5x_1 + 2x_2 - 18 \geq 0$

Muestre en un gráfico la región factible del problema de optimización.

Solución:

En la figura 1.2 se muestran las curvas de nivel de la función objetivo, las líneas rectas muestran las restricciones y la región en amarillo corresponde a la región factible.

Definición 1.3 Una solución óptima $\mathbf{x}^*$ es aquella solución factible que produce el valor óptimo de la función objetivo.

La solución óptima al problema del Ejemplo 1.2 es el vector $\mathbf{x}^{*T} = (5, 15)$ que produce el valor óptimo de $f(\mathbf{x}^*) = -96$, como se ilustra en la figura 1.3.

1.4 Propiedades de las Funciones

Continuidad

Definición 1.4 Una función $f(x)$ es continua en el punto $x = x_0$ si

1) $f(x_0)$ está definida,
2) $\lim_{x \to x_0} f(x)$ existe,
3) $\lim_{x \to x_0} f(x) = f(x_0)$.

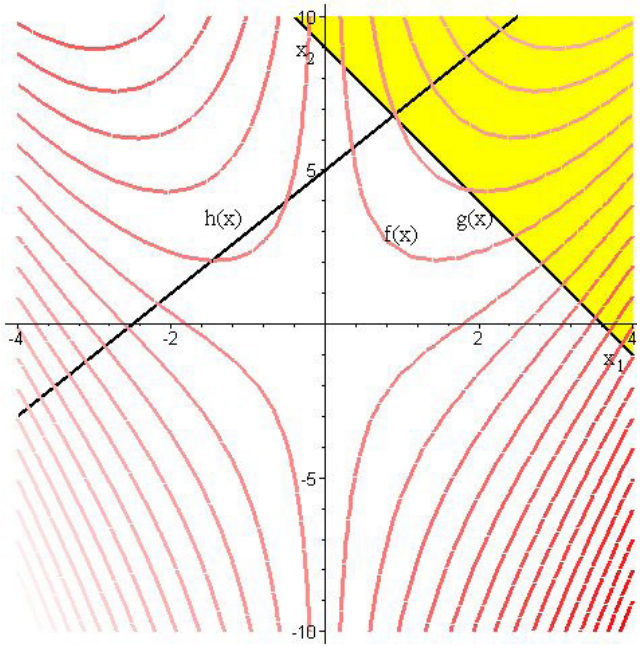


Figura 1.2 *La región factible es la línea negra contenida en la región de color amarillo a lo largo de la restricción* $h(\mathbf{x})$.

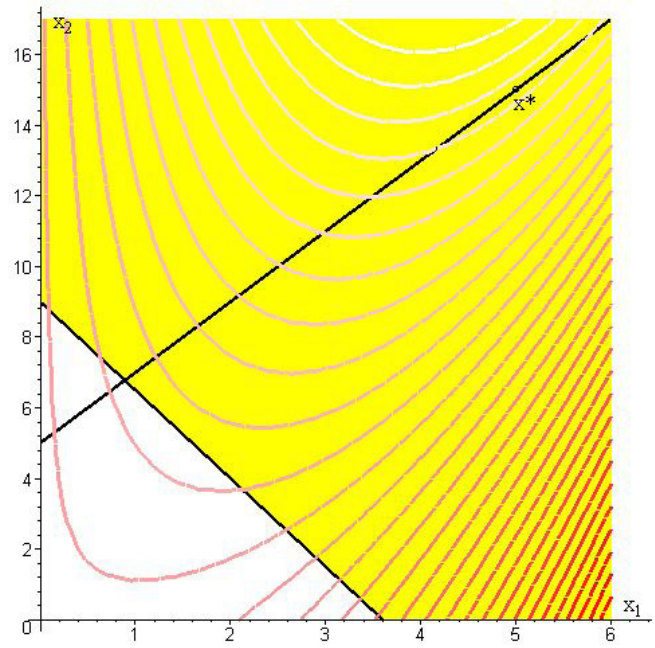


Figura 1.3 *El punto óptimo* $\mathbf{x}^*$ *en la región factible.*

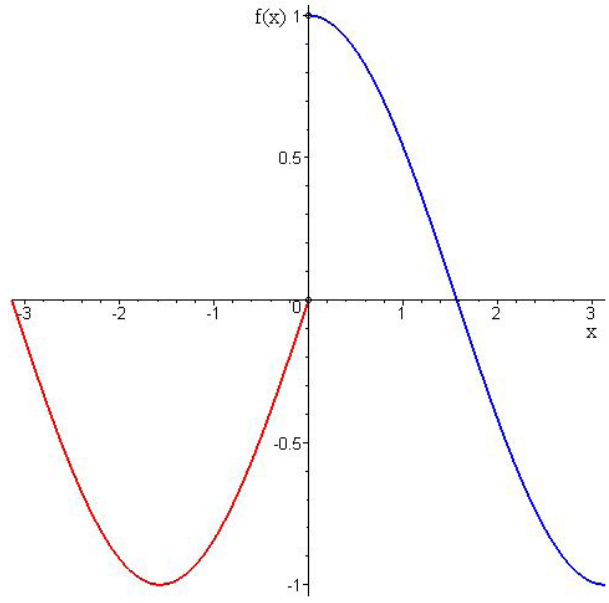


Figura 1.4a *Ejemplo de una función discontinua* $x = 0$.

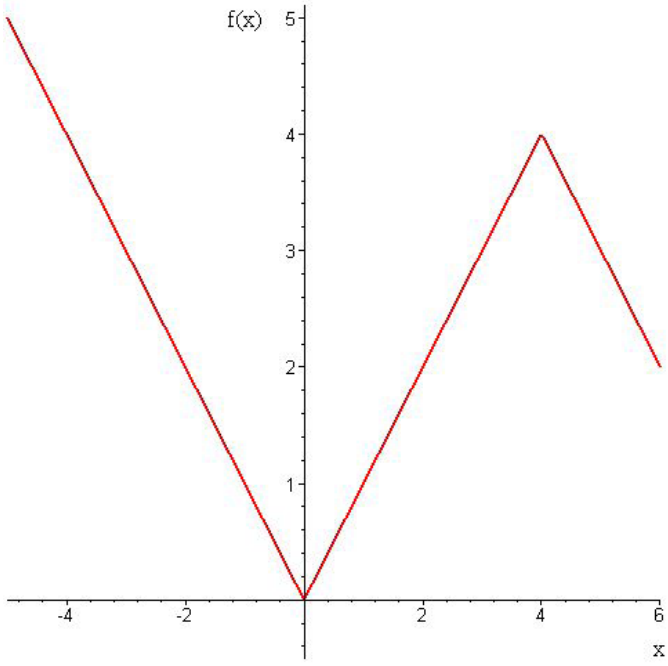


Figura 1.4b *Ejemplo de una función continua no derivable en* $x = 0$ *y* $x = 4$.

Una discontinuidad en una función puede o no causar dificultades en los métodos de optimización. En el caso de la figura 1.4a, el mínimo se da razonablemente lejos de la discontinuidad y puede o no tenerse éxito en su búsqueda. En el caso de la figura 1.4b, si se emplean métodos sin derivada entonces el pico mínimo en $f(x)$ probablemente no será importante, pero con métodos que usan derivadas podría fallar la búsqueda debido a que la derivada resulta indefinida en el punto $x = 0$ y tiene signos diferentes a cada lado de él; entonces, hacer pequeños cambios en x no lleva hacia el óptimo. Un tipo de función objetivo discontinua conocida como función discreta es aquella que solo permite valores discretos de las variables independientes; este tipo de funciones queda fuera del alcance del texto.

Modalidad

Definición 1.5 Un punto extremo es aquel en que se tiene un máximo o mínimo ya sea local o global.

Definición 1.6 El extremo global es el más grande o más pequeño de todos los puntos extremos de un conjunto. Un extremo local es cualquier extremo del conjunto.

En la figura 1.4b en el intervalo [-4,6] tenemos un mínimo y un máximo locales, y ambos extremos en el mismo intervalo son globales.

Definición 1.7 Una función $f(x)$ es unimodal si tiene un sólo extremo en el intervalo $[a,b]$ de definición.

Definición 1.8 Una función $f(x)$ es multimodal si tiene dos o más extremos en el intervalo $[a,b]$ de definición.

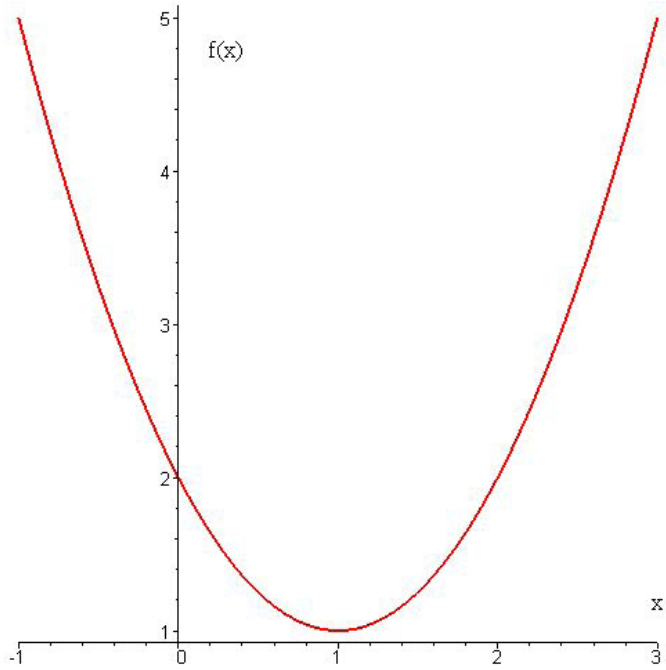


Figura 1.5a *Ejemplo de una función unimodal en el intervalo [-1,3].*

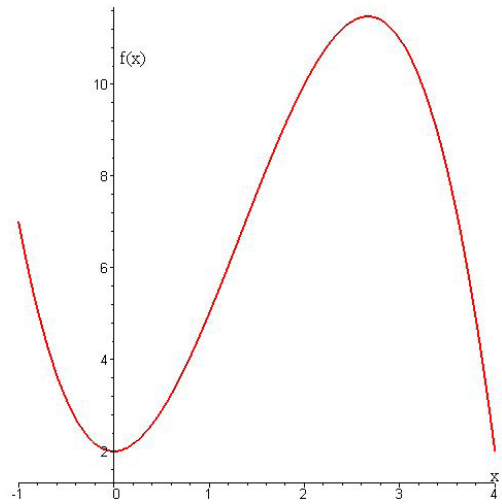


Figura 1.5b *Ejemplo de una función multimodal en el intervalo [-1,4].*

Definición 1.9 Un punto x de $f(x)$ es estacionario si en x se satisface

$$f'(x) = 0$$

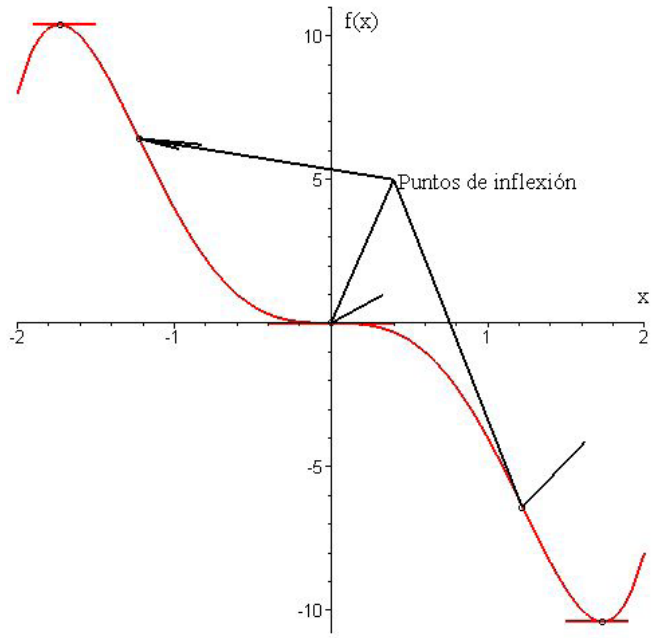


Figura 1.6 *Los puntos estacionarios de la función* $f(x) = x^5 - 5x^3$ *son el máximo, el origen y el mínimo.*

Diferenciación de una función escalar o vectorial respecto a $\mathbf{x}$

Definición 1.10 El símbolo ∇ *(nabla)* es el operador de derivada vectorial y se denota como un vector columna n-dimensional de la siguiente manera:

$$\nabla = \frac{\partial}{\partial \mathbf{x}} = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \cdots, \frac{\partial}{\partial x_n} \right)^T = \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \mathbf{e}_i \qquad (1.2)$$

Usualmente se escribe el subíndice $\mathbf{x}$ en el símbolo del operador nabla ∇ de esta manera: $\nabla_{\mathbf{x}}$, para denotar la derivada con respecto a $\mathbf{x}$. En este texto se despreciará el subíndice $\mathbf{x}$ mientras no haya confusión con respecto a la variable que se quiere derivar. Más adelante será necesario considerar derivaciones de funciones con respecto a alguna otra variable, por ejemplo para la función escalar $f(\mathbf{x},\mathbf{y}): R^{m+n} \rightarrow R$, donde $\mathbf{x}$, $\mathbf{y}$ son vectores n,m-dimensionales respectivamente. Para la diferenciación se usará la siguiente notación:

$$\nabla_{\boldsymbol{x}} f(\mathbf{x},\mathbf{y}) \quad \text{y/o} \quad \nabla_{\boldsymbol{y}} f(\mathbf{x},\mathbf{y})$$

Definición 1.11 Si la función escalar $f(\mathbf{x})$ es derivable en $\mathbf{x}$,entonces el *vector gradiente* $\mathbf{G}(\mathbf{x})$ o $\nabla f(\mathbf{x})$ de la función $f(\mathbf{x})$ es la primera derivada parcial de la función con respecto a $\mathbf{x}$ y se denota con

$$\nabla f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \cdots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right)^T \equiv \mathbf{G}(\mathbf{x}) \qquad (1.3a)$$

En notación de índices se expresa como

$$\nabla f(\mathbf{x}) = \mathbf{G}(\mathbf{x}) = \sum_{i=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_i} \mathbf{e}_i \tag{1.3b}$$

donde $\nabla f(\mathbf{x})$ es una matriz (vector) de orden n x 1 con elementos

$$G_i(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_i}; \quad i = 1, \ldots, n \tag{1.3c}$$

Geométricamente, *el vector gradiente es normal al plano tangente en el punto* $\mathbf{x}$, como se muestra en la figura 1.7 para una función de tres variables. Además, dicho vector apunta en la dirección de *máximo incremento* de la función. Estas propiedades son muy importantes y se utilizarán posteriormente en los métodos numéricos de optimización.

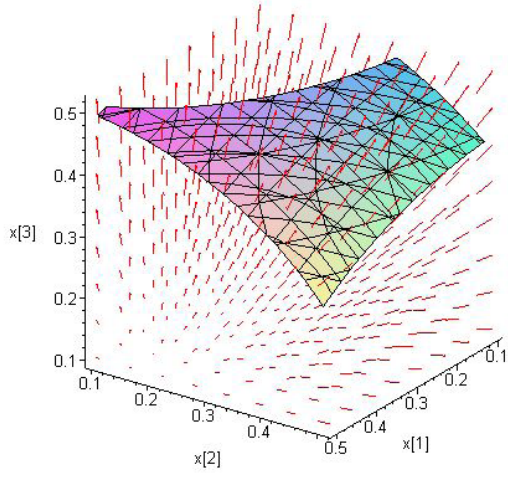


Figura 1.7 *El campo del gradiente pasa en forma perpendicular a la superficie* $f(\mathbf{x}) = x_1^2 + 2x_2^2 + 3x_3^2 - 1 = 0$ *generada por la función*

Definición 1.12 Si la función vectorial $\boldsymbol{f}(\mathbf{x})$ m-dimensional es derivable en $\mathbf{x}$, entonces la *matriz jacobiana* $\mathbf{J}(\mathbf{x})$ de la función $\boldsymbol{f}(\mathbf{x})$ es la matriz de primeras derivadas parciales de la función con respecto a $\mathbf{x}$ y se denota con

$$\left(\nabla \boldsymbol{f}^T(\mathbf{x})\right)^T = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_2(\mathbf{x})}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \frac{\partial f_m(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{pmatrix} \equiv \mathbf{J}(\mathbf{x}) \tag{1.4a}$$

y en notación de índices se expresa como

$$\mathbf{J}(\mathbf{x}) = \sum_{i,k=1}^{m,n} J_{ik}(\mathbf{x}) \mathbf{e}_i \mathbf{e}_k^T \tag{1.4b}$$

donde $\mathbf{J}$ es una matriz de orden $m\,x\,n$ con elementos

$$J_{ik}(\mathbf{x}) = \frac{\partial f_i(\mathbf{x})}{\partial x_k}; \quad \begin{matrix} i = 1,\ldots,m \\ k = 1,\ldots,n \end{matrix} \tag{1.4c}$$

Ejemplo 1.3 Para la función $\boldsymbol{f}(\mathbf{x}) = \begin{pmatrix} 3x_1^2 - 4x_1x_2 + x_2^2 \\ e^{x_1} + x_2^2 + 1 \end{pmatrix}$, calcule la matriz jacobiana en el punto $\mathbf{x}^{*T} = (1,1)$.

Solución:

La matriz jacobiana es

$$\mathbf{J}(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial f_1(\mathbf{x})}{\partial x_1} & \dfrac{\partial f_1(\mathbf{x})}{\partial x_2} \\ \dfrac{\partial f_2(\mathbf{x})}{\partial x_1} & \dfrac{\partial f_2(\mathbf{x})}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 6x_1 - 4x_2 & -4x_1 + 2x_2 \\ e^{x_1} & 2x_2 \end{pmatrix}$$

El valor de la matriz en $\mathbf{x}^{*T} = (1,1)$ es $\mathbf{J}(\mathbf{x}^*) = \begin{pmatrix} 2 & -2 \\ e & 2 \end{pmatrix}$.

Definición 1.13 Si la función escalar $f(\mathbf{x})$ es doble continuamente derivable en $\mathbf{x}$, entonces la *matriz hessiana* $\mathbf{H}(\mathbf{x})$ de la función $f(\mathbf{x})$ es la matriz de segundas derivadas parciales de la función con respecto de $\mathbf{x}$ y se denota como

$$\nabla^2 f(\mathbf{x}) = \nabla\left(\nabla^T f(\mathbf{x})\right) = \begin{pmatrix} \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{pmatrix} \quad (1.5a)$$

y en notación de índices se expresa como

$$\mathbf{H}(\mathbf{x}) = \sum_{i,j=1}^{n} h_{ij}(\mathbf{x}) \boldsymbol{e}_i \boldsymbol{e}_j^T \quad (1.5b)$$

donde $\mathbf{H}$ es una matriz simétrica de orden $n\,x\,n$ con elementos

$$h_{ij}(\mathbf{x}) = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}; \qquad \begin{matrix} i = 1,\ldots,n \\ j = 1,\ldots,n \end{matrix} \tag{1.5c}$$

Es importante hacer notar que cada elemento de la matriz hessiana es una función en si misma que debe evaluarse en el punto dado $\mathbf{x}$. Debido a que $f(\mathbf{x})$ se supone dos veces continuamente diferenciable, *las derivadas parciales cruzadas son iguales*, es decir,

$$h_{ij}(\mathbf{x}) = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} = \frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_i} = h_{ji}(\mathbf{x}); \qquad i \neq j \tag{1.5d}$$

Por lo tanto, ésta desempeña un papel muy importante en las condiciones de suficiencia de las condiciones de optimalidad que se verán más adelante. Finalmente, obsérvese que

$$\frac{\partial x_i}{\partial x_j} = \delta_{ij}\,; \qquad i, j = 1,2,...,n \tag{1.6}$$

Ejemplo 1.4 Calcule la matriz hessiana en el punto $\mathbf{x}^{*T} = (1, 2)$ para la función $f(\mathbf{x}) = x_1^3 + x_2^3 + 2x_1^2 + 3x_2^2 - x_1x_2 + 2x_1 + 4x_2$.

Solución:

La matriz hessiana es

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} 6x_1 + 4 & -1 \\ -1 & 6x_2 + 6 \end{pmatrix}$$

por lo que la matriz hessiana en el punto $\mathbf{x}^{*T} = (1, 2)$ es $\mathbf{H}(\mathbf{x}^*) = \begin{pmatrix} 10 & -1 \\ -1 & 18 \end{pmatrix}$.

Aproximación de funciones por expansión en serie de Taylor

Algunos de los procedimientos de programación matemática por analizar más adelante requieren de aproximaciones polinomiales de $f(\mathbf{x})$, $h_i(\mathbf{x})$, $i = 1, \ldots, l$ y $g_j(\mathbf{x})$, $j = 1, \ldots, m$, en la vecindad de cualquier punto en términos de su valor y sus derivadas. Por lo tanto, se repasará el desarrollo en serie de Taylor para una función de varias variables alrededor de $\mathbf{x} = \mathbf{x}_0$. Entonces, la expansión en serie de Taylor alrededor de $\mathbf{x} = \mathbf{x}_0$ para $f(\mathbf{x})$ está dada por

$$\overline{f}(\mathbf{x}) = \sum_{n=0}^{\infty} \frac{\left(\mathbf{h}^T \cdot \nabla\right)^n f(\mathbf{x}_0)}{n!}; \quad \mathbf{h} = \mathbf{x} - \mathbf{x}_0 \tag{1.7a}$$

de manera que una *aproximación lineal* o de *primer orden* de la función $f(\mathbf{x})$, puede hacerse por una serie de Taylor truncada alrededor de $\mathbf{x}_0$ como la siguiente:

$$\overline{f}(\mathbf{x}) \approx f(\mathbf{x}_0) + \left(\mathbf{x} - \mathbf{x}_0\right)^T \nabla f(\mathbf{x}_0) \tag{1.7b}$$

y una *aproximación cuadrática* o *de segundo orden* de $f(\mathbf{x})$ puede hacerse ignorando los términos de tercer orden y superiores en la serie de Taylor; se obtiene entonces

$$\overline{f}(\mathbf{x}) \approx f(\mathbf{x}_0) + \left(\mathbf{x} - \mathbf{x}_0\right)^T \nabla f(\mathbf{x}_0) + \frac{1}{2}\left(\mathbf{x} - \mathbf{x}_0\right)^T \nabla^2 f(\mathbf{x}_0)\left(\mathbf{x} - \mathbf{x}_0\right) \tag{1.7c}$$

Ejemplo 1.5 Obtenga una aproximación de segundo orden de la función $f(x)=\cos x$ alrededor del punto $x_0=0$.

Solución:

Usando la ecuación (1.7c), la expansión de Taylor de segundo orden para $\cos x$ en el punto $x_0=0$ es

$$\cos x \approx \cos 0-(x-0)\,sen\,0+\frac{1}{2}(x-0)^2(-\cos 0)=1-\frac{1}{2}x^2$$

Ejemplo 1.6 Obtenga el desarrollo de segundo orden de Taylor para la función $f(\mathbf{x})=3x_1^3x_2$ en el punto $\mathbf{x}_0^T=(1,1)$.

Solución:

El gradiente y la hessiana de la función $f(\mathbf{x})$ en el punto $\mathbf{x}_0^T=(1,1)$ usando las ecuaciones (1.3a) y (1.5c) son

$$\nabla f(\mathbf{x}_0)=\begin{pmatrix}9\\3\end{pmatrix} \quad \text{y} \quad \mathbf{H}(\mathbf{x}_0)=\begin{pmatrix}18 & 9\\9 & 0\end{pmatrix}$$

sustituyendo estas expresiones en la forma matricial de la expansión de Taylor dada por la ecuación (1.7c) y usando

$$\mathbf{x}-\mathbf{x}_0=\begin{pmatrix}x_1\\x_2\end{pmatrix}-\begin{pmatrix}1\\1\end{pmatrix}=\begin{pmatrix}x_1-1\\x_2-1\end{pmatrix}, \quad f(\mathbf{x}_0)=3$$

se obtiene una aproximación $\overline{f}(\mathbf{x})$ a $f(\mathbf{x})$ dada por

$$\overline{f}(\mathbf{x}) = 3 + (x_1 - 1, x_2 - 1)\begin{pmatrix} 9 \\ 3 \end{pmatrix} + \frac{1}{2}(x_1 - 1, x_2 - 1)\begin{pmatrix} 18 & 9 \\ 9 & 0 \end{pmatrix}\begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix}$$

Al simplificar la expresión anterior realizando las operaciones entre vectores y matrices se obtiene

$$\overline{f}(\mathbf{x}) = 9x_1^2 + 9x_1x_2 - 18x_1 - 6x_2 + 9$$

Esta expresión es una aproximación cuadrática de la función $f(\mathbf{x}) = 3x_1^3x_2$ alrededor del punto $\mathbf{x}_0^T = (1, 1)$.

1.5 Concavidad y Convexidad

La determinación de la concavidad o convexidad de una función y (o) su región de definición nos ayudará a establecer si una solución óptima local es también la solución óptima global, es decir, la mejor entre todas las soluciones. Cuando se sabe que la función objetivo tiene ciertas propiedades que se definirán después, el cálculo del óptimo puede acelerarse usando algoritmos de optimización apropiados.

Definición 1.14 Una región Ω es *convexa* si el segmento de línea que une cualesquiera dos puntos en la región cae enteramente en la región. Así, si $\mathbf{x}_1$ y $\mathbf{x}_2 \in \Omega$, todo punto de la forma

$$\mathbf{x} = \theta\mathbf{x}_2 + (1 - \theta)\mathbf{x}_1 \tag{1.8}$$

donde $0 \leq \theta \leq 1$, también está en la región.

Como ya se mencionó, en problemas de optimización con restricciones los puntos que satisfacen todas las restricciones son puntos factibles, el resto de los otros puntos son no factibles. Las restricciones de desigualdad determinan una región factible constituida por el conjunto de puntos que son factibles, mientras que las restricciones de igualdad limitan el conjunto de puntos factibles a hipersuperficies, a curvas o incluso a un solo punto.

Definición 1.15 Un *punto interior* es aquel punto que satisface todas las restricciones de desigualdad estrictamente como desigualdades.

Definición 1.16 Un *punto frontera* es aquel que satisface todas las restricciones de desigualdad como igualdades.

Los demás puntos son *puntos exteriores.*

Definición 1.17 Un *contorno o curva de nivel* de $f(\mathbf{x})$ es el conjunto de puntos donde la función objetivo tiene un valor constante, como se ilustra en la figura 1.10.

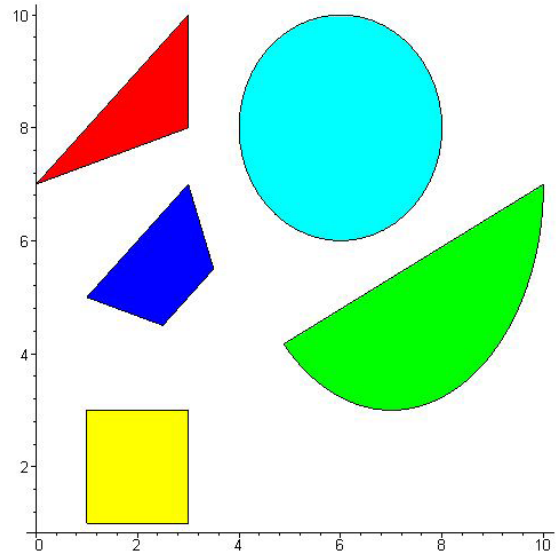


Figura 1.9a *Diferentes conjuntos que muestran regiones convexas.*

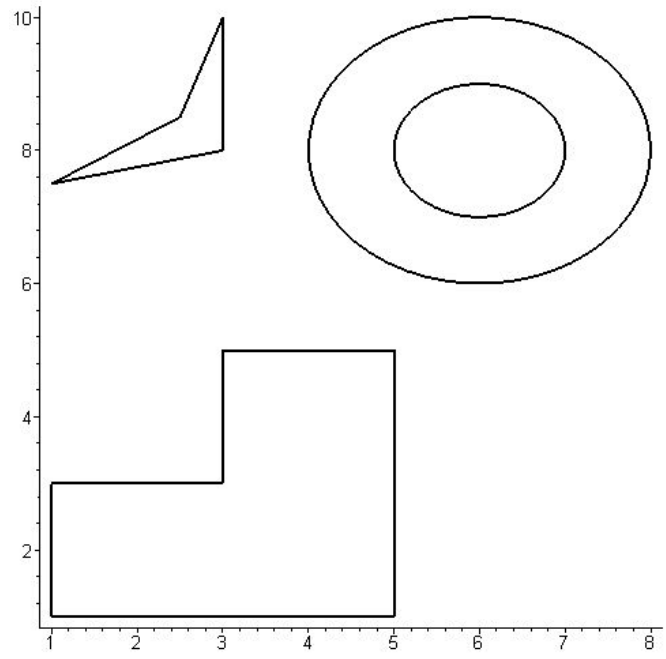


Figura 1.9b *Diferentes conjuntos que muestran regiones no convexas.*

Definición 1.18 Se dice que una *restricción de desigualdad* es activa si ésta se satisface en su frontera como una igualdad, es decir, si

$$g_j(\mathbf{x}) = 0 \text{ ; para alguna } j.$$

Si la restricción de desigualdad se satisface como una desigualdad, se dice que la restricción es *inactiva.*

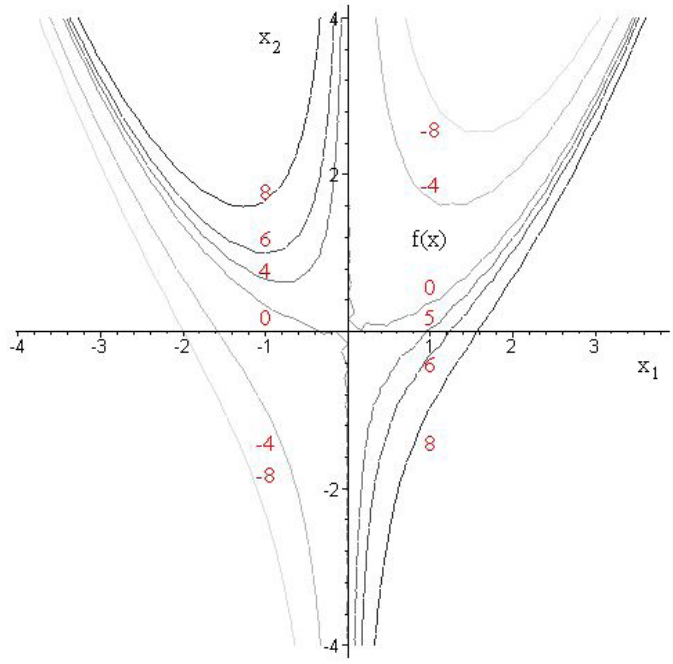


Figura 1.10 *Contornos de la función* $f(\mathbf{x}) = x_1^3 - 3x_1x_2 + 4$ *y sus valores.*

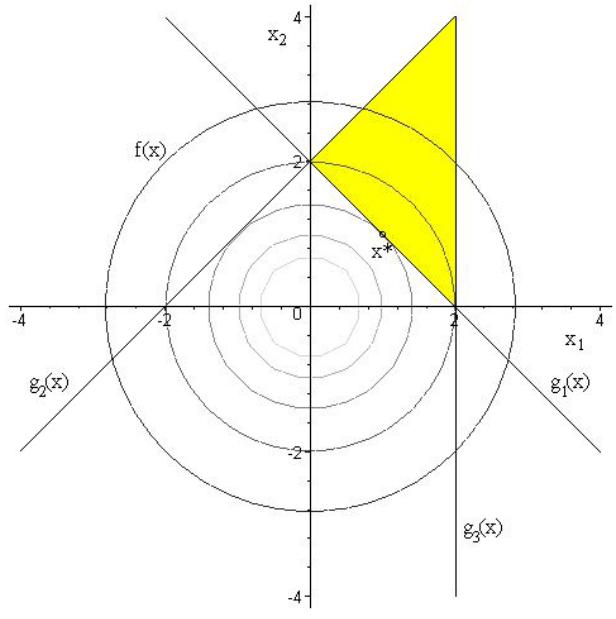


Figura 1.11 *El mínimo restringido cae sobre la restricción activa* $g_1(\mathbf{x}) = 0$.

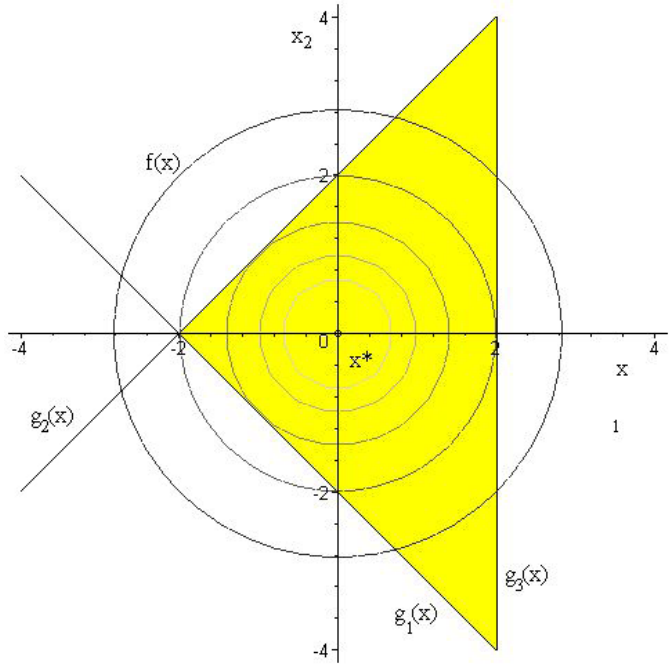


Figura 1.12 *El mínimo restringido coincide con el mínimo no restringido.*

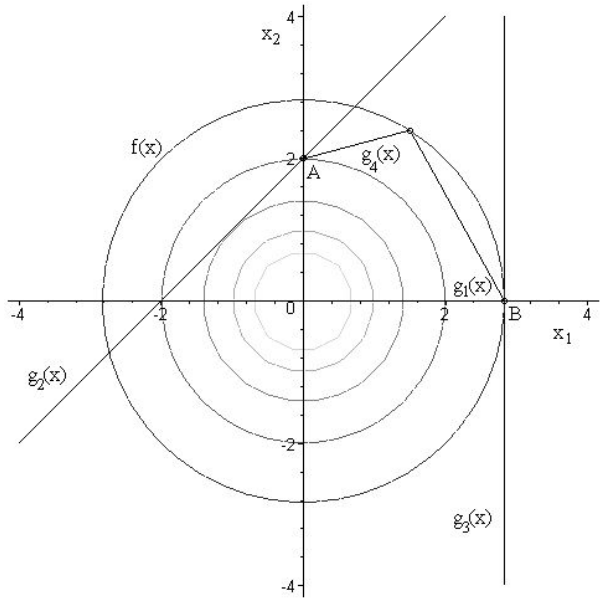


Figura 1.13 *Los mínimos locales A y B restringidos caen sobre las restricciones activas* $g_2(A)=0$, $g_4(A)=0$, $g_1(B)=0$ *y* $g_3(B)=0$

En la figura 1.11, el mínimo no restringido $\mathbf{x}^{*T} = (0, 0)$ cae fuera de la región factible y el mínimo restringido $\mathbf{x}^{*T} = (1, 1)$ cae sobre la restricción de desigualdad $g_1(\mathbf{x})$ activa; las otras dos restricciones son inactivas. Si se tuviera conocimiento a priori de ésta situación, se debería ignorar $g_2(\mathbf{x})$ y $g_3(\mathbf{x})$ y tratar este problema como un problema con una restricción de igualdad $g_1(\mathbf{x}) = 0$ únicamente. En la figura 1.12 el mínimo restringido $\mathbf{x}^{*T} = (0, 0)$ coincide con el mínimo no restringido $\mathbf{x}^{*T} = (0, 0)$. Todas las restricciones se satisfacen como desigualdades estrictas y si se tuviera conocimiento de esto, deberían ignorarse todas por ser inactivas y tratar este problema como un problema de optimización sin restricciones, también es posible para las restricciones introducir mínimos locales dentro del problema (véase la Figura 1.13). Aquí, la función tiene únicamente un mínimo no restringido $\mathbf{x}^{*T} = (0, 0)$. Sin embargo, para el problema restringido, A y B son mínimos locales ya que ningún punto factible en las cercanías inmediatas de A o B da valores menores de la función. La figura 1.13 también muestra la importancia de que las restricciones constituyan una región convexa si se desea localizar un mínimo global. La búsqueda para el mínimo de la función, si es iniciada en la parte superior de la región factible, podría terminar en el punto A, mientras que una búsqueda empezando en la parte inferior de la región podría terminar en B. Por lo tanto, se concluye que la búsqueda en una región convexa es importante para obtener resultados adecuados en optimización. De hecho, si la región factible fuera convexa, la búsqueda desde cualquier punto inicial convergería a la misma respuesta.

Definición 1.19 Se dice que una función $f(\mathbf{x})$ es *convexa* sobre la región convexa Ω si para cualesquiera dos puntos $\mathbf{x}_1, \mathbf{x}_2 \in \Omega$

$$f\left(\theta\mathbf{x}_2 + (1-\theta)\mathbf{x}_1\right) \leq \theta f\left(\mathbf{x}_2\right) + \left(1-\theta\right) f\left(\mathbf{x}_1\right) \tag{1.9}$$

para $0 \leq \theta \leq 1$. Si la desigualdad se satisface como una desigualdad estricta, entonces se dice que la función $f(\mathbf{x})$ es *estrictamente convexa.*

Para una función de una variable, la desigualdad significa que la función pasa por abajo de la cuerda que une dos puntos de su gráfica (véase la Figura 1.14).

Para una función cóncava definida sobre una región convexa, simplemente se invierte la desigualdad para obtener

$$f\left(\theta \mathbf{x}_2 + (1-\theta)\mathbf{x}_1\right) \geq \theta f\left(\mathbf{x}_2\right) + \left(1-\theta\right) f\left(\mathbf{x}_1\right) \tag{1.10}$$

Se dice que la función es estrictamente cóncava si la desigualdad se satisface como una desigualdad estricta. Una función de este tipo pasa por arriba de la cuerda que une dos puntos sobre su gráfica (véase la Figura 1.15).

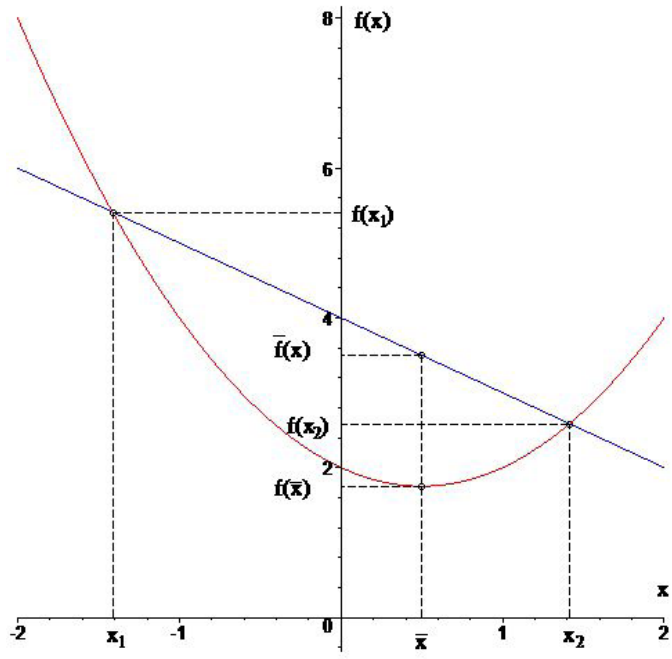


Figura 1.14 *Función convexa según la definición 1.19,* $\overline{x} = \theta x_2 + \left(1-\theta\right) x_1$ *y* $\overline{f}\left(x\right) = \theta f\left(x_2\right) + \left(1-\theta\right) f\left(x_1\right)$

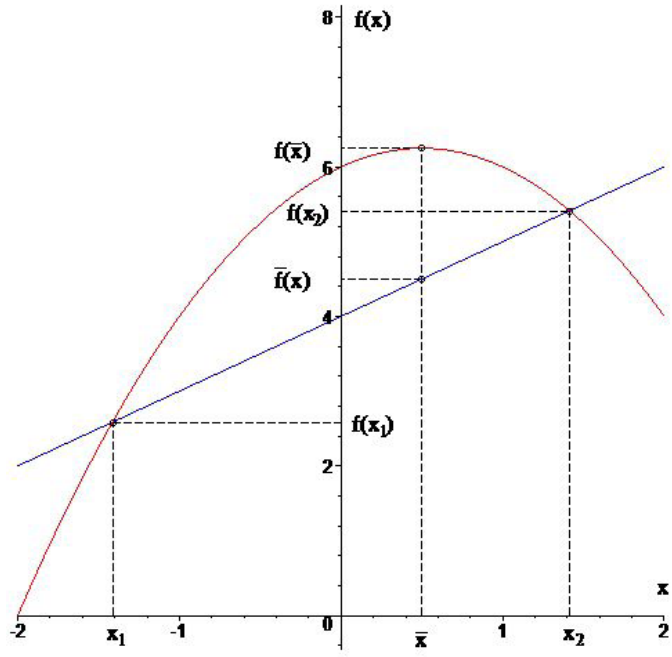


Figura 1.15 *Función cóncava según la definición 1.19,* $\overline{x} = \theta x_2 + (1-\theta)x_1$ *y* $\overline{f}(x) = \theta f(x_2) + (1-\theta) f(x_1)$.

Ejemplo 1.7 Considere el siguiente problema de optimización:

$$\textit{Minimizar} \quad f(\mathbf{x}) = \left(x_1 - \frac{1}{2}\right)^2 + \left(x_2 - \frac{1}{2}\right)^2$$

$$\textit{Sujeta a}: \qquad g(\mathbf{x}) = \frac{1}{x_1} + \frac{1}{x_2} - 2 \leq 0$$

La figura 1.16 muestra los contornos de la función objetivo de dos variables y la frontera de la restricción $g(\mathbf{x}) = 0$; ahora considere los puntos $\mathbf{x}_1$ y $\mathbf{x}_2$ sobre la frontera de la restricción definidos por

$$\mathbf{x}_1 = \begin{pmatrix} 0.667 \\ 2.000 \end{pmatrix} \text{ y } \mathbf{x}_2 = \begin{pmatrix} 3.000 \\ 0.600 \end{pmatrix}$$

La función de restricción es convexa si cualquier punto $\mathbf{x} = \theta\mathbf{x}_2 + (1-\theta)\mathbf{x}_1$ sobre la línea de conexión de los puntos $\mathbf{x}_1$ y $\mathbf{x}_2$ corresponde a un valor de $g(\mathbf{x}) \leq 0$ donde los dos puntos pueden estar en cualquier parte de la superficie de restricción. Considérese un valor de $\theta = 0.5$;se sabe que $g(\mathbf{x}_1) = 0$ y $g(\mathbf{x}_2) = 0$ ya ya que caen en la frontera de $g(\mathbf{x})$. Entonces,

$$\mathbf{x} = \theta\mathbf{x}_2 + (1-\theta)\mathbf{x}_1 = 0.5\begin{pmatrix}3.000\\0.600\end{pmatrix} + (1-0.5)\begin{pmatrix}0.667\\2.000\end{pmatrix} = \begin{pmatrix}1.833\\1.300\end{pmatrix}$$

y

$$g(\mathbf{x}) = \frac{1}{1.833} + \frac{1}{1.300} - 2 = -0.685 \leq 0$$

En verdad, se puede mostrar que para $0 \leq \theta \leq 1$, el valor de la restricción siempre será menor o igual que cero y así la función de restricción es convexa. En forma similar, la evaluación de $f(\mathbf{x})$ en $\mathbf{x}_1$ y $\mathbf{x}_2$ da

$$f(\mathbf{x}_1) = 2.278 \quad \text{y} \quad f(\mathbf{x}_2) = 6.260$$

otra vez, considerando un valor de $\theta = 0.5$ se tiene

$$\theta f(\mathbf{x}_2) + (1-\theta) f(\mathbf{x}_1) = 4.269 \text{ y } f(\theta\mathbf{x}_2 + (1-\theta)\mathbf{x}_1) = 2.417$$

de la ecuación (1.9),

$$f(\theta\mathbf{x}_2 + (1-\theta)\mathbf{x}_1) = 2.417 \leq 4.269 = \theta f(\mathbf{x}_2) + (1-\theta) f(\mathbf{x}_1)$$

Se ve que la desigualdad se satisface (y se satisface para cualquier $0 \leq \theta \leq 1$) y así la función objetivo satisface el requerimiento de convexidad. Existen, además, otras propiedades importantes de las funciones convexas o cóncavas que pueden deducirse usando las desigualdades anteriores.

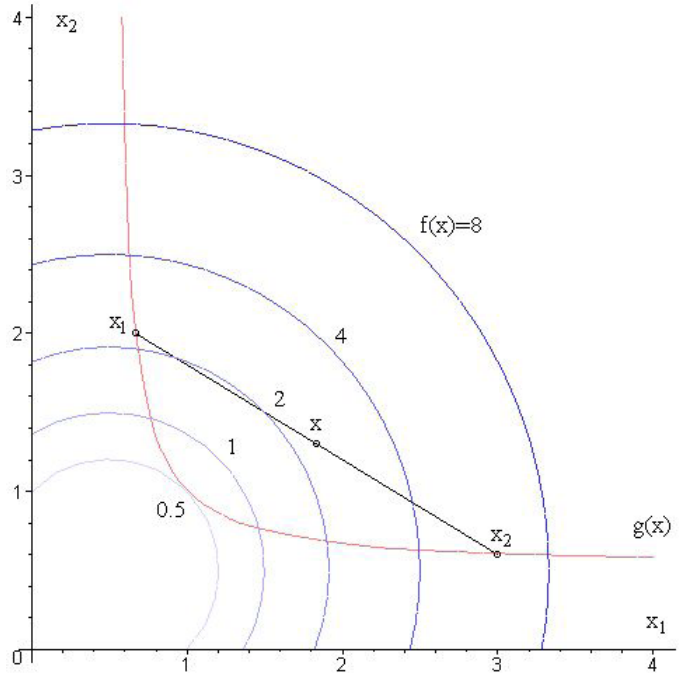


Figura 1.16 *Una función convexa en una región convexa.*

Proposición 1.1 Si $f(\mathbf{x})$ es convexa sobre la región convexa Ω y $\mathbf{x}_1, \mathbf{x}_2 \in \Omega$, entonces

$$f(\mathbf{x}_2) \geq f(\mathbf{x}_1) + (\mathbf{x}_2 - \mathbf{x}_1)^T \nabla f(\mathbf{x}_1) \tag{1.11}$$

Para funciones *convexas* de una o dos variables, la desigualdad dice que tales funciones pasan por arriba de cualquier línea o plano tangente a las funciones, como se ilustra en la figura 1.17.

Para una función *cóncava* la desigualdad se invierte (ver Figura 1.18).

$$f(\mathbf{x}_2) \leq f(\mathbf{x}_1) + (\mathbf{x}_2 - \mathbf{x}_1)^T \nabla f(\mathbf{x}_1) \tag{1.12}$$

Las ecuaciones de desigualdad 1.9 a 1.12 en la definición de funciones convexas y cóncavas no son muy convenientes cuando se usan para probar las propiedades de convexidad o concavidad; en su lugar puede usarse la segunda derivada de $f(x)$ o la matriz hessiana de $f(\mathbf{x})$ si $\mathbf{x}$ es un vector.

Ejemplo 1.8 El gradiente de la función del ejemplo 1.7 evaluado en el punto $\mathbf{x}_1$ esta dado por

$$\nabla f(\mathbf{x}_1) = \begin{pmatrix} 2x_1 - 1 \\ 2x_2 - 1 \end{pmatrix} = \begin{pmatrix} 0.334 \\ 3.000 \end{pmatrix}$$

y la diferencia de vectores es

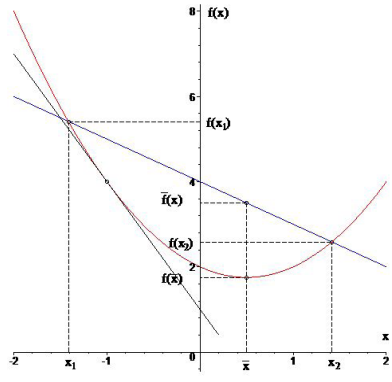


Figura 1.17 *Función convexa según la proposición 1.1*

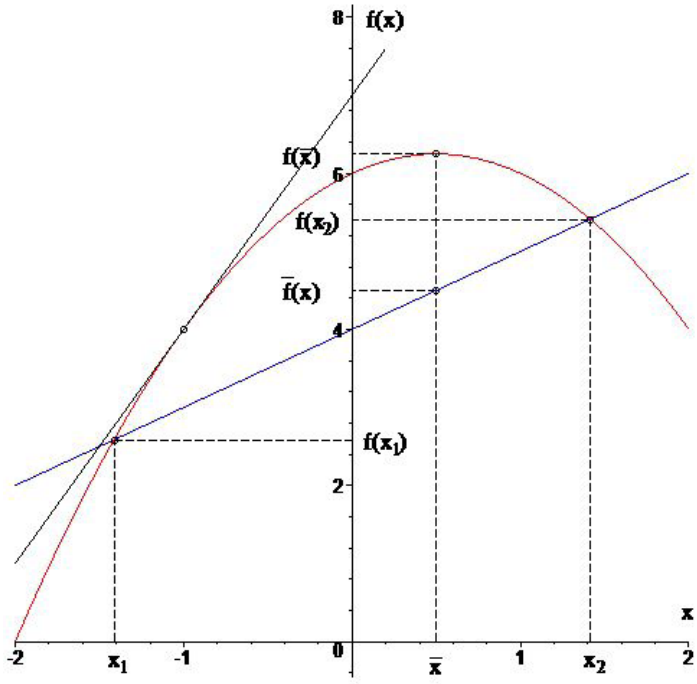


Figura 1.18 *Función cóncava según la proposición 1.1*

$$\mathbf{x}_2 - \mathbf{x}_1 = \begin{pmatrix} 2.333 \\ -1.400 \end{pmatrix}$$

Por otro lado, la evaluación de $f(\mathbf{x})$ en $\mathbf{x}_1$ y $\mathbf{x}_2$ fue de

$$f(\mathbf{x}_1) = 2.278 \text{ y } f(\mathbf{x}_2) = 6.260$$

y entonces

$$f(\mathbf{x}_1) + (\mathbf{x}_2 - \mathbf{x}_1)^T \nabla f(\mathbf{x}_1) = 2.278 - 3.420 = -1.142$$

de la ecuación (1.11)

$$f(\mathbf{x}_2) = 6.260 \geq -1.142 = f(\mathbf{x}_1) + (\mathbf{x}_2 - \mathbf{x}_1)^T \nabla f(\mathbf{x}_1)$$

y se ve que la desigualdad se satisface de manera estricta, así que la función objetivo satisface el requerimiento de convexidad estricta.

Proposición 1.2 Una función es *convexa* si su matriz hessiana $\mathbf{H} \equiv (h_{ij})_{nxn}$ es *positiva semidefinida* y es *estrictamente convexa* si la matriz hessiana es *positiva definida*.

Para funciones convexas de una variable, esto dice que la segunda derivada es positiva de manera que la primera derivada es una función creciente que puede anularse sólo en un punto. De esta manera, tal función tendrá un punto mínimo, como se ilustra en la figura 1.19.

Proposición 1.3 Una función es *cóncava* si la matriz hessiana es negativa *semidefinida* y es *estrictamente cóncava* si la matriz hessiana es *negativa definida.*

Para funciones cóncavas de una variable, esto dice que la segunda derivada es negativa de manera que la primera derivada es una función decreciente que puede anularse sólo en un punto. De esta manera, tal función tendrá un punto máximo, como se ilustra en la figura 1.20.

Ejemplo 1.9 La matriz hessiana de la función objetivo del ejemplo 1.7 esta dada por

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

Los determinantes de los menores principales son positivos $(D_1 = 2, D_2 = 4)$, o bien sus valores propios son positivos $(\lambda_1 = 2, \lambda_2 = 2)$ de manera que la matriz hessiana es positiva definida y de aquí que la función objetivo satisfaga el requerimiento de convexidad.

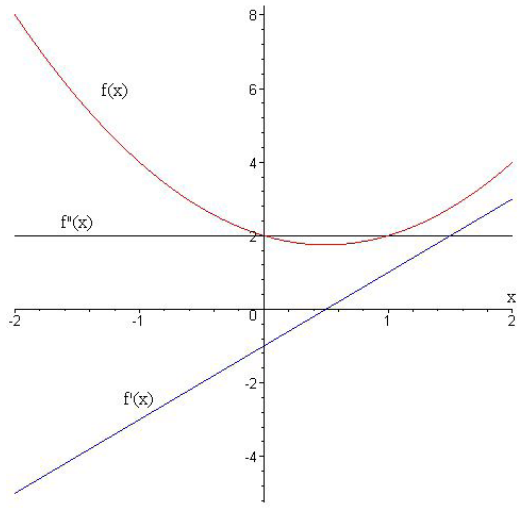


Figura 1.19 *Función convexa con derivada creciente y segunda derivada positiva.*

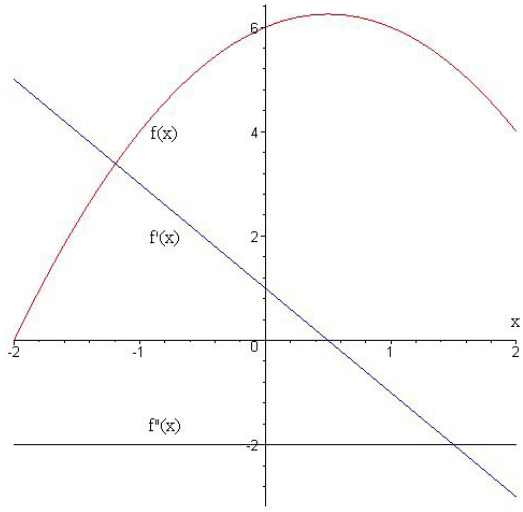


Figura 1.20 *Función cóncava con derivada decreciente y segunda derivada negativa.*

Proposición 1.4 Si las restricciones de desigualdad $g_j(\mathbf{x}) \leq 0$ con $j = 1, 2, \ldots, m$ son funciones *convexas* sobre la región convexa Ω, entonces

$$p(\mathbf{x}) = \sum_{j=1}^{m} \alpha_j g_j(\mathbf{x})$$

donde $\alpha_j \geq 0$, es también *convexa*.

Una combinación lineal de funciones convexas produce una nueva función convexa.

Proposición 1.5 Sea Ω la región de restricción definida por las $g_j(\mathbf{x}) \leq 0$ con $j = 1, 2, \ldots, m$, donde las $g_j(\mathbf{x})$ son *convexas*, entonces la región de restricción Ω es *convexa*.

Ejemplo 1.10 ¿Es convexa la siguiente región construida por las cuatro restricciones?

$$\begin{aligned} x_2 &\geq 1 - x_1 \\ x_2 &\leq 1 + 0.5x_1 \\ x_1 &\leq 2 \\ x_2 &\geq 0 \end{aligned}$$

Solución:

Si se reescriben e identifican las restricciones de la forma $g_j(\mathbf{x}) \leq 0$, con $j = 1,\ldots,4$ se tiene

$$g_1(\mathbf{x}) = 1 - x_1 - x_2 \leq 0$$
$$g_2(\mathbf{x}) = x_2 - 0.5x_1 - 1 \leq 0$$
$$g_3(\mathbf{x}) = x_1 - 2 \leq 0$$
$$g_4(\mathbf{x}) = -x_2 \leq 0$$

Todas las funciones de restricción son lineales y éstas son funciones tanto convexas como cóncavas; por tanto, la región construida por ellas es convexa. Como se ilustra en la figura 1.21.

Proposición 1.6 Si $f(\mathbf{x})$ es una función *convexa* sobre la región restringida por $g_j(\mathbf{x}) \leq 0$, donde las $g_j(\mathbf{x})$ son *convexas*, entonces un mínimo local de $f(\mathbf{x})$ en esta región es su mínimo global en la misma región.

Esta proposición caracteriza el conocido problema de programación convexa.

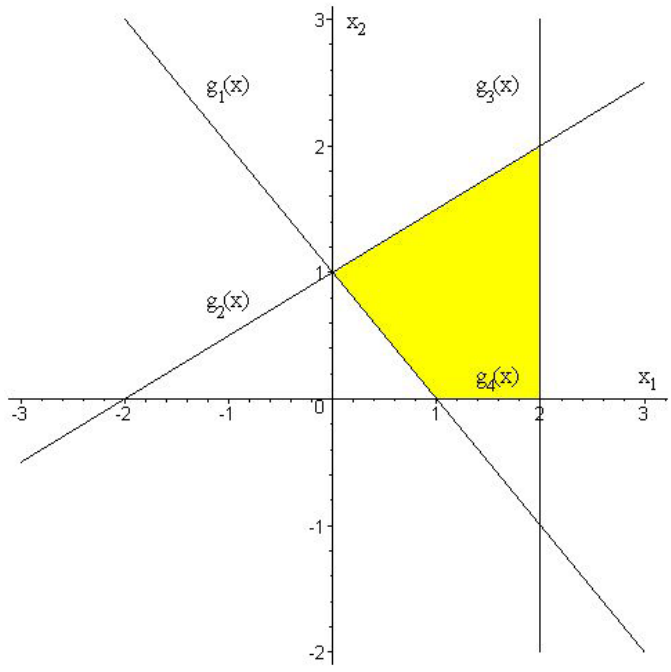


Figura 1.21 *Región convexa construida por las restricciones del ejemplo 1.10.*

Ejemplo 1.11 Considere el problema de optimización

$$\begin{aligned} \textit{Minimizar} \quad & f(\mathbf{x}) = (x_1 - 1)^2 + (x_2 - 1)^2 \\ \textit{Sujeta a} \quad & g_1(\mathbf{x}) = 1 - x_1 - x_2 \leq 0 \\ & g_2(\mathbf{x}) = x_2 - 0.5x_1 - 1 \leq 0 \\ & g_3(\mathbf{x}) = x_1 - 2 \leq 0 \\ & g_4(\mathbf{x}) = -x_2 \leq 0 \end{aligned}$$

donde $f(\mathbf{x})$ y $g_j(\mathbf{x})$ con $j = 1,\ldots,4$ son funciones convexas sobre la región convexa formada por las restricciones. Muestre que el mínimo local de $f(\mathbf{x})$ en esta región es su mínimo global (ayuda: Bosqueje una gráfica del problema).

Solución:

La figura 1.22 ilustra el problema de optimización convexa planteado.

1.6 Condiciones Necesarias y Suficientes para Máximos y Mínimos sin Restricciones

Para aplicar los conceptos matemáticos y técnicas numéricas necesarias de la teoría de optimización en problemas concretos de Ingeniería, es necesario definir previamente lo que se pretende optimizar. El enunciado general de un problema de programación matemática con restricciones podría ser:

Funciones de una Variable

Definición 1.20 Una función $f(\mathrm{x})$ definida sobre un conjunto $S \subset R$ tiene un *mínimo local* (*mínimo relativo*) en el valor $x^* \in S$, si existe un valor positivo δ tal que

$$\text{si} \quad \left|x - x^*\right| < \delta, \text{ entonces } f\left(x^*\right) < f\left(x\right)$$

Definición 1.21 La función $f(\mathrm{x})$ definida sobre un conjunto $S \subset R$ tiene un *mínimo global* en $x^* \in S$

si $f\left(x^*\right) < f\left(x\right)$ para todos los valores de $x \in S$.

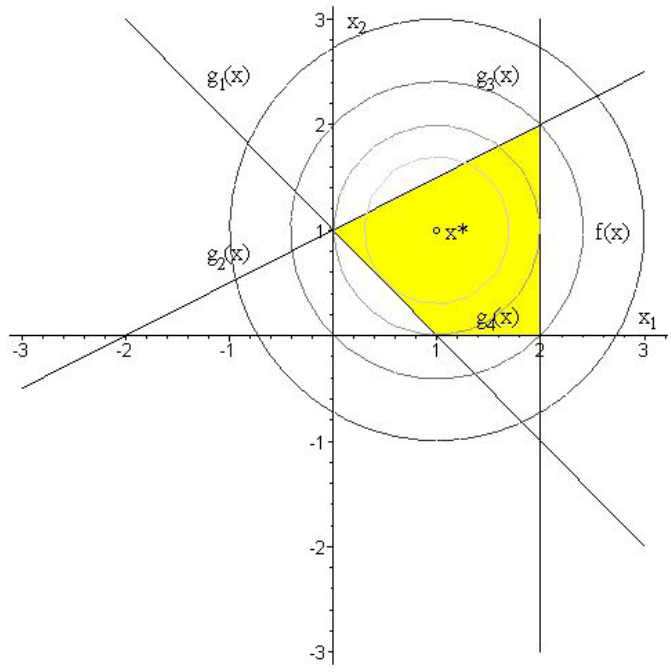


Figura 1.22 *El mínimo restringido coincide con el no restringido en el problema convexo.*

Definición 1.22 Un *punto estacionario* es un punto x^* en el cual

$$f'\left(x^*\right)=0 \tag{1.13}$$

Un *punto de inflexión* o *un punto de silla* es un punto estacionario que no corresponde a un óptimo local (mínimo o máximo). La figura 1.23 muestra la representación gráfica de una función $f(x)$ que tiene un mínimo local en x_4 y un mínimo global en x_2 dentro del intervalo [-2,4]. El enfoque clásico al problema de hallar los valores de los puntos estacionarios x_4 y x_2 es encontrar las ecuaciones que deberán satisfacerse para x_4 y x_2. La función $f(x)$ y su derivada representadas en la figura 1.23 son continuas y se ve que en ambos puntos, $x=x_4$ y $x=x_2$: $f'\left(x\right)=0$.

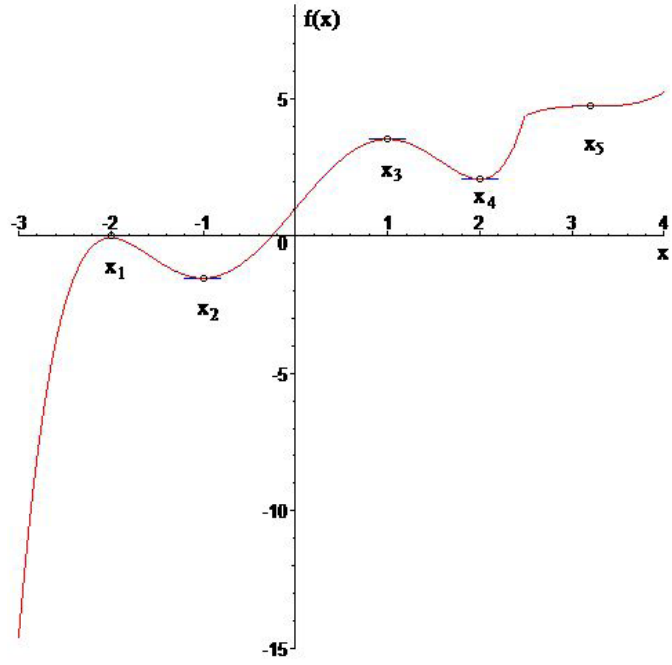


Figura 1.23 *Puntos estacionarios de la función* $f(x)$ *en [-2,4].*

Así, x_2 y x_4 son soluciones de la ecuación anterior. Los valores x_1 y x_3 son puntos en los cuales hay un máximo local y un máximo global respectivamente dentro del intervalo [-2,4], mientras que en x_5 se encuentra un punto de inflexión en el intervalo [3,4]; estos puntos también satisfacen la ecuación anterior. De esta manera, la ecuación (1.13) es sólo una *condición necesaria* para un mínimo, pero no es suficiente.

Sin embargo, obsérvese que en x_4 y $x_2, f'(x)$ cambia de signo negativo a positivo, en x_1 y x_3 el cambio de $f'(x)$ es de positivo a negativo, mientras que en x_5 la derivada no cambia de signo al pasar x a través de x_5. Así, en un mínimo la derivada es una función creciente, y ya que la rapidez de incremento de $f'(x)$ es medida por la segunda derivada, se esperaría que

$$f''(x) > 0 \quad \text{para} \quad x = x_4 \text{ ó } x_2$$

y

$$f''(x) < 0 \quad \text{para} \quad x = x_1 \text{ ó } x_3$$

Pero si la segunda derivada es cero, la situación permanece ambigua. Los resultados intuitivos anteriores pueden ponerse sobre una base firme considerando el desarrollo en serie de Taylor de $f(x)$ alrededor de $x = x^*$; esto por supuesto, considerando la continuidad de $f(x)$ y sus derivadas.

$$f(x) - f(x^*) = (x - x^*) f'(x^*) + \frac{1}{2}(x - x^*)^2 f''(x^*) + \cdots$$

Si x^* da un mínimo, el lado izquierdo es positivo para toda $(x - x^*)$ pequeña. $(|x - x^*| < \delta)$ y $f'(x^*)$ debe ser nula por la condición necesaria (1.13), ya que el próximo término implica un $(x - x^*)^2$ y vemos que si en verdad $f(x)$ tiene un mínimo en x^*, entonces

$$f''(x) > 0 \tag{1.14}$$

Por lo tanto, la ecuación (1.14) es una condición suficiente. Si

$$f'(x) = 0$$

y

$$f''(x) < 0 \tag{1.15}$$

entonces por argumentos similares x^* da un máximo (x_1 , $x_3 = x^*$). Se deberán comparar $f(x_4)$ y $f(x_2)$ para distinguir entre un mínimo local y un mínimo global. El caso ambiguo donde $f''(x)=0$ puede establecerse cuando se da, continuando la expansión en serie de Taylor.

$$f(x) - f(x^*) = (x - x^*) f'(x^*) + \frac{1}{2}(x - x^*)^2 f''(x^*) + \frac{1}{6}(x - x^*)^3 f'''(x^*) + \cdots$$

de ella podemos derivar el siguiente:

Teorema 1.1 Si $f(x)$ y sus derivadas son continuas, entonces x^* es un punto extremo (máximo o mínimo) si y solo si n es par, donde n es el orden de la derivada superior no nula que primero aparece evaluada en x^*.

Es decir, para n par

$$\begin{aligned} &si \quad f^n\left(x^*\right) > 0, \quad x^* \quad da\ un\ mínimo \\ &si \quad f^n\left(x^*\right) < 0, \quad x^* \quad da\ un\ máximo \end{aligned} \tag{1.16}$$

y para n impar

$$si \quad f^n\left(x^*\right) \neq 0, \quad x^* \quad da\ un\ punto\ de\ silla \tag{1.17}$$

Ejemplo 1.12 Halle el punto crítico y determine su naturaleza para la función

$$f(x) = (x-1)^6$$

Solución:

Derivando la función, igualando a cero y resolviendo para x se tiene

$$f'(x) = 6(x-1)^5 = 0 \quad \Rightarrow \quad x^* = 1$$

¿Lo anterior es máximo o mínimo? Para responder a esta pregunta se deriva otra vez la función y se evalúa en el punto crítico. Se tiene

$$f''(x) = 30(x-1)^4 \quad \Rightarrow \quad f''\left(x^*\right) = 0$$

En este momento sería erróneo concluir que se tiene un punto de inflexión; para asegurarse de que puede ser así, se tiene que seguir derivando la función hasta que ésta no se anule en el punto crítico; entonces

$$\begin{aligned} f^3(x) &= 120(x-1)^3 \quad \Rightarrow \quad f^3\left(x^*\right) = 0 \\ f^4(x) &= 360(x-1)^2 \quad \Rightarrow \quad f^4\left(x^*\right) = 0 \\ f^5(x) &= 720(x-1) \quad \Rightarrow \quad f^5\left(x^*\right) = 0 \\ f^6(x) &= 720 \quad \Rightarrow \quad f^6\left(x^*\right) = 6! > 0 \end{aligned}$$

Así, $f(x)$ tiene un mínimo cuando $x^* = 1$ porque $n = 6$ es par.

Funciones de varias variables

Definición 1.23 $f(\mathbf{x})$ tiene un *mínimo local* en $\mathbf{x}^*$ si existe una $\delta > 0$ tal que para $\|\mathbf{x} - \mathbf{x}^*\| < \delta$ entonces $f(\mathbf{x}) \geq f\left(\mathbf{x}^*\right)$.

Definición 1.24 $f(\mathbf{x})$ tiene un *mínimo global* en $\mathbf{x}^*$ si $f(\mathbf{x}) > f\left(\mathbf{x}^*\right)$ para todos los valores de $\mathbf{x}$.

Definición 1.25 Si $f(\mathbf{x}) \geq f\left(\mathbf{x}^*\right)$ se dice que $\mathbf{x}^*$ es un *mínimo débil*, si $f(\mathbf{x}) > f\left(\mathbf{x}^*\right)$ se dice que $\mathbf{x}^*$ es un *mínimo fuerte.*

Con estas definiciones y ciertas suposiciones de diferenciabilidad, se tiene el desarrollo en serie de Taylor para el caso de una función de n variables;

$$f(\mathbf{x}) - f(\mathbf{x}^*) = (\mathbf{x} - \mathbf{x}^*)^T \nabla f(\mathbf{x}^*) + \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T \mathbf{H}(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) + \cdots$$

Entonces, si $\mathbf{x}^*$ da un mínimo para $f(\mathbf{x})$, cada una de las primeras derivadas parciales $\partial f / \partial x_i \, (i = 1, \ldots, n)$ debe anularse en $\mathbf{x}^*$. Así, una *condición necesaria* para un mínimo en $\mathbf{x}^*$ es

$$\nabla f(\mathbf{x}^*) = \mathbf{0} \tag{1.18}$$

y entonces el signo de $f(\mathbf{x}) - f(\mathbf{x}^*)$ queda determinado por el de $(\mathbf{x} - \mathbf{x}^*)^T \mathbf{H}(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)$. Así, una *condición suficiente* para un mínimo en $\mathbf{x}^*$ es que

$$(\mathbf{x} - \mathbf{x}^*)^T \mathbf{H}(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \geq 0 \tag{1.19}$$

es decir, que $\mathbf{H}(\mathbf{x}^*)$ sea *positiva definida* o *positiva semidefinida*.

Para un máximo en $\mathbf{x}^*$ las condiciones *necesarias* y *suficientes* son:

$$\nabla f(\mathbf{x}^*) = \mathbf{0} \tag{1.20a}$$

$$(\mathbf{x} - \mathbf{x}^*)^T \mathbf{H}(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \leq 0 \tag{1.20b}$$

es decir, que $\mathbf{H}(\mathbf{x}^*)$ sea *negativa definida* o *negativa semidefinida*.

Conclusión:

Si $\mathbf{H}\left(\mathbf{x}^*\right)$ es positiva definida, $\mathbf{x}^*$ es un *mínimo fuerte.*

Si $\mathbf{H}\left(\mathbf{x}^*\right)$ es positiva semidefinida, $\mathbf{x}^*$ es un *mínimo débil.*

Si $\mathbf{H}\left(\mathbf{x}^*\right)$ es negativa definida, $\mathbf{x}^*$ es un *máximo fuerte.*

Si $\mathbf{H}\left(\mathbf{x}^*\right)$ es negativa semidefinida, $\mathbf{x}^*$ es un *máximo débil.*

Si $\mathbf{H}\left(\mathbf{x}^*\right)$ es indefinida, $\mathbf{x}^*$ es un *punto de silla.*

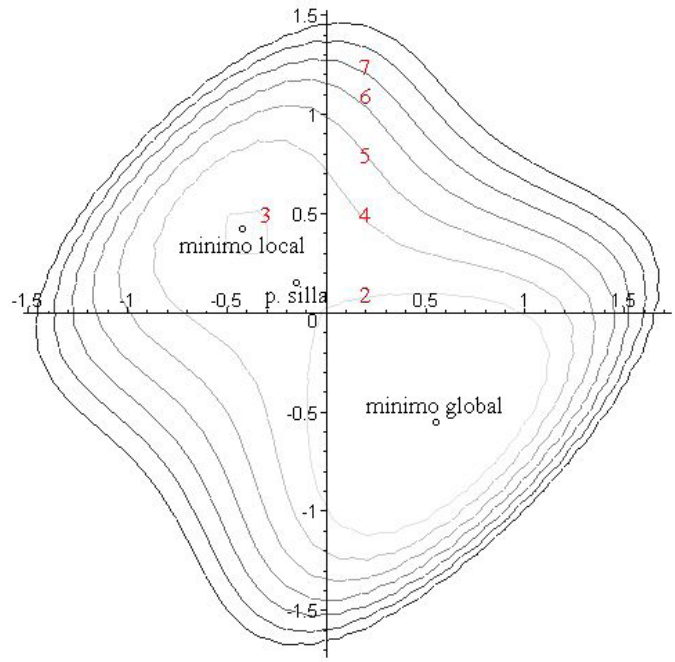


Figura 1.24 *Contornos de la función $f\left(\mathbf{x}\right)$ del ejemplo 1.13 con tres puntos estacionarios: Uno es punto de silla y los otros dos son mínimos.*

Ejemplo 1.13 Examine la naturaleza de los puntos estacionarios de la función

$$f(\mathbf{x}) = (x_2 - x_1)^4 + 8x_1x_2 - x_1 + x_2 + 3$$

Solución:

De la condición necesaria se obtienen los puntos extremos de la función; entonces

$$\nabla f(\mathbf{x}) = \begin{pmatrix} -4(x_2 - x_1)^3 + 8x_2 - 1 \\ 4(x_2 - x_1)^3 + 8x_1 + 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \mathbf{0}$$

y resolviendo el sistema de ecuaciones no lineales numéricamente se tienen tres soluciones dadas por

$$\mathbf{x}_1^* = \begin{pmatrix} 0.553 \\ -0.553 \end{pmatrix}, \quad \mathbf{x}_2^* = \begin{pmatrix} -0.418 \\ 0.418 \end{pmatrix} \text{ y } \mathbf{x}_3^* = \begin{pmatrix} -0.127 \\ 0.127 \end{pmatrix}$$

La matriz hessiana de la función es

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} 12(x_2 - x_1)^2 & -12(x_2 - x_1)^2 + 8 \\ -12(x_2 - x_1)^2 + 8 & 12(x_2 - x_1)^2 \end{pmatrix}$$

evaluada en los puntos críticos el lector puede verificar por los determinantes de los menores principales o los valores propios de las matrices que $\mathbf{x}_1^*$ y $\mathbf{x}_2^*$ son mínimos fuertes y que $\mathbf{x}_3$* es un punto de silla. Por último, evaluando en la función objetivo, se tiene

$$f\left(\mathbf{x}_1^*\right) = 0.943 \quad \text{y} \quad f\left(\mathbf{x}_2^*\right) = 2.926$$

de lo que se concluye que $\mathbf{x}_1^*$ es un mínimo fuerte global y $\mathbf{x}_2^*$ es un mínimo fuerte local (véase la Figura 1.24 para este problema).

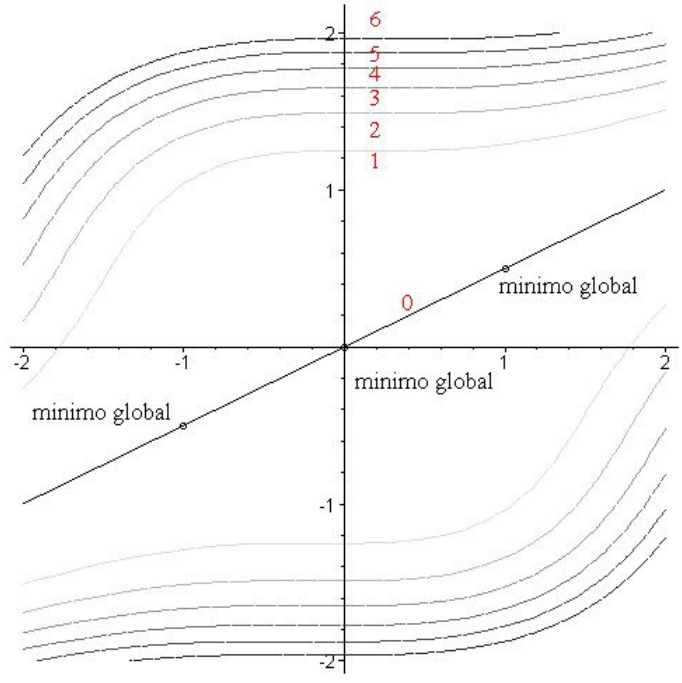


Figura 1.25 *Contornos de la función* $f\left(\mathbf{x}\right)$ *del ejemplo 1.14 con infinidad de puntos estacionarios, todos ellos extremos.*

Ejemplo 1.14 Examine la naturaleza de los puntos estacionarios de la función

$$f(\mathbf{x}) = 0.1\left(x_1^2 + x_1x_2 + x_2^2\right)\left(x_1 - 2x_2\right)^2$$

Solución:

De la condición necesaria se obtienen los puntos extremos de la función, y resolviendo el sistema de ecuaciones no lineales se tiene un número infinito de soluciones dadas por la ecuación $x_1 - 2x_2 = 0$; por ejemplo, tres soluciones serían

$$\mathbf{x}_1^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad , \quad \mathbf{x}_2^* = \begin{pmatrix} 1 \\ \frac{1}{2} \end{pmatrix} \quad \text{y} \quad \mathbf{x}_3^* = \begin{pmatrix} -1 \\ -\frac{1}{2} \end{pmatrix}$$

De la condición suficiente se puede determinar su naturaleza, en general, si se evalúa usando la solución $x_1 - 2x_2 = 0$, en términos de x_2 resulta

$$\mathbf{H}\left(\mathbf{x}^*\right) = x_2^2 \begin{pmatrix} 1.4 & -2.8 \\ -2.8 & 5.6 \end{pmatrix}$$

los determinantes de los menores principales de la matriz hessiana son $D_1 = 1.4x_2^2 \geq 0$ y $D_2 = 0x_2^2 = 0$ para todo x_2 por lo que, la matriz hessiana es positiva semidefinida y todos los puntos son mínimos globales débiles con el mismo valor funcional de $f\left(\mathbf{x}^*\right) = 0.0$ (véase la Figura 1.25).

1.7 Optimización Cuadrática sin Restricciones

Funciones de una variable

Considérese la función cuadrática general

$$f(x) = ax^2 + bx + c$$

Donde a, b y c son constantes. Obteniendo sus derivadas resulta

$$f'(x) = 2ax + b \text{ y } f''(x) = 2a$$

Para que x^* sea un valor extremo de la función $f(x)$, debe tenerse que

$$f'(x^*) = 0 \quad \Rightarrow \quad 2ax^* + b = 0$$
$$\therefore \quad x^* = -\frac{b}{2a}$$

y para que x^* sea un mínimo, entonces

$$f''(x^*) > 0 \quad \Rightarrow \quad 2a > 0 \quad \Leftrightarrow \quad a > 0$$

por lo tanto,

$$\boxed{x^* = -\frac{b}{2a}; \quad \begin{cases} es\ mínimo\ si\ a > 0 \\ es\ máximo\ si\ a < 0 \end{cases}} \tag{1.21}$$

Funciones de varias variables

Se minimizará la función cuadrática de n variables dada por

$$f(\mathbf{x}) = c + \mathbf{x}^T\mathbf{b} + \mathbf{x}^T\mathbf{A}\mathbf{x}$$

donde c es una constante, $\mathbf{b}$ es un vector constante y $\mathbf{A}$ es una matriz no simétrica de elementos constantes. Para que $\mathbf{x}^*$ sea un punto estacionario, debe satisfacerse

$$\nabla f\left(\mathbf{x}^*\right) = \mathbf{0}$$

entonces

$$\nabla f(\mathbf{x}) = \nabla c + \nabla\left(\mathbf{x}^T\mathbf{b}\right) + \nabla\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right)$$

donde

$$\nabla c = \mathbf{0} \quad , \quad \nabla\left(\mathbf{x}^T\mathbf{b}\right) = \mathbf{b} \quad \text{y} \quad \nabla\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right) = \left(\mathbf{A} + \mathbf{A}^T\right)\mathbf{x}$$

para $\mathbf{A}$ no simétrica, luego

$$\nabla f(\mathbf{x}) = \mathbf{b} + \left(\mathbf{A} + \mathbf{A}^T\right)\mathbf{x}$$

Evaluando en $\mathbf{x}^*$ el gradiente, igualando a cero y despejando a $\mathbf{x}^*$ se tiene que

$$\mathbf{x}^* = -\left(\mathbf{A} + \mathbf{A}^T\right)^{-1}\mathbf{b} \tag{1.22}$$

es el punto extremo. Se calcula ahora la matriz hessiana

$$\mathbf{H}(\mathbf{x}) = \nabla\left(\nabla^T f(\mathbf{x})\right) = \nabla\left(\mathbf{b}^T + \mathbf{x}^T\left(\mathbf{A} + \mathbf{A}^T\right)^T\right) = \mathbf{0} + \left(\mathbf{A} + \mathbf{A}^T\right) = \mathbf{A} + \mathbf{A}^T$$

estos resultados pueden verificarse usando la notación de índices. Evaluando en $\mathbf{x}^*$ se tiene que

$$\mathbf{H}\left(\mathbf{x}^*\right) = \mathbf{A} + \mathbf{A}^T \tag{1.23}$$

Así, el resultado es una matriz de elementos constantes; por otro lado,

$$\begin{aligned}\left(\mathbf{A} + \mathbf{A}^T\right)^{-1} &= \left(\mathbf{A} + \mathbf{A}^T\right)^{-1}\mathbf{I} = \left(\mathbf{A} + \mathbf{A}^T\right)^{-1}\mathbf{H}\mathbf{H}^{-1} \\ &= \left(\mathbf{A} + \mathbf{A}^T\right)^{-1}\left(\mathbf{A} + \mathbf{A}^T\right)\mathbf{H}^{-1} = \mathbf{I}\mathbf{H}^{-1} = \mathbf{H}^{-1}\end{aligned}$$

por lo que

$$\mathbf{H}^{-1} = \left(\mathbf{A} + \mathbf{A}^T\right)^{-1} \tag{1.24}$$

y sustituyendo en la ecuación 1.22, el resultado de $\mathbf{x}^*$ es

$$\mathbf{x}^* = -\mathbf{H}^{-1}\mathbf{b} \tag{1.25}$$

Se deja como ejercicio al lector mostrar que se llega al mismo resultado (Ec. 1.25), si la matriz $\mathbf{A}$ de la función cuadrática es simétrica.

Concluyendo:

Si $\mathbf{H}\left(\mathbf{x}^*\right)$ es positiva definida, también lo es $\left(\mathbf{A}+\mathbf{A}^T\right)$.

Si $\mathbf{H}\left(\mathbf{x}^*\right)$ es positiva definida, también lo es $\mathbf{H}^{-1}\left(\mathbf{x}^*\right)$ y en consecuencia $\left(\mathbf{A}+\mathbf{A}^T\right)^{-1}$ es positiva definida y $\mathbf{x}^*$ es un *mínimo fuerte.*

Si $\left(\mathbf{A}+\mathbf{A}^T\right)^{-1}$ es positiva semidefinida, $\mathbf{x}^*$ es un *mínimo débil.*

Si $\left(\mathbf{A}+\mathbf{A}^T\right)^{-1}$ es negativa definida o semidefinida entonces $\mathbf{x}^*$ es un *máximo fuerte* o *débil* respectivamente.

Finalmente, si $\left(\mathbf{A}+\mathbf{A}^T\right)^{-1}$ es indefinida entonces $\mathbf{x}^*$ es un *punto de silla.*

Ejemplo 1.15 Examine la naturaleza de las siguientes funciones

a) $f\left(\mathbf{x}\right)=x_1^2-x_1x_2+x_2^2$

b) $f\left(\mathbf{x}\right)=\frac{1}{2}x_1^2-2x_1x_2+2x_2^2$

c) $f\left(\mathbf{x}\right)=x_1^2-2x_1x_2-\frac{1}{2}\left(x_2^2-1\right)$

Solución:

Se deja como ejercicio al lector, mostrar que el problema del inciso a) tiene un mínimo fuerte como se muestra en la figura 1.26, el del inciso b) tiene un número infinito de mínimos débiles, figura 1.27 y en el inciso c) se tiene un punto de silla, figura 1.28. .

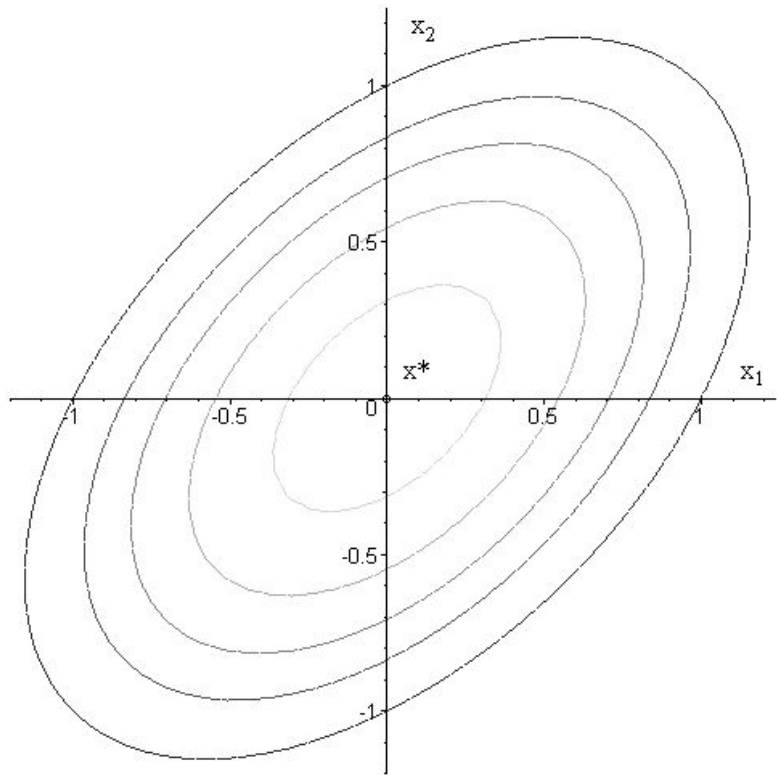


Figura 1.26 *Contornos de la función* $f(\mathbf{x}) = x_1^2 - x_1 x_2 + x_2^2$ *mostrando el mínimo fuerte.*

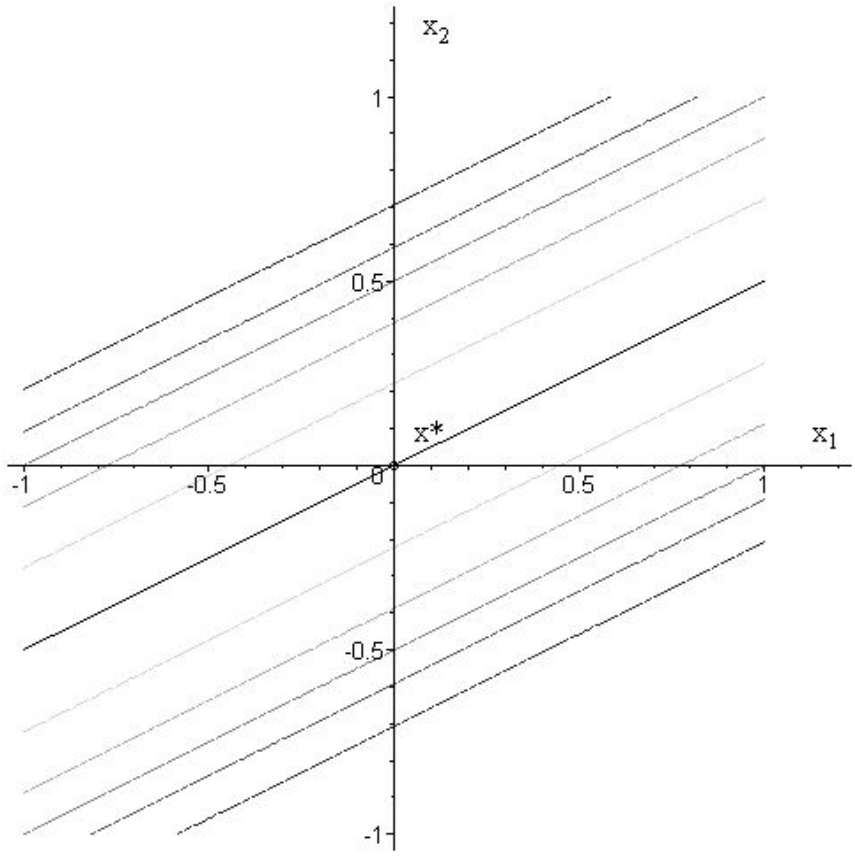


Figura 1.27 *Contornos de la función* $f(\mathbf{x}) = \frac{1}{2}x_1^2 - 2x_1x_2 + 2x_2^2$ *mostrando un mínimo débil, la línea que pasa por el origen contiene a todos los mínimos.*

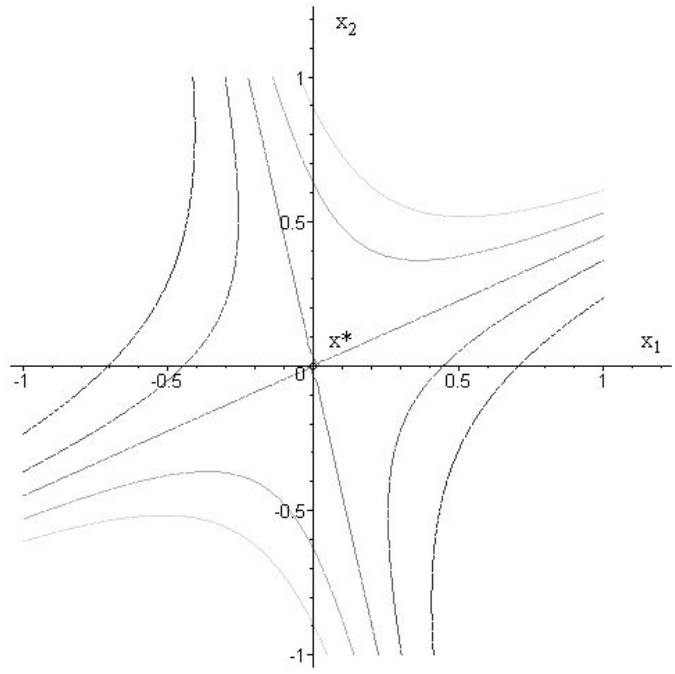


Figura 1.28 *Contornos de la función* $f(\mathbf{x}) = x_1^2 - 2x_1x_2 - \frac{1}{2}\left(x_2^2 - 1\right)$ *mostrando el punto de silla.*

Problemas

La estrategia para resolver problemas de optimización consiste en que deben asignarse símbolos a todas las magnitudes a determinar, establecer la función objetivo para la variable que debe optimizarse, reducir en la medida de lo posible la función objetivo a una ecuación con solo una variable independiente, esto puede exigir el uso de ecuaciones auxiliares o de restricción que relacionen las variables independientes de la función objetivo, y la determinación del dominio de la misma, para que el problema planteado tenga sentido, por último determinar el óptimo mediante las técnicas dadas.

1.1 Un cartel debe contener 300 cm^2 de texto impreso con márgenes superior e inferior de 6 cm y 4 cm en los laterales. Encuentre las dimensiones del cartel que minimizan el área total.

1.2 Una caja con base cuadrada y sin tapa debe retener 1000 cm^3. Determine las dimensiones que requieren el menor material para construir la caja.

1.3 Un granjero tiene 300 m de malla para cercar dos corrales rectangulares iguales y que comparten un lado de la cerca. Halle las dimensiones de los corrales para que el área cercada sea máxima.

1.4 Un terreno tiene la forma de un rectángulo con dos semicírculos en los extremos. Si el perímetro del terreno es de 50 m, determinar las dimensiones de manera que se tenga el área máxima.

1.5 Halle el volumen del cilindro circular recto más grande que puede inscribirse dentro de una esfera de radio R.

1.6 Se pide calcular el volumen máximo de un paquete rectangular enviado por correo, que posee una base cuadrada y cuya suma de la anchura más la altura más la longitud sea de 108 cm.

1.7 Se desea construir un recipiente cilíndrico de metal con tapa que tenga una superficie total de 100 cm^2. Encuentre las dimensiones de modo que tenga el mayor volumen posible.

1.8 Inscribir en una esfera de radio 1 m, un cilindro circular que tenga: a) volumen máximo y b) área lateral máxima. En ambos casos encuentre las dimensiones, radio de la base y altura.

1.9 Un alambre de 100 cm. de longitud, se corta en dos partes formando con una de ellas un círculo y con la otra un cuadrado. Cómo debe cortarse el alambre para que:

a) la suma de las áreas de las dos figuras sea máxima y

b) la suma de las áreas de las dos figuras sea mínima.

1.10 Dos pasillos de 2 y 3 m de ancho están unidos en ángulo recto. Encuentre la longitud de la barra recta más larga que puede pasarse horizontalmente de un pasillo a otro por una esquina.

1.11 El trabajo teórico para un compresor adiabático de dos etapas donde el gas se enfría a la temperatura de entrada entre las etapas está dado por

$$W = \frac{kp_1V_1}{k-1}\left[\left(\frac{p_2}{p_1}\right)^{(k-1)/k} - 2 + \left(\frac{p_3}{p_2}\right)^{(k-1)/k}\right]$$

donde $k = C_p / C_v$ = 1.4
p_1 = presión a la entrada = 1 atm
p_2 = presión de la etapa intermedia
p_3 = presión a la salida = 4 atm
V_1 = volumen de entrada

Se desea optimizar la presión intermedia de manera que el trabajo sea mínimo. Muestre que con los valores dados para p_1 y p_3 , p_2^{opt} = 2 atm.

1.12 La función objetivo para el requerimiento de trabajo de un compresor de tres etapas puede expresarse como (p es la presión)

$$f = \left(\frac{p_2}{p_1}\right)^{0.286} + \left(\frac{p_3}{p_2}\right)^{0.286} + \left(\frac{p_4}{p_3}\right)^{0.286}$$

donde p_1 = 1 atm y p_4 = 10 atm. El mínimo ocurre a una razón de presión de $\sqrt[3]{10}$ para cada etapa. ¿ f es convexa para 1 ≤ p_2 ≤ 10 y 1 ≤ p_3 ≤10?

CAPÍTULO 2

Métodos univariables

2.1 Introducción

Las funciones más simples con las que se inicia el estudio de los métodos de optimización no lineal son las funciones de una sola variable. Estas funciones suelen llamarse funciones univariadas o unidimensionales porque sólo dependen de una variable independiente. Aunque la minimización de funciones unidimensionales es en sí misma de importancia práctica, el área principal de aplicación de estas técnicas en el contexto de la optimización es su uso como una herramienta auxiliar para abordar de manera eficiente subproblemas de minimización multidimensionales.

2.2 Errores

En la práctica del cálculo numérico, es importante tener en cuenta que las soluciones calculadas a través de una computadora no son soluciones matemáticamente exactas. La precisión de una solución numérica puede verse disminuida por diversos factores, algunos de naturaleza sutil, por ejemplo los errores de truncamiento y redondeo, y la comprensión de estas dificultades puede guiarnos a menudo a desarrollar o a construir algoritmos numéricos adecuados.

Definición 2.1 Si $\mathbf{x}$ es una aproximación a $\mathbf{x}^*$. El error absoluto de la aproximación es

$$e = \left\| \mathbf{x} - \mathbf{x}^* \right\| \tag{2.1}$$

y el error relativo es:

$$r = \frac{\left\| \mathbf{x} - \mathbf{x}^* \right\|}{\left\| \mathbf{x}^* \right\|}, \quad si \quad \mathbf{x}^* \neq \mathbf{0}. \tag{2.2}$$

El error absoluto no es más que la distancia entre el valor exacto $\mathbf{x}^*$ y el valor aproximado $\mathbf{x}$, mientras que el error relativo mide el error entendido como una fracción del valor exacto. Por lo general, interesa el error absoluto y no el error relativo; pero cuando el valor exacto de una cantidad *es muy pequeño o muy grande*, los errores relativos son más significativos

2.3 Convergencia de los algoritmos de optimización

El diseño de los mejores algoritmos de minimización está de alguna manera asociado con la idea de la minimización eficiente de una función cuadrática con una matriz hessiana positiva definida. En la vecindad inmediata de un mínimo, muchas funciones pueden considerarse esencialmente cuadráticas por el concepto de convexidad ya comentado.

Rapidez de convergencia

Será importante tener una indicación teórica de la rapidez de convergencia hacia una solución. En adelante, el interés estará en ver cuál es el orden de convergencia de un método dado sin ahondar con profundidad en este tema, es decir, en este texto algunas relaciones de rapidez de convergencia se establecerán sin prueba.

Definición 2.2 Sea la sucesión $\{\mathbf{x}_k\}$ convergente a $\mathbf{x}^*$. Se denomina orden de convergencia de $\{\mathbf{x}_k\}$ como al máximo de los números positivos p que satisfagan que

$$0 \leq \lim_{k \to \infty} \frac{\left\| \mathbf{x}_{k+1} - \mathbf{x}^* \right\|}{\left\| \mathbf{x}_k - \mathbf{x}^* \right\|^p} = \beta < \infty \tag{2.3}$$

Aquí $\mathbf{x}_k$ es el punto alcanzado en la k-ésima iteración y $\mathbf{x}^*$ es el mínimo. La constante $\beta\left(0 \leq \beta < 1\right)$ se conoce como la *rapidez (tasa o razón)* de convergencia. Las tasas de convergencia rápidas están asociadas con valores grandes de p y valores pequeños de β. El caso $p = 1$ se denomina *rapidez de convergencia lineal* y la tasa de primer orden más rápida posible se denomina *superlineal* que es cuando $\beta = 0$ ó $\beta \to 0$. El caso $p = 2$ se denomina rapidez de convergencia *cuadrática*, $p = 3$ es *cúbica*, etc. Debe observarse que el valor de p depende del algoritmo, mientras que el valor de β depende de la función que está minimizándose.

Número de condición de la matriz $\mathbf{A}$

Definición 2. El número de condición de una matriz $\mathbf{A}$ se expresa como

$$\gamma = \left\|\mathbf{A}\right\|_i \left\|\mathbf{A}^{-1}\right\|_i \tag{2.4}$$

donde $i = \infty, 1, 2, e\left(de\ espectral\right)$

Una definición muy usual del número de condición está dada en términos de la norma espectral de la matriz $\mathbf{A}$; luego

$$\gamma = \left\|\mathbf{A}\right\|_e \left\|\mathbf{A}^{-1}\right\|_e = \lambda_{máx} \cdot \lambda'_{máx}$$

puede mostrarse (se deja al lector como ejercicio) que el valor propio más grande de $\mathbf{A}^{-1}$ es igual al recíproco del valor propio más pequeño de $\mathbf{A}$, es decir

$$\lambda'_{máx} = \frac{1}{\lambda_{mín}}$$

de manera que

$$\gamma = \frac{\lambda_{máx}}{\lambda_{mín}}$$

donde $\lambda_{máx} = \underset{i}{máx}\{\lambda_i\}$ y $\lambda_{mín} = \underset{i}{mín}\{\lambda_i\}$ son los valores propios más grande y más pequeño de la matriz $\mathbf{A}$. Si $\mathbf{A}$ representa a la matriz hessiana $\mathbf{H}$ de una función, entonces $\lambda_{máx}$ y $\lambda_{mín}$ son los valores propios máximo y mínimo de la matriz hessiana en el mínimo, y γ es un índice de la razón de la máxima a la mínima curvatura en dos direcciones en el mínimo. Para varios algoritmos, a mayor γ corresponde una mayor β y será más difícil realizar la minimización.

Una función objetivo con un valor bastante grande de γ causa problemas significativos y se dice que está *mal condicionada*, mientras que una función con un valor relativamente pequeño de γ se dice que está *bien escalada*. Por ejemplo, en un problema mal condicionado los contornos de la función alrededor del mínimo son elipses muy aplanadas, pero en un problema bien escalado los contornos son casi circulares.

Cuando γ es grande, las operaciones realizadas con la matriz hessiana o matrices relacionadas están sujetas a mayores errores de redondeo que cuando γ es pequeña. En el caso límite de una matriz singular, $\gamma \longrightarrow \infty$ debido a la presencia de un valor propio nulo. El número de condición es, por consiguiente, una medida de cuán bien se puede esperar la distinción de una solución aproximada *"buena"* de otra *"mala"*.

Ejemplo 2.1 ¿Cuál es el orden y tasa de convergencia de las siguientes sucesiones?

a) $x_k = 1 + \dfrac{1}{2^k}$

b) $x_k = 1 + \dfrac{1}{k!}$

Solución:

a) Sin pérdida de generalidad, veamos algunos puntos de la sucesión

$$x_0 = 2, \quad x_1 = \frac{3}{2} = 1.500, \quad x_2 = \frac{5}{4} = 1.250, \quad x_3 = \frac{9}{8} = 1.125, \quad \ldots$$

de manera que

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty}\left(1 + \frac{1}{2^k}\right) = \lim_{k\to\infty} 1 + \lim_{k\to\infty} \frac{1}{2^k} = 1 + 0 = 1$$
$$\therefore \quad x^* = 1$$

Usando la definición 2.2 y suponiendo que $p = \dfrac{1}{2}$, se tiene

$$\lim_{k\to\infty} \frac{\left|x_{k+1} - x^*\right|}{\left|x_k - x^*\right|^{1/2}} = \lim_{k\to\infty} \frac{\left|1 + \frac{1}{2^{k+1}} - 1\right|}{\left|1 + \frac{1}{2^k} - 1\right|^{1/2}} = \lim_{k\to\infty} \frac{2^{k/2}}{2^{k+1}} = \lim_{k\to\infty} \frac{2}{2^{k/2}} = 0 < \infty$$

Suponiendo que $p=1$, se tiene

$$\lim_{k\to\infty}\frac{\left|x_{k+1}-x^*\right|}{\left|x_k-x^*\right|}=\lim_{k\to\infty}\frac{\left|1+\frac{1}{2^{k+1}}-1\right|}{\left|1+\frac{1}{2^k}-1\right|}=\lim_{k\to\infty}\frac{2^k}{2^{k+1}}=\lim_{k\to\infty}\frac{1}{2}=\frac{1}{2}<\infty$$

y suponiendo ahora que $p=2$, se tiene

$$\lim_{k\to\infty}\frac{\left|x_{k+1}-x^*\right|}{\left|x_k-x^*\right|^2}=\lim_{k\to\infty}\frac{2^{2k}}{2^{k+1}}=\lim_{k\to\infty}\frac{2^k}{2}=\frac{\lim\limits_{k\to\infty}2^k}{\lim\limits_{k\to\infty}2}=\frac{\infty}{2}=\infty$$

en general,

$$\lim_{k\to\infty}\frac{|x_{k+1}-x^*|}{|x_k-x^*|^p}=\begin{cases}0, & p<1\\ \frac{1}{2}, & p=1\\ +\infty, & p>1\end{cases}$$

es decir, la sucesión dada converge linealmente $(p=1)$ hacia $x^*=1$ con una rapidez de convergencia de $\beta=\frac{1}{2}=0.5$.

b) Algunos puntos de la sucesión son

$$x_0=x_1=2,\quad x_2=\frac{3}{2}=1.500,\quad x_3=\frac{7}{6}=1.166,\quad x_4=\frac{25}{24}=1.041,\quad \ldots$$

de manera que

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty}\left(1+\frac{1}{k!}\right) = \lim_{k\to\infty} 1 + \lim_{k\to\infty}\frac{1}{k!} = 1+0=1$$
$$\therefore \quad x^* = 1$$

Usando la definición 2.2 y suponiendo que $p=1$, se tiene

$$\lim_{k\to\infty}\frac{\left|x_{k+1}-x^*\right|}{\left|x_k-x^*\right|} = \lim_{k\to\infty}\frac{\left|1+\frac{1}{(k+1)!}-1\right|}{\left|1+\frac{1}{k!}-1\right|} = \lim_{k\to\infty}\frac{k!}{(k+1)!} = \lim_{k\to\infty}\frac{1}{k+1} = 0 < \infty$$

Suponiendo ahora que $p=2$, se tiene

$$\lim_{k\to\infty}\frac{\left|x_{k+1}-x^*\right|}{\left|x_k-x^*\right|^2} = \lim_{k\to\infty}\frac{(k!)^2}{(k+1)!} = \lim_{k\to\infty}\frac{k!}{k+1} = \frac{\lim_{k\to\infty} k!}{\lim_{k\to\infty}(k+1)} \to \infty$$

es decir, la sucesión dada converge en forma superlineal $(p=1)$ a $x^*=1$ con una rapidez de convergencia de $\beta=0$.

Ejemplo 2.2 Dadas las siguientes funciones, determine el número de condición de la matriz hessiana asociada a éstas.

a) $f(\mathbf{x}) = x_1^2 + x_2^2$

b) $f(\mathbf{x}) = 100x_1^2 + x_2^2$

Solución:

a) La matriz hessiana de $f(\mathbf{x})$ es

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

cuyos valores propios son (recuerde la propiedad de una matriz diagonal) .

$$\lambda_1 = \lambda_2 = 2 = \lambda_{máx} = \lambda_{mín}$$
$$\therefore \quad \gamma = \frac{\lambda_{máx}}{\lambda_{mín}} = \frac{2}{2} = 1$$

b) La matriz hessiana para este caso es

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} 200 & 0 \\ 0 & 2 \end{pmatrix}$$

y sus valores propios son

$$\lambda_1 = \lambda_{máx} = 200 \quad \text{y} \quad \lambda_2 = \lambda_{mín} = 2$$
$$\therefore \quad \gamma = \frac{\lambda_{máx}}{\lambda_{mín}} = \frac{200}{2} = 100$$

De ambos resultados se concluye que, en general, la segunda función daría más problemas para determinar numéricamente su mínimo que para la primera función según la comparación de sus números de condición, el cual es mayor para la segunda de ellas. Además, debe tomarse en cuenta también que esto depende del punto inicial elegido, debido a que los contornos de la primera función son circulares, mientras que los de la segunda función son elípticos; el lector puede verificar esto como un ejercicio.

2.4 Criterios de Convergencia o Terminación

En las próximas secciones se analizarán diferentes algoritmos para resolver el problema de minimizar una función no restringida o restringida en un intervalo dado. Una parte crítica de todo el proceso es determinar cuándo detener la búsqueda hacia el punto óptimo. El criterio de terminación que se elige tiene un mayor efecto en la eficiencia y confiabilidad del proceso de optimización. A continuación se presentarán diferentes criterios de terminación de un proceso iterativo.

Número máximo de iteraciones

Un criterio de terminación que siempre debe usarse es el número máximo de iteraciones. Éste se aplica cuando el número de iteraciones n en un proceso iterativo excede a un número grande predeterminado $N_{máx}$; si esto sucede, entonces el proceso de búsqueda terminará. Esto asegura que si el proceso es extremadamente lento ya sea por dificultades numéricas o algorítmicas o bien debido a simples errores de programación, entonces el programa ya no continuará iterando en forma indefinida. Obsérvese que esta regla se aplica a cualquier proceso iterativo.

Cambio absoluto o relativo en los valores de las abscisas

El segundo criterio que puede usarse es una verificación del progreso de la optimización. Hay dos criterios que pueden usarse con respecto al cambio absoluto o relativo en los valores de las variables de interés, y estos criterios de convergencia se basan en la definición del error absoluto y el error relativo de orden n en un proceso iterativo, dados por las siguientes ecuaciones:

$$e_n = \left\| \mathbf{x}_n - \mathbf{x}^* \right\| \leq \varepsilon \quad \text{y} \quad r_n = \frac{\left\| \mathbf{x}_n - \mathbf{x}^* \right\|}{\left\| \mathbf{x}^* \right\|} \leq \varepsilon$$

donde $\mathbf{x}_n$ es el punto alcanzado en la *n-ésima* iteración, $\mathbf{x}^*$ es el mínimo y $\varepsilon > 0$ es la magnitud de una cantidad pequeña que suele denominarse *tolerancia* y que indica un límite en el tamaño que debe tener el error. Sin embargo, como no se conoce *a priori* el valor de $\mathbf{x}^*$, entonces se elige una estimación del error a través de la comparación entre dos iteraciones sucesivas en el valor de $\mathbf{x}$, como se establece a continuación:

$$e_n = \left\| \mathbf{x}_n - \mathbf{x}_{n-1} \right\| \leq \varepsilon \tag{2.5}$$

y

$$r_n = \frac{\left\| \mathbf{x}_n - \mathbf{x}_{n-1} \right\|}{\left\| \mathbf{x}_n \right\|} \leq \varepsilon \tag{2.6}$$

Cambio absoluto o relativo en el valor de la función objetivo

Un tercer criterio que puede utilizarse es el cambio absoluto o relativo en el valor de la función objetivo, y también en este caso se tienen dos criterios para indicar la convergencia. El primero es comparar el valor absoluto de $f\left(\mathbf{x}\right)$ sobre dos iteraciones sucesivas. En este caso, la convergencia se satisface si

$$\left| f\left(\mathbf{x}_n\right) - f\left(\mathbf{x}_{n-1}\right) \right| \leq \varepsilon \tag{2.7}$$

donde ε es una tolerancia predeterminada. Ésta puede ser constante, por ejemplo, $\varepsilon = 0.0001$, o puede ser una fracción del valor de la función objetivo en el punto $\mathbf{x}_0$, por ejemplo, $\varepsilon = 0.001\left| f\left(\mathbf{x}_0\right) \right|$, donde el valor absoluto se toma para considerar el posible valor negativo de la función objetivo.

El otro criterio que puede usarse es la verificación en el cambio relativo de $f(\mathbf{x})$ entre iteraciones sucesivas. Aquí la convergencia se satisface si

$$\frac{\left|f(\mathbf{x}_n) - f(\mathbf{x}_{n-1})\right|}{máx\left\{\left|f(\mathbf{x}_n)\right|, 10^{-10}\right\}} \leq \varepsilon \tag{2.8}$$

donde ε representa un cambio fraccional predeterminado, por ejemplo, $\varepsilon = 0.001$. El cambio absoluto se divide entre el máximo entre la magnitud de $f(\mathbf{x}_n)$ y 10^{-10}; con esto se asegura que no habrá división entre cero en el evento de que $f(\mathbf{x}_n) \to 0$.

Valor absoluto en el valor del gradiente de la función objetivo

El criterio de convergencia final que puede usarse es la verificación de las condiciones necesarias para la optimalidad. En el caso de la minimización no restringida o restringida en un intervalo y considerada en el presente capítulo, este criterio simplemente requiere que el gradiente de $f(\mathbf{x})$ se verifique para ver si cada componente está suficientemente cercana a cero para indicar que se alcanzó el mínimo. Este criterio se establece como sigue:

$$\left\|\nabla f(\mathbf{x}_n)\right\|_\infty \leq \varepsilon \tag{2.9}$$

Aquí la convergencia se satisface si cada componente de $\nabla f(\mathbf{x}_n)$ es menor o igual en magnitud que una constante predeterminada ε, por ejemplo, $\varepsilon = 10^{-4}$. Este criterio se emplea con facilidad cuando se usan métodos con derivada o gradiente. Cuando se usan métodos sin derivada o gradiente, es preferible utilizar los criterios antes vistos.

2.5 Métodos con derivadas

Ahora se considerarán algunos procedimientos numéricos simples que en forma indirecta localizan el mínimo de una función $f(x)$. Los métodos indirectos son aquellos que determinan los puntos extremos usando sólo *condiciones necesarias*, *derivadas analíticas* y *valores de la función*.

Con estos métodos, básicamente se busca el mínimo de $f(x)$ en un intervalo (a, b) en el cual se sospecha que se encuentra dicho mínimo o bien partiendo de un solo punto x_0 o dos de ellos x_0 y x_1 , los cuales se encuentran suficientemente cercanos al punto mínimo x^*. La búsqueda se hace determinando las raíces de la función

$$f'(x) = 0 \tag{2.10}$$

en puntos elegidos dentro del intervalo, o bien mediante un punto o puntos elegidos suficientemente cerca del óptimo ya que la idea es tratar de obtener el objetivo de la manera más eficiente posible. Una vez encontradas las raíces de la función $f'(x)$, se usa el criterio de la segunda derivada para determinar si estos valores extremos son máximos o mínimos. En caso de no poder determinar la segunda derivada o que ésta no exista, se utilizan puntos vecinos a los valores extremos y se comparan los valores de la función para determinar la naturaleza de los puntos extremos. Debe recordarse que además de la unimodalidad y continuidad en las funciones que se quieren optimizar, se requiere también la derivabilidad de las mismas. A continuación se verán algunos métodos numéricos simples para resolver la ecuación $f'(x) = 0$ y se comentará el error y/o la rapidez de convergencia de estos métodos.

Métodos que usan intervalos

En esta sección se analizarán los métodos que aprovechan el hecho de que una función $f'(x)$ cambia de signo en la vecindad de una raíz. Éstas técnicas se conocen como *métodos que usan intervalos* porque se necesita de dos valores iniciales para la búsqueda subsecuente de la raíz. Como su nombre lo indica, estos valores deben *encerrar* a la raíz. Los métodos descritos sobre este punto emplean diferentes estrategias para reducir sistemáticamente el tamaño del intervalo también conocido como *intervalo de incertidumbre* y así converger en la respuesta correcta. Se verá que en el desarrollo de los métodos que usan intervalos se supuso que en un principio se conoce el intervalo que contiene a la raíz.

En la práctica, dicho intervalo normalmente debe localizarse primero mediante cálculos preliminares. Entre los métodos simples disponibles para hacer esto se encuentran los *métodos gráficos* y los *métodos de búsqueda incremental*. Como preámbulo a los métodos que usan intervalos, examinaremos brevemente los métodos gráficos para obtener las graficas de las funciones y estimar sus raíces.

Métodos gráficos

Los métodos gráficos, además de su utilidad para determinar valores iniciales, también son útiles para visualizar las propiedades de las funciones y el comportamiento de los métodos numéricos.

Un método simple para obtener una aproximación a la raíz de la ecuación $f'(x) = 0$ consiste en graficar la función derivada y observar dónde cruza el eje x. Este punto, que representa el valor de x para el cual $f'(x) = 0$, proporciona una aproximación inicial de la raíz y, en consecuencia, del punto extremo de la función $f(x)$. Las interpretaciones geométricas, además de proporcionar aproximaciones iniciales de la raíz, son herramientas importantes en la asimilación de las propiedades de las funciones pues prevén las fallas de los métodos numéricos. Por ejemplo, las figuras 2.1 y 2.2 muestran algunas formas distintas en que la raíz de $f'(x) = 0$ puede encontrarse en un intervalo definido por un límite inferior a y un límite superior b. En general, si $f'(a)$ y $f'(b)$ tienen signos opuestos, esto indica que existe un número impar de raíces dentro del intervalo definido por los valores de a y b, pero si $f'(a)$ y $f'(b)$ tienen el mismo signo, esto indica que no hay raíces o que hay un número par de ellas entre los valores dados.

Métodos de búsqueda incremental

Además de verificar una respuesta individual, debe determinarse si se localizaron todas las raíces posibles. Como se mencionó antes, por lo general una gráfica de la función ayudará en esta tarea.

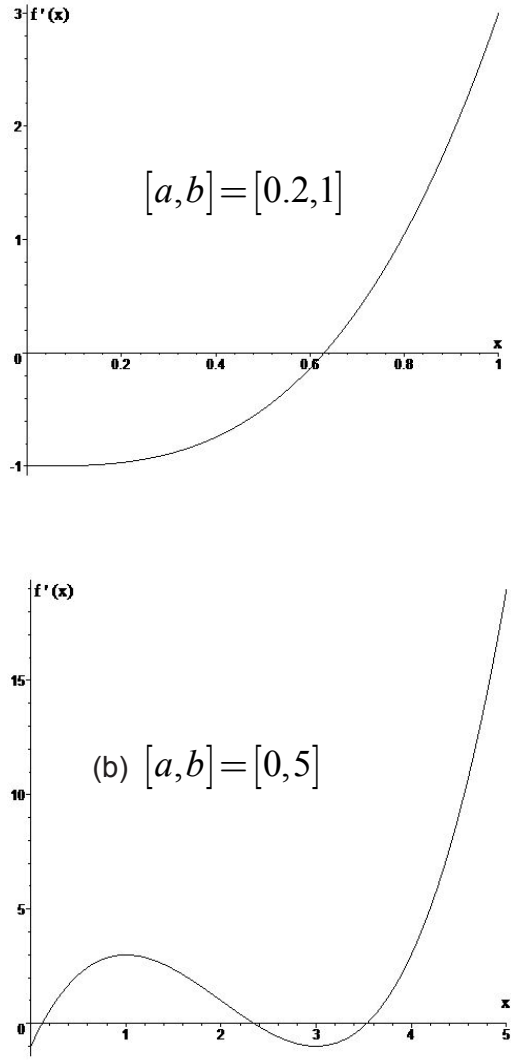


Figura 2.1 *En (a) y (b) existe un número impar de raíces si* $f'(a)$ *y* $f'(b)$ *tienen signos opuestos.*

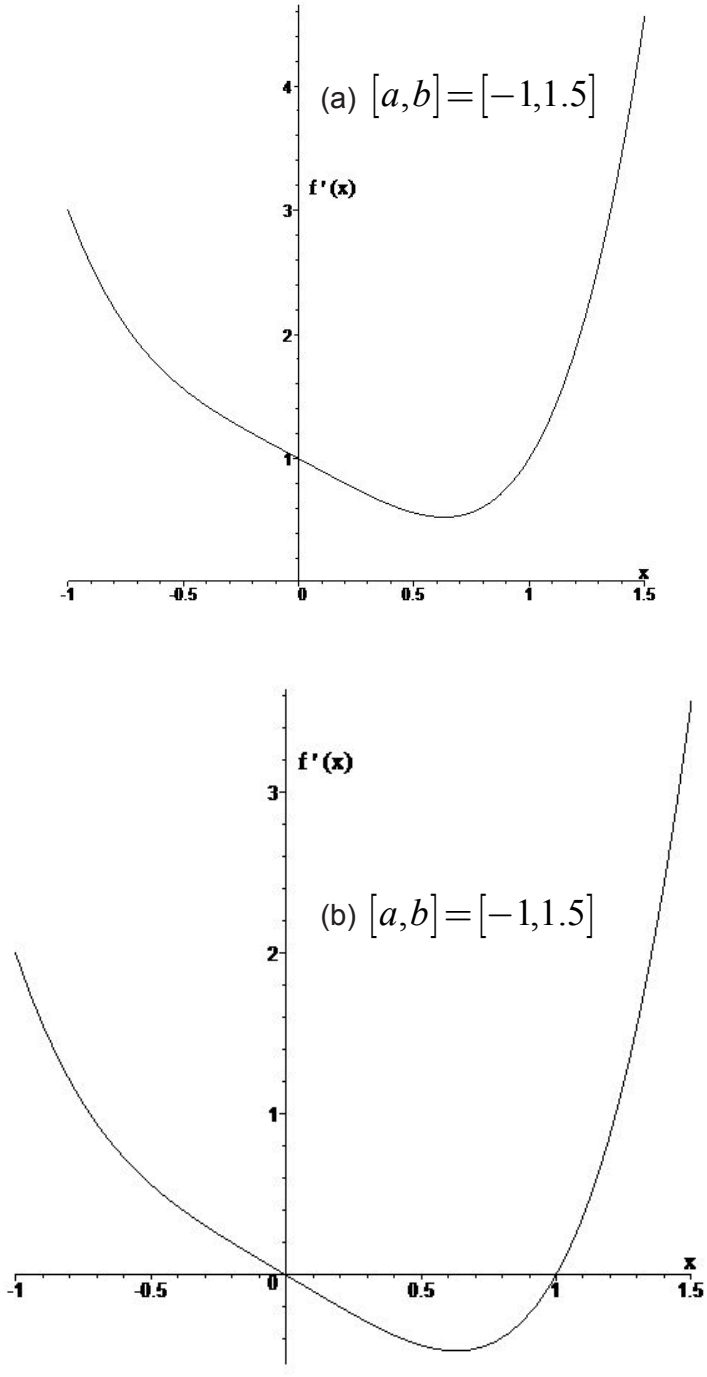


Figura 2.2 *(a) No hay o (b) existe un número par de raíces si* $f'(a)$ *y* $f'(b)$ *tienen signos iguales.*

Otra opción es incorporar una búsqueda incremental al inicio de la estrategia de solución. Esto consiste en empezar en una zona de la región de interés y realizar evaluaciones de la función con pequeños intervalos a lo largo de la región. Cuando la función cambia de signo, se supone que la raíz cae dentro del incremento. Los valores de x en los extremos del intervalo pueden servir de valores iniciales para una de las técnicas que usan intervalos y que se describirán más adelante.

Un problema en potencia aunado a los métodos de búsqueda incremental es escoger la longitud del incremento. Si la longitud es muy pequeña, la búsqueda puede consumir demasiado tiempo; por otro lado, si la longitud es muy grande, existe la posibilidad de que las raíces muy cercanas entre sí pasen inadvertidas. El problema se combina con la existencia de raíces múltiples. Aunque el empleo de un incremento muy fino puede solucionar en parte el problema de la acotación de las raíces, debe aclararse que los métodos sencillos, como el de la búsqueda incremental, no son infalibles. Se debe ser prudente al utilizar tales técnicas automáticas con cualquier otra información que provea la visualización en la localización de las raíces (puntos extremales).

Método de bisección

Cuando la primera derivada de la función objetivo esta disponible en los puntos extremos de un intervalo de incertidumbre, es necesario evaluar la función en sólo un punto interior para reducir este intervalo. Esto se debe a que es posible decidir si un intervalo acota un mínimo simplemente al observar los valores de la función f_a ,f_b y sus derivadas $f_a^{'}$, $f_b^{'}$ en los puntos extremos a , b .

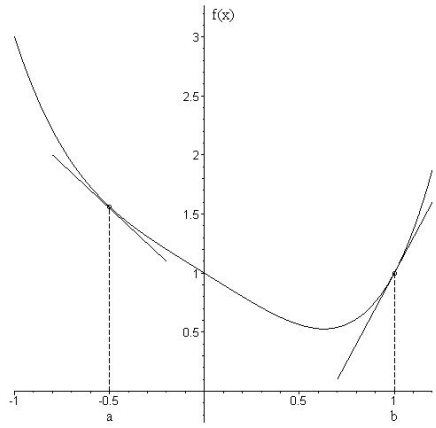


Figura 2.3 *Existe un mínimo si ambas pendientes en* a *y* b *satisfacen que:* $f_a^{'} < 0$ *y* $f_b^{'} > 0$

Las condiciones por satisfacer en los puntos a y b son que $f_a^{'} < 0$ y $f_b^{'} > 0$ como se ilustra en la figura 2.3, o que $f_a^{'} < 0$, $f_b^{'} < 0$ y $f_b > f_a$ como se muestra en la figura 2.4,

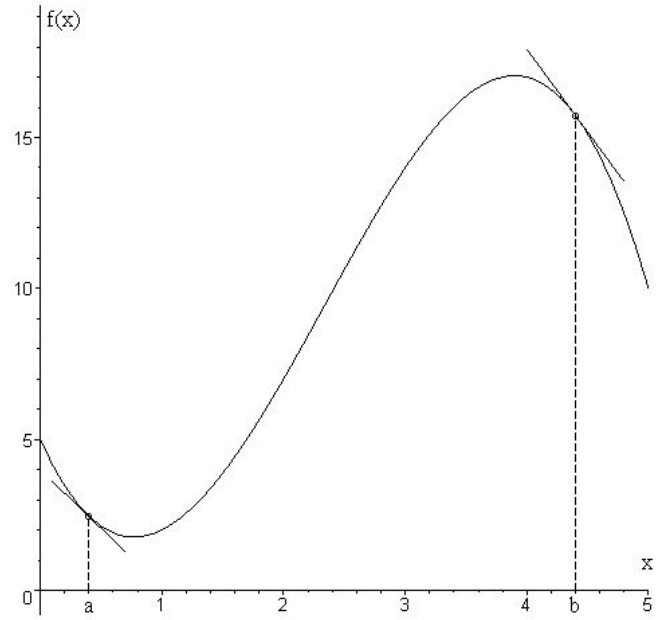


Figura 2.4 *Existe un mínimo si en los puntos* a *y* b *se satisface que:* $f_a^{'} < 0$, $f_b^{'} < 0$ *y* $f_b > f_a$.

o bien que $f_a^{'} > 0$, $f_b^{'} > 0$ y $f_b < f_a$ como se ilustra en la figura 2.5. El método de bisección parte de un intervalo inicial de incertidumbre $[a_0, b_0]$ en el cual se encuentra una de las raíces de la función $f'(x) = 0$; es decir, $f'(a_0)$ y $f'(b_0)$ tienen signos opuestos. El método consiste en evaluar la función $f'(x)$, continúa y derivable, en el punto medio del intervalo $[a_0, b_0]$, como se ilustra en la figura 2.6. Si $f'(a_0)$ y $f'(x_0)$ tienen signos opuestos, se reducirá el intervalo de $[a_0, b_0]$ a $[a_0, x_0]$; si no se satisface la condición de los signos opuestos, entonces se reducirá el intervalo de $[a_0, b_0]$ a $[x_0, b_0]$ ya que la raíz buscada se encuentra dentro de este nuevo intervalo. Se repite este proceso hasta lograr que el intervalo sea más pequeño que una tolerancia prefijada, y el último valor x_n será una buena aproximación de la raíz. Este procedimiento tiene la garantía de que converge a la raíz hasta una pre-

cisión deseada una vez que la raíz haya sido acotada y que la función sea unimodal en el intervalo de interés. Si la función no es unimodal, la condición $f'(a_0)f'(b_0) < 0$ se satisface siempre que el intervalo tenga un número impar de raíces, en este caso, el método de bisección encontrará una de las raíces separadas en el intervalo dado y tal vez será no deseada.

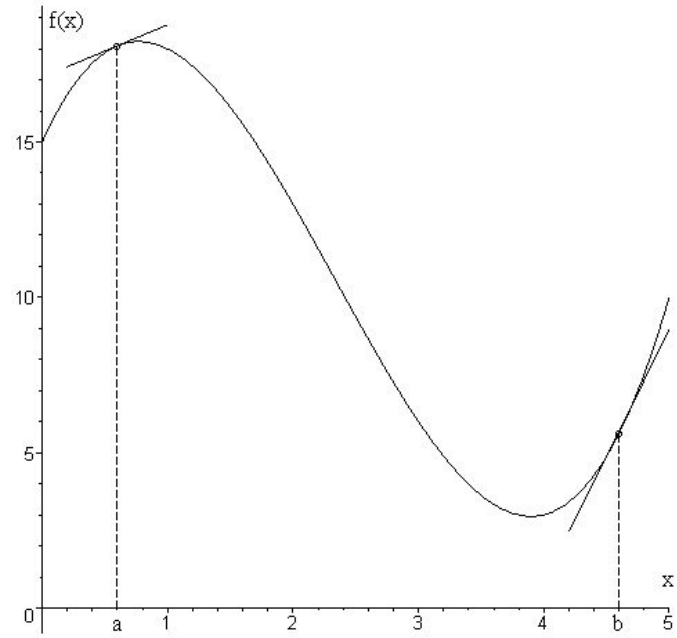


Figura 2.5 *Existe un mínimo si en los puntos* a *y* b *se satisface que:* $f_a > 0$, $f_b' > 0$ *y* $f_b < f_a$

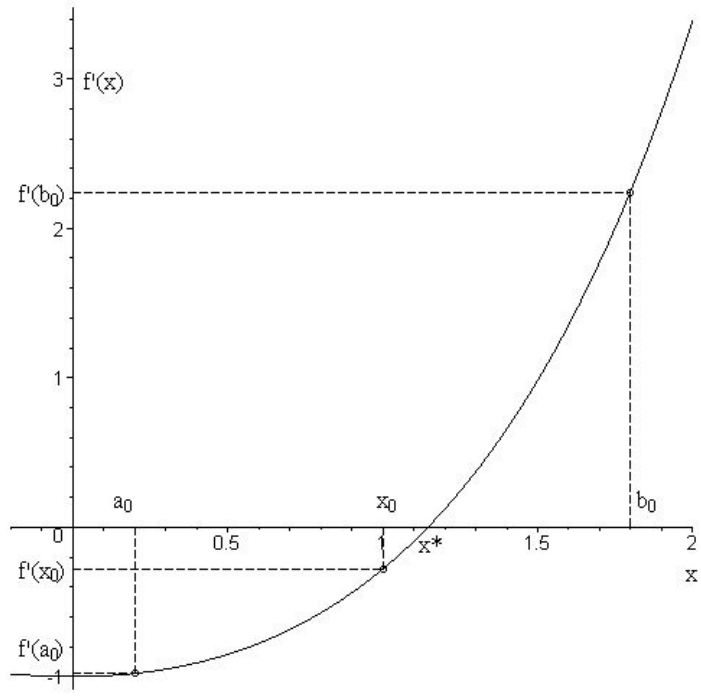


Figura 2.6 *El proceso de decisión en el método de bisección.*

El método de bisección también tiene problemas con las raíces dobles (o múltiples) debido a que estas funciones tocan el eje x de manera tangencial en estas raíces; otro defecto del método de bisección es que éste puede atrapar una singularidad como si fuera una raíz, debido a que dicho método no reconoce la diferencia entre una raíz y una singularidad (un *punto singular* es aquel en el que el valor de la función tiende a infinito, por ejemplo que la derivada no esta definida en el punto extremo de una función continua). Sin embargo, una gran ventaja es que también es útil para funciones no analíticas; en resumen, se puede decir que es un método robusto.

El algoritmo de bisección puede establecerse como sigue:

Algoritmo 2.1 Método de bisección:

Dados ε y una función $f'(x)$ continua sobre el intervalo $[a_0, b_0]$ tal que

$$f'(a_0) f'(b_0) < 0$$

Para $k = 0,1,2,...$ hasta donde se satisfaga, hacer:

1. $x_k = \dfrac{a_k + b_k}{2}$
2. Si $|b_k - x_k| \leq \varepsilon$ entonces $x_k \approx x^*$ y terminar.
3. Si $f'(a_k) f'(x_k) \leq 0$ entonces

   $a_{k+1} = a_k \quad b_{k+1} = x_k$

   En otro caso,

   $a_{k+1} = x_k \ , b_{k+1} = b_k$

   Fin del si
4. Regresar al paso 1.

Error límite en el método de bisección

Sean a_n, b_n y x_n los *n-ésimos* valores calculados a partir de a_0, b_0 y x respectivamente; entonces

$$b_n - a_n = \frac{1}{2}(b_{n-1} - a_{n-1})$$

luego

$$b_n - a_n = \frac{1}{2}(b_{n-1} - a_{n-1}) = \frac{1}{2^2}(b_{n-2} - a_{n-2}) = \frac{1}{2^3}(b_{n-3} - a_{n-3}) = \cdots$$

por lo que

$$b_n - a_n = \frac{1}{2^n}(b_0 - a_0), \quad n \geq 0 \tag{2.11}$$

donde $(b_0 - a_0)$ representa la longitud del intervalo original con el que se inició. Ya que la raíz x^* esta entre $[a_n, x_n]$ o $[x_n, b_n]$ se sabe que el error de orden n esta dado por

$$e_n = |x_n - x^*| \leq x_n - a_n = b_n - x_n = \frac{1}{2}(b_{n-1} - a_{n-1})$$

Combinando este resultado con la ecuación 2.11, se tiene

$$e_n = |x_n - x^*| \leq \frac{1}{2^n}(b_0 - a_0) \tag{2.12}$$

Esta ecuación expresa el error límite y muestra que x_n converge a x^* si $n \to \infty$ Para mostrar cuántas iteraciones serán necesarias para que x_n se aproxime a x^* con la precisión deseada, se necesita que

$$e_n \leq \varepsilon$$

donde $\varepsilon > 0$ es el criterio de tolerancia deseado; entonces

$$\frac{1}{2^n}(b_0 - a_0) \leq \varepsilon$$

por tanto,

$$n \geq \frac{\ln\left(\frac{b_0 - a_0}{\varepsilon}\right)}{\ln 2} \tag{2.13}$$

que expresa el número máximo de iteraciones para reducir el intervalo inicial a la precisión deseada.

Ejemplo 2.3 ¿Cuántas iteraciones se requieren para resolver un problema dónde el intervalo $[a_0, b_0] = [0,1]$ y la tolerancia $\varepsilon = 0.01$?

Solución:

Aplicando (2.13)

$$n \geq \frac{\ln\left(\frac{b_0 - a_0}{\varepsilon}\right)}{\ln 2} = 6.64$$

por lo que se requieren $n = 7$ iteraciones para reducir el intervalo original a menos o igual que 0.01 de precisión.

Rapidez de convergencia del método de bisección

Tomando el error máximo posible de orden n como

$$e_n = \left| x_n - x^* \right| = \frac{1}{2^n}(b_0 - a_0)$$

aplicando la ecuación (2.3), el método de bisección converge linealmente a x^* con rapidez de convergencia de $\beta = 0.5$. Estas predicciones teóricas se pueden verificar con los resultados de problemas resueltos y se dejan como ejercicio al lector.

Finalmente, otros métodos que usan intervalo son los métodos de *regula falsi* y *regula falsi modificada* por mencionar algunos y los cuales están fuera del alcance de este libro.

Métodos abiertos

En contraste con los métodos anteriores, los métodos abiertos se basan en fórmulas que requieren de un solo valor x o un par de ellos, pero que no necesariamente encierran la raíz. Como tales, algunas veces divergen y se alejan de la raíz x^* a medida que crece el número de iteraciones; sin embargo, cuando éstos convergen, en general lo hacen mucho más rápido que los métodos que usan intervalos, entre estos métodos se encuentra el método de la secante, Newton-Raphson, Newton modificado, Halley, Chebyshev, entre otros.

Método de Newton-Raphson

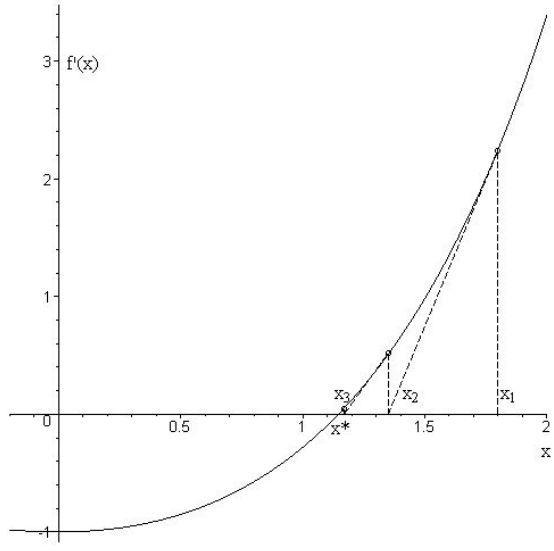
En esencia, la técnica de Newton-Raphson consiste en el uso de rectas tangentes; es por esto que al método se le conoce también como método de la tangente. Esta estrategia permite usar únicamente un punto inicial cercano a la raíz, sin requerir cosas adicionales dentro de la vecindad donde se encuentra.

Para la descripción de este método considérese las gráficas (a) y (b) de la figura 2.7. Por lo general, se inicia con una estimación de x^* que se denota con x_0; para mejorar esta estimación, considérese la línea recta que es tangente a la gráfica de $f'(x)$ en el punto $(x_0, f'(x_0))$; si x_0 esta cerca de x^*, esta línea tangente casi coincidirá con la gráfica de $f'(x)$ para puntos x alrededor de x^*. Entonces si x_1 es la raíz de la línea tangente, x_1 será casi igual a x^*. En general, para determinar una fórmula para x_{n+1}, considérese la pendiente de la línea tangente. Usando la derivada $f''(x)$ de $f'(x)$, se sabe del cálculo elemental que la pendiente de la tangente en $(x_n, f'(x_n))$ es $f''(x_n)$; esto guía a la ecuación

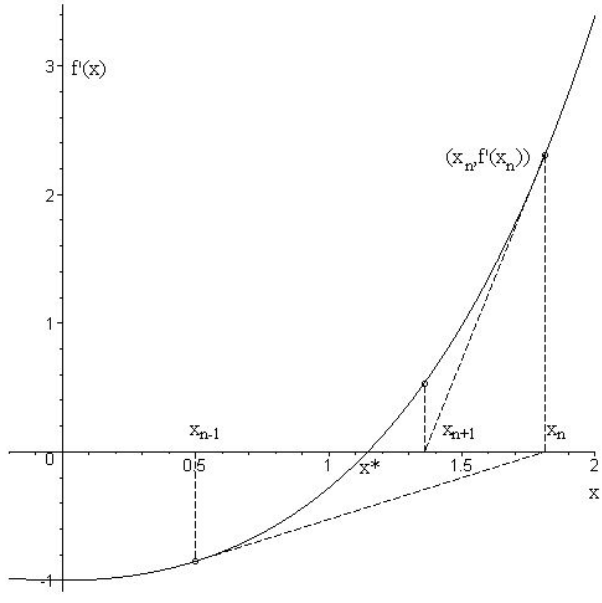
$$\tan\theta_n = f''(x_n) = \frac{0 - f'(x_n)}{x_{n+1} - x_n}$$

y , en consecuencia,

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)} \quad , \quad n = 0,1,2,... \tag{2.14}$$



(a) El punto inicial x_1 se encuentra a la derecha de x^*.



(b) El punto x_{n-1} se encuentra a la izquierda de x^*.

Figura 2.7 *El proceso iterativo en el método de Newton-Raphson.*

El método de Newton-Raphson utiliza en forma iterativa las rectas tangentes que pasan por las aproximaciones consecutivas de la raíz; además, este método requiere una buena estimación inicial, pues de otro modo la solución iterativa puede divergir o converger a una solución irrelevante. La razón de convergencia iterativa del método de Newton-Raphson es alta cuando funciona. La desventaja de este método se da en el caso de abordar raíces múltiples cuya convergencia suele ser lenta, o aún en el caso de raíces simples se puede tener problemas cuando $f''(x_n) \to 0$, lo que implica que la raíz puede estar cerca de un punto de inflexión. En resumen, se puede decir que la convergencia del método es muy dependiente de la naturaleza de la función y del valor de la aproximación inicial.

Algoritmo 2.2 Método de Newton-Raphson:

Dados ε , un punto inicial x_0 y una función $f'(x)$ continua y diferenciable. Para k = 0,1,2... hasta donde se satisfaga, hacer:

1. $x_{k+1} = x_k - \dfrac{f'(x_k)}{f''(x_k)}$

2. Si $\left|x_{k+1} - x_k\right| \leq \varepsilon$ o $\left|f'(x_{k+1})\right| \leq \varepsilon$ entonces $x_{k+1} \approx x^*$ y terminar.

3. Regresar al paso 1.

Rapidez de convergencia del método de Newton-Raphson

Suponiendo que $f'(x)$ tiene derivadas continuas para todo x en algún intervalo alrededor de la raíz x^*, por el teorema de Taylor podemos escribir

$$f'(x^*) = f'(x_n) + (x^* - x_n) f''(x_n) + \frac{1}{2}(x^* - x_n)^2 f'''(x_n) + \cdots$$

obsérvese que $f'\left(x^*\right)=0$ por suposición, y luego se divide por $f''\left(x_n\right)$ para obtener

$$0=\frac{f'\left(x_n\right)}{f''\left(x_n\right)}+\left(x^*-x_n\right)+\left(x^*-x_n\right)^2\frac{f'''\left(x_n\right)}{2f''\left(x_n\right)}+\cdots$$

Como

$$\frac{f'\left(x_n\right)}{f''\left(x_n\right)}=x_n-x_{n+1}$$

entonces

$$0=x_n-x_{n+1}+x^*-x_n+\left(x^*-x_n\right)^2\frac{f'''\left(x_n\right)}{2f''\left(x_n\right)}+\cdots$$

Resolviendo para $e_{n+1}=\left|x_{n+1}-x^*\right|$ se tiene

$$e_{n+1}=\left|x_{n+1}-x^*\right|=\left|\left(x_n-x^*\right)^2\frac{f'''\left(x_n\right)}{2f''\left(x_n\right)}+\cdots\right|\approx\left|\left(x_n-x^*\right)^2\frac{f'''\left(x_n\right)}{2f''\left(x_n\right)}\right|$$

donde $M=\left|\dfrac{f'''\left(x_n\right)}{2f''\left(x_n\right)}\right|$. Tomando el límite cuando $n\to\infty$, resulta

$$\lim_{n\to\infty}\frac{e_{n+1}}{e_n^2}=\lim_{n\to\infty}\frac{\left|x_{n+1}-x^*\right|}{\left|x_n-x^*\right|^2}=\left|\frac{f'''\left(x^*\right)}{2f''\left(x^*\right)}\right|=M^*$$

en consecuencia si $M^*<\infty$ el método de Newton-Raphson converge cuadráticamente a x^*.

2.6 Métodos sin derivada

Al resolver la ecuación no lineal encontrando la raíz, la meta fue determinar la variable x que diera *cero* en la función $f'(x)$. La *optimización de una sola variable* tiene como objetivo encontrar el valor de x que generará un *extremo*, ya sea éste un máximo o un mínimo. Como en la localización de raíces, la optimización en una dimensión sin derivadas se puede dividir en métodos de comparación de los valores de la función y en métodos de interpolación como se describirá en las próximas secciones. Por lo tanto, ahora la meta es encontrar el valor de la variable x que minimiza a $f(x)$ restringida a un intervalo.

2.7 Intervalo de incertidumbre

Considérese el problema de hallar x^* para

$$\begin{aligned} &Minimizar \quad f(x) \\ &sujeta\ a \quad a_0 \le x \le b_0 \end{aligned} \tag{2.15}$$

Debido a que la posición exacta del mínimo x^* de $f(x)$ sobre $[a_0, b_0]$ no se conoce, este intervalo se llama *intervalo de incertidumbr*e. Durante el procedimiento de búsqueda hacia el mínimo, pueden excluirse porciones de este intervalo que no contengan al mismo; entonces se dice que se ha reducido el intervalo de incertidumbre. En general, $[a_i, b_i]$ para $i = 0,...,n$, se conoce como intervalo de incertidumbre si un punto mínimo x^* cae en $[a_i, b_i]$ aunque su valor exacto no se conozca.

Teorema 2.1

Sea $f(x)$ una función estrictamente unimodal sobre el intervalo $[a,b]$. Sean x_1, $x_2 \in [a,b]$ de manera que $x_1 < x_2$.

Si $f(x_1) > f(x_2)$, entonces $f(z) \geq f(x_2)$ para todo $z \in [a,x_1]$, como se ilustra en la figura 2.8.

Si $f(x_1) \leq f(x_2)$, entonces $f(z) \geq f(x_1)$, para todo $z \in [x_2,b]$ como se ilustra en la figura 2.9.

Del teorema que se acaba de enunciar, bajo unimodalidad estricta se tiene que si $f(x_1) > f(x_2)$ el nuevo intervalo de incertidumbre es $[x_1,b]$ pero si $f(x_1) \leq f(x_2)$ entonces el nuevo intervalo de incertidumbre es $[a,x_2]$.

2.8 Razón de reducción y eficiencia

Una medida de la eficacia de un método de búsqueda por intervalos se conoce como razón de reducción, RR, propuesta por Wilde (1964), quién la definió como la razón del intervalo original de incertidumbre al intervalo final después de n ensayos de búsqueda o evaluaciones de la función.

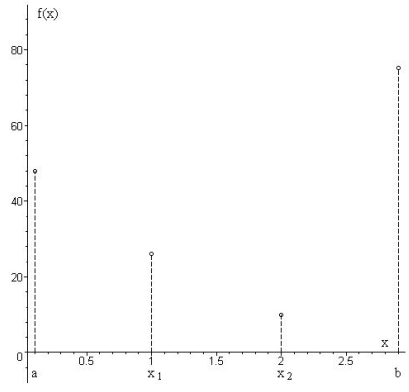


Figura 2.8 *El Intervalo de incertidumbre es* $[x_1,b]$

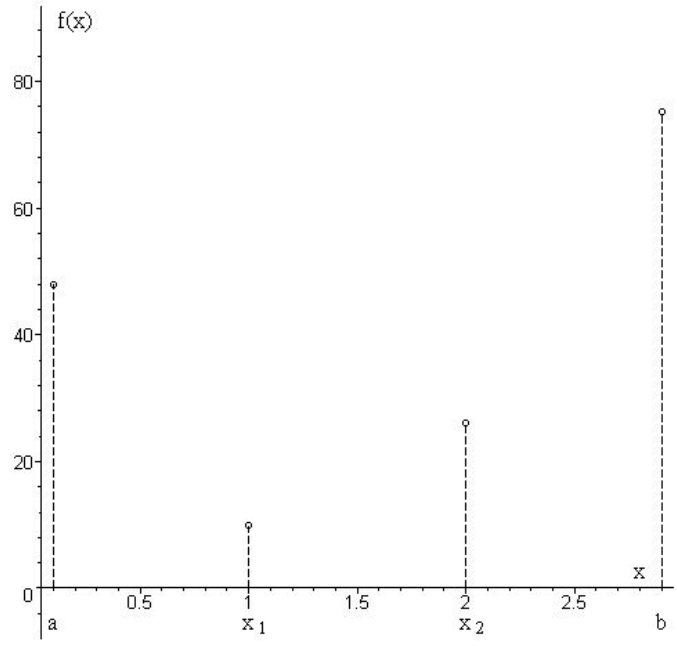


Figura 2.9 *El Intervalo de incertidumbre es* $[a, x_2]$

$$RR \equiv \frac{I_0}{I_n} = \frac{b_0 - a_0}{b_n - a_n} \tag{2.16}$$

donde I_0, I_n son los intervalos inicial y final de incertidumbre respectivamente, y a_i y b_i para $i = 0,...,n$ son los puntos extremos del i-ésimo intervalo. Como puede observarse, RR es una función monótonamente creciente con el número n de ensayos.

Wilde (1964) también definió la *eficiencia* η para n evaluaciones funcionales como el recíproco de RR

$$\eta \equiv \frac{1}{RR} = \frac{I_n}{I_0}$$

en este caso, $\eta \to 0$ como $n \to \infty$. (2.17)

De aquí en adelante solo se usará el concepto de eficiencia.

2.9 Métodos de comparación de la función objetivo

Se considerarán ahora algunos procedimientos numéricos que localizan en forma directa el mínimo (en general, extremos) de una función $f(x)$. Básicamente, con estos métodos la búsqueda se hace evaluando la función en puntos elegidos dentro de un intervalo $[a,b]$, donde se sabe que se encuentra el punto óptimo. La idea general es obtener nuestro objetivo de la manera más eficiente posible, es decir, con *el menor número de evaluaciones de la función*. Una característica general de los métodos de aproximación es que el punto preciso en el cual ocurre el óptimo nunca será conocido, y lo mejor que puede obtenerse es determinar el intervalo final de incertidumbre.

Por lo tanto, nuestro objetivo primordial es obtener el intervalo final de incertidumbre para establecer la eficiencia η del método aplicado. Este intervalo de incertidumbre siempre prevalece debido a que tales métodos calculan el valor de la función únicamente en valores discretos de las variables independientes. Para aplicar éstas técnicas, sólo se necesita conocer el intervalo inicial de incertidumbre $I_0 = b_0 - a_0$ y asegurarse de que $f(x)$ sea unimodal en el intervalo de interés. A continuación se verán brevemente dos métodos clásicos para establecer las ideas básicas de los métodos más eficientes de búsqueda univariada sin derivadas.

Método de la sección áurea

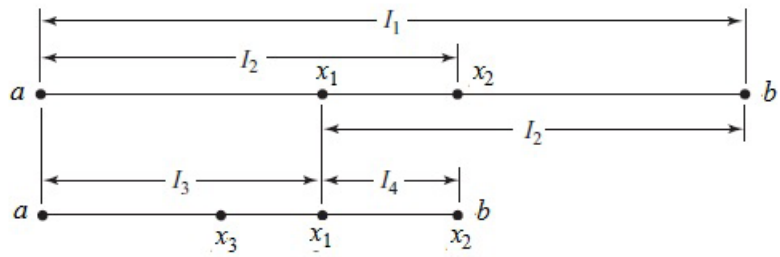


Figura 2.10 *El Intervalo de incertidumbre es* $[a, x_2]$, *si* $f(x_1) < f(x_2)$.

Refiriéndose a la figura 2.10 basada en la aplicación del teorema 2.1, el intervalo inicial de incertidumbre $I_1 = [a, b]$, es dividido por los puntos x_1 y x_2 en la primera evaluación, suponiendo que $f(x_1) < f(x_2)$, el nuevo intervalo de incertidumbre será $I_2 = [a, x_2]$, en la segunda evaluación se coloca el punto x_3 simétrico a x_1 con respecto a cualquiera de los extremos del intervalo $[a, x_2]$ como se muestra en la figura 2.10. Siguiendo esta estrategia con la aplicación del teorema 2.1 en cada evaluación subsecuente, se desea mantener

$$I_1 = I_2 + I_3$$

$$I_2 = I_3 + I_4$$

Cuando se han realizado j evaluaciones, entonces por el mismo razonamiento, se tiene

$$I_{j-1} = I_j + I_{j+1} \tag{2.18}$$

si además se impone la condición de mantener la razón de intervalos sucesivos constante, es decir,

$$\frac{I_1}{I_2} = \frac{I_2}{I_3} = \frac{I_3}{I_4} = \cdots = \tau$$

en general

$$\frac{I_{j-1}}{I_j} = \frac{I_j}{I_{j+1}} = \frac{I_{j+1}}{I_{j+2}} = \cdots = \tau \tag{2.19}$$

entonces, dividiendo la ecuación (2.18) por I_j se tiene

$$\frac{I_{j-1}}{I_j} = 1 + \frac{I_{j+1}}{I_j}$$

y de la ecuación (2.19),

$$\tau = 1 + \frac{1}{\tau} \quad \Rightarrow \quad \tau^2 - \tau - 1 = 0$$

Resolviendo la ecuación cuadrática se obtiene

$$\tau = \frac{1 \pm \sqrt{5}}{2}$$

y tomando la raíz positiva por razones obvias resulta

$$\tau = \frac{1 + \sqrt{5}}{2} \approx 1.618$$

la constante que representa a la razón de intervalos sucesivos.

El nombre de sección aurea viene de la ecuación (2.19); es decir, vemos que I_{j-1} se divide en dos partes, de manera que: la razón del intervalo total entre la mayor parte es igual a la razón de la mayor parte entre la menor parte, ésta es la llamada sección áurea de los antiguos griegos. Una vez determinado el valor de la razón τ, puede determinarse ahora el intervalo final de incertidumbre I_n; obsérvese que la razón

$$\frac{I_0}{I_1} = \tau$$
$$\frac{I_0}{I_2} = \frac{I_0}{I_1}\frac{I_1}{I_2} = \tau^2$$
$$\frac{I_0}{I_3} = \tau^3$$
$$\vdots$$
$$\frac{I_0}{I_n} = \tau^n$$

por lo tanto,

$$I_n = \tau^{-n} I_0 \tag{2.20}$$

y por consiguiente, la eficiencia será

$$\eta = \left(\frac{1}{\tau}\right)^n = \left(0.618\right)^n. \tag{2.21}$$

De esta manera, para reducir el intervalo final de incertidumbre al 1% del valor original son necesarias $n = 9.56 \approx 10$ iteraciones u 11 evaluaciones de la función. Para entender el mecanismo de este método, supóngase que se tiene el intervalo de incertidumbre $I_0 = (b_0 - a_0)$, como se muestra en la figura 2.11.
Si se consideran los resultados de las dos evaluaciones de la función, puede determinarse que intervalo se investigará posteriormente. Este intervalo contendrá uno de los puntos previos y el próximo punto será puesto de manera simétrica con respecto a éste, y así se continuará sucesivamente.

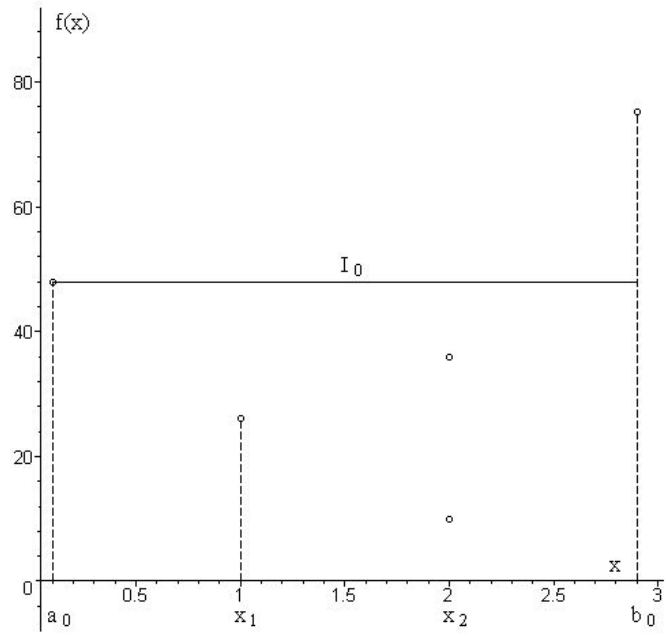


Figura 2.11 *Colocación de los primeros dos puntos* x_1 *y* x_2 *en* $[a_0, b_0]$.

Para iniciar, el primer punto se pone a una distancia

$$I_1 = \frac{1}{\tau} I_0$$

de un extremo y el segundo a la misma distancia del otro extremo. Se evalúan $f(x_1)$ y $f(x_2)$ y se tienen dos casos por considerar:

1) si $f(x_1) < f(x_2)$ el nuevo intervalo será (a, x_2), véase la figura 2.11 considerando el punto más alto para x_2.

2) si $f(x_1) > f(x_2)$ el nuevo intervalo será (x_1, b), véase la figura 2.11 considerando el punto más bajo para x_2.

Rapidez de convergencia

De la ecuación (2.20), la razón entre intervalos sucesivos en términos del intervalo inicial será

$$\frac{I_{k+1}}{I_k} = \frac{\tau^{-k-1} I_0}{\tau^{-k} I_0} = \frac{1}{\tau}$$

y, por lo tanto,

$$\lim_{k\to\infty} \frac{I_{k+1}}{I_k} = \lim_{k\to\infty} \frac{1}{\tau} = \frac{1}{\tau} = 0.618$$

Puede mostrarse sin dificultad que $\lim_{k\to\infty} \frac{I_{k+1}}{I_k^2} \to \infty$; por consiguiente, el método de la sección áurea converge linealmente con rapidez de convergencia $\beta \approx 0.618$.

Algoritmo 2.3 Método de la sección áurea:

Dada una función $f(x)$ continua sobre el intervalo $[a_1, b_1]$ y una tolerancia ε.

Calcular $c_1 = a_1 + (1-1/\tau)(b_1 - a_1)$, $d_1 = b_1 - (1-1/\tau)(b_1 - a_1)$, $f_c = f(c_1)$, $f_d = f(d_1)$

Para $k = 1,2,3,...$ hasta donde se satisfaga, hacer:

1. Si $f_c < f_d$, entonces

tomar $a_{k+1} = a_k$, $b_{k+1} = d_k$, $d_{k+1} = c_k$,

$$c_{k+1} = a_{k+1} + (1-1/\tau)(b_{k+1} - a_{k+1}),\ f_d = f_c,$$

$$f_c = f(c_{k+1})$$

En otro caso tomar

$$a_{k+1} = c_k,\ b_{k+1} = b_k,\ c_{k+1} = d_k,$$

$$d_{k+1} = b_{k+1} - (1-1/\tau)(b_{k+1} - a_{k+1}),\ f_c = f_d,$$

$$f_d = f(d_{k+1}).$$

Fin si

2. Si $|b_{k+1} - a_{k+1}| \leq \varepsilon$, entonces tomar

$x^* = c_{k+1}$ si $f_c < f_d$

o $x^* = d_{k+1}$ si $f_d < f_c$ y terminar.

Fin si

3. Tomar $k = k+1$ y regresar a 1.

Por último, existe una diversidad de métodos que comparan los valores de la función objetivo y que usan el concepto de intervalo de incertidumbre como los métodos de búsqueda uniforme, dicotómico secuencial, trisección, Fibonacci, etc.

2.10 Métodos de interpolación

En el método anterior se trató de determinar un intervalo pequeño en el cual se localizó el mínimo de la función. En el próximo método se adoptará un enfoque diferente. La clase de algoritmos conocidos como *métodos de interpolación polinomial* aproximan la función objetivo en un intervalo que se sabe acota un mínimo local por un polinomio cuadrático o cúbico que tienen los mismos valores de la función y quizás del gradiente en puntos particulares del intervalo. Luego, el mínimo del polinomio se utiliza para predecir el mínimo de la función objetivo y, de hecho, aquél reemplaza uno de los puntos previos guiando a un nuevo conjunto de puntos dispuestos en un intervalo menor al original, que posteriormente acotará el mínimo de la función.

Se comienza por estudiar un método que usa sólo valores de la función.

Método de interpolación cuadrática de Powell

Supóngase que se conocen los valores de una función $f(x)$ en tres puntos distintos a , b y c (los extremos y un punto del intervalo de incertidumbre) como $f_a = f(a)$, $f_b = f(b)$ y $f_c = f(c)$ respectivamente; entonces se puede aproximar $f(x)$ por una función cuadrática de la forma

$$p(x) = \alpha x^2 + \beta x + \gamma \tag{2.22}$$

donde α, β y γ están determinados por las ecuaciones

$$\begin{pmatrix} a^2 & a & 1 \\ b^2 & b & 1 \\ c^2 & c & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} p_a \\ p_b \\ p_c \end{pmatrix} = \begin{pmatrix} f_a \\ f_b \\ f_c \end{pmatrix} \tag{2.23}$$

En realidad no se necesita la forma explícita de $p(x)$ ya que sólo es de interés la posición de su mínimo $\hat{x}^*$. Resolviendo para α, β y γ se obtiene

$$\alpha = \frac{(c-b)f_a + (a-c)f_b + (b-a)f_c}{\Delta}$$

$$\beta = \frac{\left(b^2-c^2\right)f_a + \left(c^2-a^2\right)f_b + \left(a^2-b^2\right)f_c}{\Delta}$$

$$\gamma = \frac{bc(c-b)f_a + ac(a-c)f_b + ab(b-a)f_c}{\Delta} \qquad (2.24)$$

donde $\Delta = (a-b)(b-c)(c-a)$. Claramente, $p(x)$ tendrá un mínimo en

$$\hat{x}^* = -\frac{\beta}{2\alpha} \text{ si, } \alpha > 0$$

es decir,

$$p'(x) = 2\alpha x + \beta = 0 \text{ da } \hat{x}^* = -\frac{\beta}{2\alpha}$$

De este modo, la posición del mínimo de $f(x)$ se aproxima con

$$\hat{x}^* = \frac{1}{2}\frac{\left(b^2-c^2\right)f_a + \left(c^2-a^2\right)f_b + \left(a^2-b^2\right)f_c}{(b-c)f_a + (c-a)f_b + (a-b)f_c} \qquad (2.25)$$

Observe que $\hat{x}^*$ cae en $[a,b]$ si $f_c < f_a$ y $f_c < f_b$. La función objetivo se evalúa en $\hat{x}^*$ y uno de los puntos extremos a o b debe reemplazarse por $\hat{x}^*$ para empezar la próxima iteración de manera que los 3 nuevos puntos estén dispuestos sobre un intervalo que acote el mínimo de $f(x)$. La figura 2.12 muestra que no siempre el punto con el mayor valor de la función es el apropiado para descartarse.

Puede mostrarse que el orden de convergencia del método de interpolación cuadrática es de 1.3 (Luenberger, 1973). Esto es mejor que la convergencia superlineal pero no tan rápida como la de segundo orden. En el caso de que la tolerancia sea muy pequeña, entonces a, b, c y f_a, f_b, f_c, estarán muy cerca unos de otros y la ecuación (2.25) puede fallar.

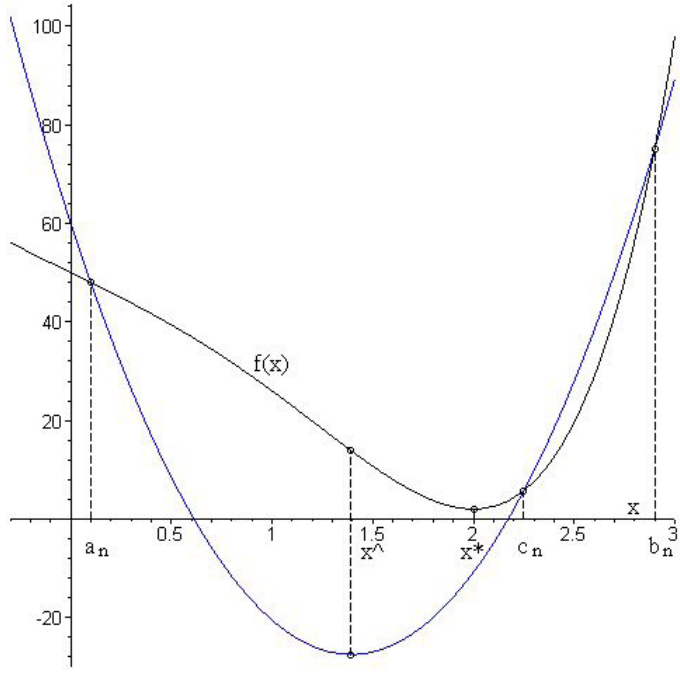


Figura 2.12a *Se descarta el punto con mayor valor de la función.*

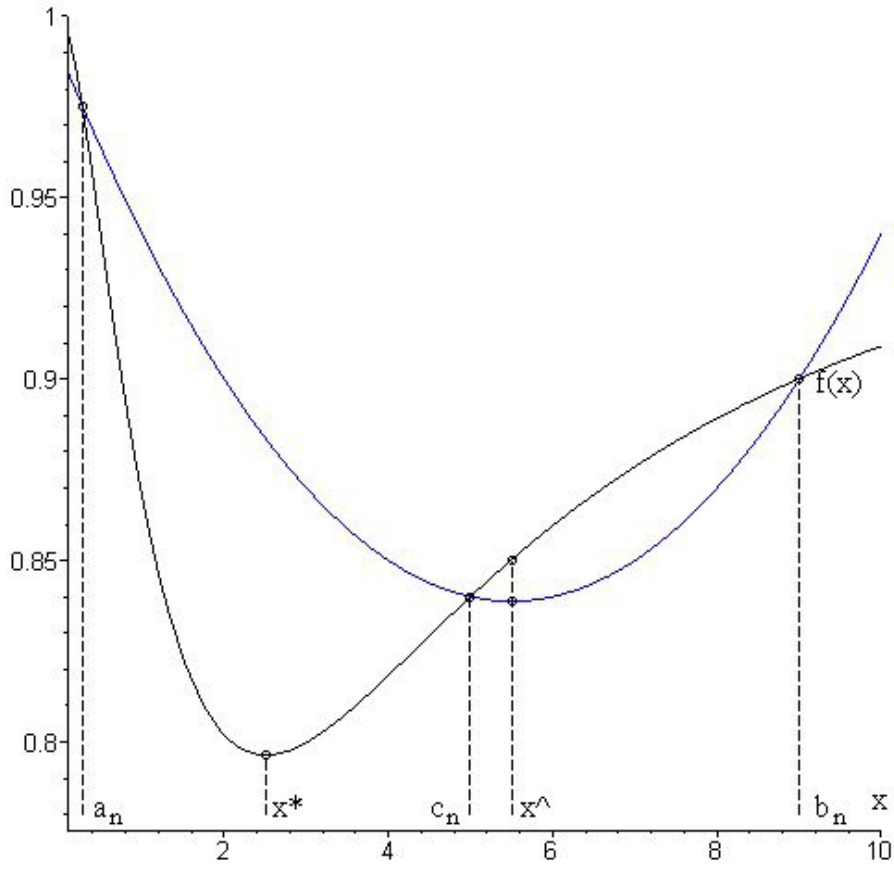


Figura 2.12b *No se descarta el punto con mayor valor de la función.*

Algoritmo 2.4 Método de interpolación cuadrática:

Dada una función $f(x)$ continua y los valores a_1, b_1, c_1, ε_x y ε_f.

Calcular $f_a = f(a_1)$, $f_b = f(b_1)$ y $f_c = f(c_1)$.

Para $k = 1,2,3,...$ hasta donde se satisfaga, hacer:

1. $\hat{x}^* = \frac{1}{2}\frac{\left(b_k^2 - c_k^2\right)f_a + \left(c_k^2 - a_k^2\right)f_b + \left(a_k^2 - b_k^2\right)f_c}{\left(b_k - c_k\right)f_a + \left(c_k - a_k\right)f_b + \left(a_k - b_k\right)f_c}$, $f_x = f\left(\hat{x}^*\right)$

2. Si $\hat{x}^* < c_k$ y $f_x < f_c$, entonces

   tomar $a_{k+1} = a_k$, $b_{k+1} = c_k$, $c_{k+1} = \hat{x}^*$, $f_b = f_c$, $f_c = f_x$.

   Pero si $\hat{x}^* > c_k$ y $f_x > f_c$, entonces

   tomar $a_{k+1} = a_k$, $b_{k+1} = \hat{x}^*$, $c_{k+1} = c_k$, $f_b = f_x$.

   Pero si $\hat{x}^* < c_k$ y $f_x > f_c$, entonces

   tomar $a_{k+1} = \hat{x}^*$, $b_{k+1} = b_k$, $c_{k+1} = c_k$, $f_a = f_x$.

   En otro caso,

   tomar $a_{k+1} = c_k$, $b_{k+1} = b_k$, $c_{k+1} = \hat{x}^*$, $f_a = f_c$, $f_c = f_x$.

   Fin si

3. Si $\left|b_{k+1} - a_{k+1}\right| \le \varepsilon_x$ o $\left[\frac{f(c_k) - f(c_{k+1})}{f(c_k)}\right] \le \varepsilon_f$, entonces

   tomar $x^* = c_{k+1}$ y terminar.

   Fin si

4. Tomar $k = k + 1$ y regresar a 1.

Finalmente, se pueden estudiar otros métodos de interpolación de mayor orden, por ejemplo, para usar una interpolación cúbica se pueden desarrollar dos métodos alternativos, el primero usando valores de la función objetivo en donde se requiere de cuatro puntos establecidos en el intervalo de incertidumbre, y el segundo, usando dos valores de la función objetivo y dos valores de la derivada de la función, este último método se conoce como método de interpolación cúbica de Davidon y éste tiene un orden de convergencia cuadrático.

Problemas

Para cada uno de los siguientes problemas encuentre la solución numéricamente con cualquiera de los métodos propuestos y compárela con el valor de la solución exacta (analítica si existe).

2.1 El alcance R de un proyectil lanzado con velocidad inicial v_0 y con un ángulo θ respecto de la horizontal es $R(\theta) = \left(v_0^2 sen2\theta\right)/g$, donde g=9.81 m/s^2 es la aceleración de la gravedad. Encuentre el ángulo θ que produce alcance máximo, si v_0 = 40 m/s.

2.2 Se lanza un cuerpo hacia arriba con velocidad inicial 40 m/s, ¿Calcule cuál es la máxima altura que alcanzará si la aceleración gravitacional g = 10 m/s? La ecuación que describe la altura en función del tiempo es:

$$h(t) = vt - \frac{g}{2}t^2$$

2.3 La energía potencial V (r) de un gas, esta dada por la expresión del potencial de Lennard-Jones,

$$V(r) = 4 \in \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

donde $\in, \sigma$ son constantes, tales que $\in$, σ> 0.

Para argón $\mathrm{A} = 6.19 \text{ x } 10^{-3} \text{kJ} (\mathring{\mathrm{A}})^6 / mol = 4\varepsilon\sigma^6$ y

$$\mathrm{B} = 9.58 \text{ x } 10^{-6} \text{kJ } (\mathring{\mathrm{A}})^{12} / mol = 4\varepsilon\sigma^{12}$$

a) ¿Tiene el potencial de Lennard-Jones $V(r)$ un punto o puntos estacionarios? Si es así, localícelo(s).

b) Identifique la naturaleza del punto o puntos estacionarios (mínimo, máximo, etc.)

c) ¿Cuál es la magnitud de la energía potencial en los puntos estacionarios?

2.4 La trayectoria de una pelota se calcula por medio de la ecuación

$$y = y_0 + \left(\tan\theta_0\right)x - \frac{g}{2v_0\cos^2\theta_0}x^2$$

donde y = altura (m), θ_0 = ángulo inicial (radianes), v_0 = velocidad inicial (m/s), g = constante gravitacional = 9.81 m/s^2, y y_0 = altura inicial (m). Use el método de la búsqueda de la sección dorada para determinar la altura máxima dado que y_0 = 1 m, v_0= 25 m/s y θ_0 = 50º. Haga iteraciones hasta que el error aproximado esté por debajo del 1%, con el uso de valores iniciales de a = 0 y b = 60 m.

2.5 Hay que separar una mezcla de benceno y tolueno en un reactor flash. ¿A qué temperatura deberá operarse el reactor para obtener la mayor pureza de tolueno en la fase líquida (maximizar x_T)? La presión en el reactor es de 800 mm Hg. Las unidades en la ecuación de Antoine son mm Hg y °C para presión y temperatura, respectivamente.

$$x_B P_{sat,B} + x_T P_{sat,T} = P$$

$$\log_{10}\left(P_{sat,B}\right) = 6.905 - \frac{1211}{T + 221}$$

$$\log_{10}\left(P_{sat,T}\right) = 6.953 - \frac{1344}{T + 219}$$

2.6 A se convertirá en B en un reactor con agitación. El producto B y la sustancia sin reaccionar A se purifican en una unidad de separación. La sustancia A que no entró en la reacción se recicla al reactor. Un ingeniero de procesos ha encontrado que el costo inicial del sistema es una función de la conversión, x. Encuentre la conversión que dará el sistema de menor costo. C es una constante de proporcionalidad.

$$Costo = C\left[\left(\frac{1}{(1-x)}\right)^{0.6} + 6\left(\frac{1}{x}\right)^{0.6}\right]$$

CAPÍTULO 3

Métodos multivariables

3.1 Introducción

En este capítulo se considerará el problema de minimizar una función $f(\mathbf{x})$ de varias variables. Los métodos que se describirán más adelante procederán en la siguiente forma: dado un vector $\mathbf{x}$, se determina un vector de dirección $\mathbf{d}$ elegido en forma adecuada; luego $f(\mathbf{x})$ se minimiza desde $\mathbf{x}$ en la dirección de $\mathbf{d}$ por una de las técnicas ya revisadas èn el capítulo anterior, lo cual se conoce como búsqueda lineal. En lo que sigue del resto de este texto, se requerirá resolver un problema de búsqueda lineal de la forma

$$\begin{aligned} \textit{Minimizar} \quad & f(\mathbf{x}+\alpha\mathbf{d}) \\ \textit{sujeta a} \quad & \alpha \in I \end{aligned} \tag{3.1}$$

donde $I=\mathbb{R}$, ó $I=\{\alpha:\alpha\geq 0\}$, ó $I=\{\alpha:a\leq\alpha\leq b\}$, con el fin de evitar los métodos con tamaño de paso discreto como inicialmente fueron diseñados, ya que presentan una convergencia muy lenta.

Ejemplo 3.1 Dada la función $f(\mathbf{x})=(x_1-1)^2+(x_2-2)^2+(x_3-3)^2$ y el punto $\mathbf{x}=(4,3,2)^T$, encuentre α^* para minimizar $f(\mathbf{x}+\alpha\mathbf{d})$ en la dirección $\mathbf{d}=(1,1,1)^T$.

Solución:

Se tiene que

$$\mathbf{x}+\alpha\mathbf{d}=\begin{pmatrix}4\\3\\2\end{pmatrix}+\alpha\begin{pmatrix}1\\1\\1\end{pmatrix}=\begin{pmatrix}4+\alpha\\3+\alpha\\2+\alpha\end{pmatrix}$$

y sustituyendo en la función objetivo resulta

$$\begin{aligned}f\left(\mathbf{x}+\alpha\mathbf{d}\right)&=\left(4+\alpha-1\right)^2+\left(3+\alpha-2\right)^2+\left(2+\alpha-3\right)^2\\&=\left(\alpha+3\right)^2+\left(\alpha+1\right)^2+\left(\alpha-1\right)^2=3\alpha^2+6\alpha+11=\phi\left(\alpha\right)\end{aligned}$$

derivando $f\left(\mathbf{x}+\alpha\mathbf{d}\right)$ con respecto a α se tiene.

$$\phi'\left(\alpha\right)=6\alpha+6=0\quad\Rightarrow\quad\alpha^*=-1$$

Por lo tanto, $f\left(\mathbf{x}+\alpha^*\mathbf{d}\right)=8$ es el valor mínimo; cualquier otro valor de α incrementa el valor de $f\left(\mathbf{x}+\alpha\mathbf{d}\right)$.

La función $f\left(\mathbf{x}+\alpha\mathbf{d}\right)=\phi\left(\alpha\right)$ es una función de una sola variable y puede optimizarse numéricamente con cualquier técnica vista en los capítulos anteriores, sobre todo cuando no se puede resolver en forma analítica el problema si la función es altamente no lineal.

3.2 Método de comparación de la función objetivo

Se ha dedicado mucho esfuerzo en desarrollar métodos de búsqueda directos para localizar el mínimo de una función de n variables. Recuérdese que un método directo de búsqueda es aquel que usa únicamente valores de la función.

En esta sección, se considerará con detalle sólo un método; la experiencia ha mostrado que este método es robusto y capaz de aplicarse a una gran variedad de problemas de pequeña escala. Para éste método se han construido un gran número de funciones que, dada su naturaleza, proveen severas pruebas para el mismo. Ejemplos de algunas de ellas son:

1). La función de Rosenbrock: $f(\mathbf{x}) = 100\left(x_2 - x_1^2\right)^2 + (1 - x_1)^2$ con $\mathbf{x}^* = (1,1)^T$, la cual se ilustra en la Figura 3.1.

2). La función de Powell:

$$f(\mathbf{x}) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4$$

con $\mathbf{x}^* = (0,0,0,0)^T$.

3). La función de Cragg- Levi:

$$f(\mathbf{x}) = \left(e^{x_1} - x_2\right)^4 + 100(x_2 - x_3)^6 + \tan^4(x_3 - x_4) + x_1^8 + (x_4 - 1)^2$$

con $\mathbf{x}^* = \left(0,1,1,(1 \pm n\pi)\right)^T$.

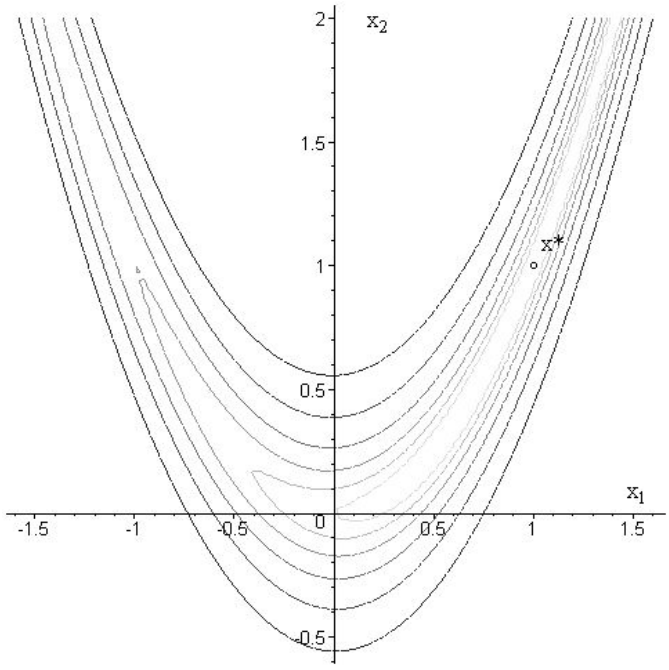


Figura 3.1 *Contornos de la función de Rosenbrock,* $\mathbf{x}^* = (1,1)^T$.

Método de Nelder y Mead (o del Poliedro Flexible)

El método de Nelder y Mead (1964) es una modificación del método simplex de Spendley, Hext y Himsworth (1962). Este método data desde 1964. Un conjunto de $(n+1)$ puntos mutuamente equidistantes en un espacio n-dimensional se conoce como un simplex o poliedro regular. Esta figura sirve de fundamento en el método de Spendley, Hext y Himsworth (1962). De esta manera, en dos dimensiones el simplex es un triángulo equilátero y en tres dimensiones es un tetraedro regular. La idea del método es comparar los valores de la función en los $(n+1)$ vértices del simplex y moverlos hacia el punto óptimo en cada estado. Nelder y Mead (1964) propusieron varias modificaciones al método que permiten a los simplejos formar figuras *no regulares*. El resultado es un método directo de búsqueda muy robusto si el número de variables no excede de 5 ó 6. El movimiento del simplejo en este método se realiza por la aplicación de tres operaciones básicas: reflexión, expansión y contracción. La idea fundamental de estas operaciones resultará clara cuando se consideren los pasos del siguiente procedimiento:

Dada una función $f(\mathbf{x})$, los $(n+1)$ puntos $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n+1}$ y la tolerancia ε .

Se calculan $f_1 = f(\mathbf{x}_1)$, $f_2 = f(\mathbf{x}_2), \ldots,$ $f_{n+1} = f(\mathbf{x}_{n+1})$.

1) Se procede a buscar el valor más grande de la función $f_h = f(\mathbf{x}_h)$, el siguiente valor superior $f_g = f(\mathbf{x}_g)$ y el valor más pequeño $f_l = f(\mathbf{x}_l)$, es decir,

$$f(\mathbf{x}_l) = \underset{1 \le i \le n+1}{mín} f(\mathbf{x}_i)$$

$$f(\mathbf{x}_g) = \underset{\substack{1 \le i \le n+1 \\ i \ne h}}{máx} f(\mathbf{x}_i)$$

$$f(\mathbf{x}_h) = \underset{1 \le i \le n+1}{máx} f(\mathbf{x}_i)$$

y los correspondientes puntos $\mathbf{x}_h$, $\mathbf{x}_g$ y $\mathbf{x}_l \in \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n+1}\}$.

2) Se determina el centroide de todos los puntos exceptuando $\mathbf{x}_h$. Sea este punto $\mathbf{x}_0$ y se evalúa $f_0 = f(\mathbf{x}_0)$, donde

$$\mathbf{x}_0 = \frac{1}{n} \sum_{\substack{i=1 \\ i \ne h}}^{n+1} \mathbf{x}_i \tag{3.2}$$

3) Como parece razonable tratar de moverse lejos de $\mathbf{x}_h$. Esto se hace *reflejando* $\mathbf{x}_h$ en $\mathbf{x}_0$ para hallar $\mathbf{x}_r$ y $f_r = f(\mathbf{x}_r)$. La reflexión se ilustra en la figura 3.2 para el caso de una función de dos variables.

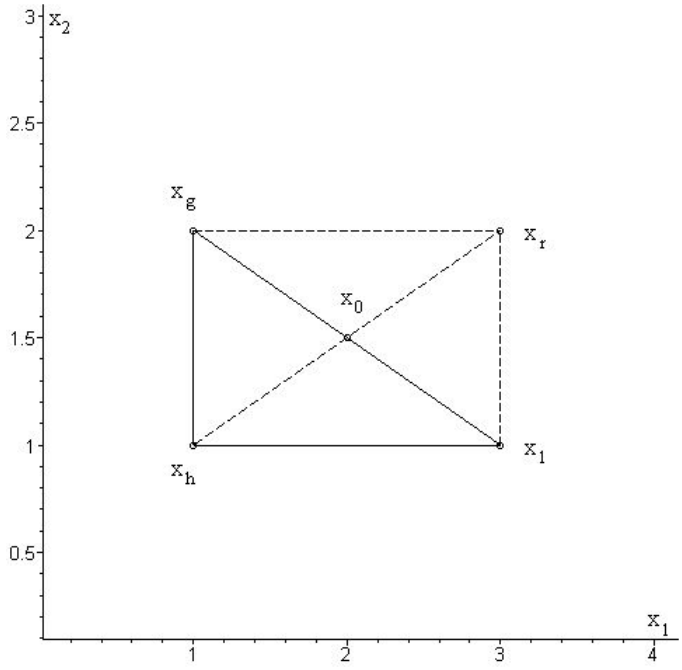


Figura 3.2 *Reflexión de* $\mathbf{x}_h$ *a través de* $\mathbf{x}_0$ *para obtener* $\mathbf{x}_r$.

Sea $\alpha > 0$ el *factor de reflexión*; se busca $\mathbf{x}_r$ de manera que

$$\mathbf{x}_r - \mathbf{x}_0 = \alpha\left(\mathbf{x}_0 - \mathbf{x}_h\right) \tag{3.3}$$

es decir,

$$\mathbf{x}_r = \left(1+\alpha\right)\mathbf{x}_0 - \alpha\mathbf{x}_h \tag{3.4}$$

4) Se compara f_r con f_l .

a) Si $f_r < f_l$, se obtuvo el valor más pequeño de la función. La dirección desde $\mathbf{x}_0$ hasta $\mathbf{x}_r$ parece ser buena para moverse a lo largo de ella. Por lo tanto, se hace una expansión en esta dirección para hallar $\mathbf{x}_e$ y $f_e = f(\mathbf{x}_e)$. La figura 3.3 ilustra la expansión del simplex:

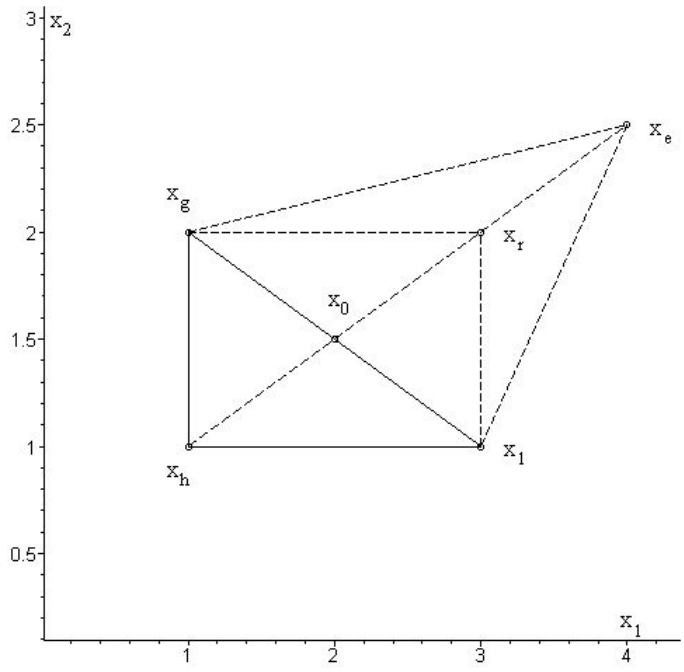


Figura 3.3 *Expansión de* $\mathbf{x}_r$ *a través de* $\mathbf{x}_0$ *para obtener* $\mathbf{x}_e$.

Con un *factor de expansión* $\gamma > 1$ se tiene

$$\mathbf{x}_e - \mathbf{x}_0 = \gamma(\mathbf{x}_r - \mathbf{x}_0) \tag{3.5}$$

es decir,

$$\mathbf{x}_e = \gamma\mathbf{x}_r + (1-\gamma)\mathbf{x}_0 \tag{3.6}$$

i) Si $f_e < f_l$, se hace $\mathbf{x}_h = \mathbf{x}_e$ y se prueban los $(n+1)$ puntos del simplex para la convergencia hacia el mínimo (paso 8). Si se ha convergido, se termina el proceso; en caso contrario, se regresa al paso 2.

ii) Si $f_e \geq f_l$, se abandona $\mathbf{x}_e$ debido a que se ha hecho un movimiento muy lejos en la dirección de $\mathbf{x}_0$ a $\mathbf{x}_r$. En su lugar, se hace $\mathbf{x}_h = \mathbf{x}_r$, se prueba la convergencia y si no se satisface, se regresa al paso 2.

b) Si a) no se satisface, es decir, si $f_r > f_l$ pero $f_r \leq f_g$, $\mathbf{x}_r$ es una mejoría sobre los dos peores puntos del simplex y se hace $\mathbf{x}_h = \mathbf{x}_r$, se prueba la convergencia y si no se satisface, se regresa al Paso 2.

c) Si $f_r > f_l$ y $f_r > f_g$, se procede al paso 5.

5) Se comparan f_r y f_h .

a) Si $f_r < f_h$, se hace $\mathbf{x}_h = \mathbf{x}_r$ y $f_h = f_r$. Recuérdese que $f_r > f_g$ (paso 4c).

Luego se procede al paso 5b. En caso contrario, es decir, si $f_r > f_h$ se procede directamente a la *contracción* (paso 5b).

b) Si $f_r < f_h$, se hace $\mathbf{x}_h = \mathbf{x}_r$ y $f_h = f_r$. Se realiza el proceso de contracción con un factor $0 < \beta < 1$, es decir, se determina $\mathbf{x}_c$ de

$$\mathbf{x}_c - \mathbf{x}_0 = \beta\left(\mathbf{x}_r - \mathbf{x}_0\right) \tag{3.7}$$

es decir,

$$\mathbf{x}_c = \beta\mathbf{x}_r + \left(1-\beta\right)\mathbf{x}_0 \tag{3.8}$$

la figura 3.4 ilustra este proceso.

En caso contrario, es decir, si $f_r > f_h$ se procede directamente a la contracción y se determina $\mathbf{x}_c$ de

$$\mathbf{x}_c - \mathbf{x}_0 = \beta\left(\mathbf{x}_h - \mathbf{x}_0\right) \tag{3.9}$$

es decir,

$$\mathbf{x}_c = \beta\mathbf{x}_h + \left(1-\beta\right)\mathbf{x}_0 \tag{3.10}$$

la figura 3.5 ilustra este caso.

6) Se comparan f_c y f_h.

a) Si $f_c < f_h$, se hace $\mathbf{x}_h = \mathbf{x}_c$ y $f_h = f_c$, se verifica la convergencia y si no se satisface, se regresa al paso 2.

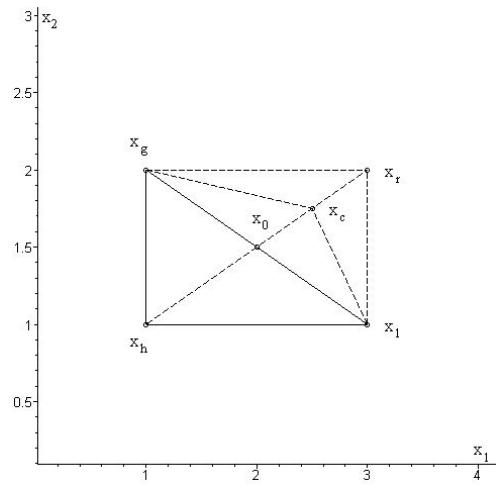


Figura 3.4 *Contracción de* $\mathbf{x}_r$ *a través de* $\mathbf{x}_0$ *para obtener* $\mathbf{x}_c$ *(*$\beta > 0$*).*

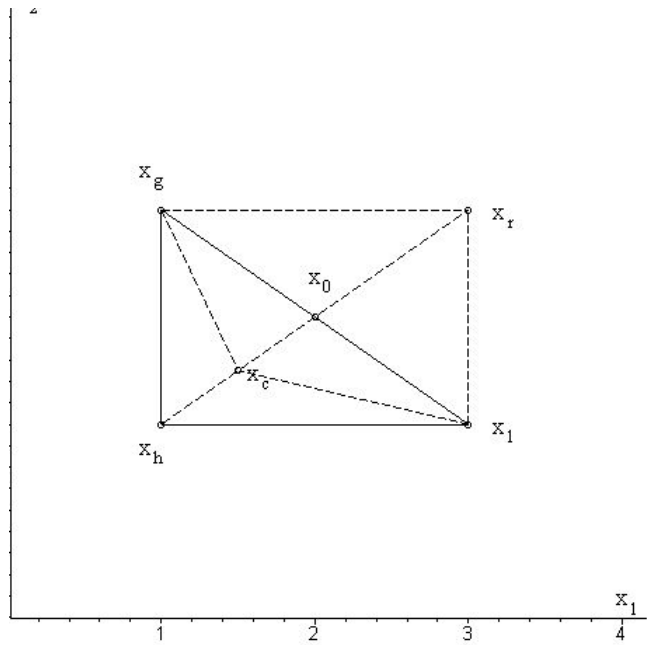


Figura 3.5 *Contracción de* $\mathbf{x}_r$ *a través de* $\mathbf{x}_0$ *para obtener* $\mathbf{x}_c$ *(*β<0 *).*

b) Si $f_c > f_h$, parecería que todos los esfuerzos por hallar un valor de $f < f_h$ fallaron, de manera que se va al paso 7.

7) En este paso, se reduce el tamaño del simplex dividiendo en dos la distancia de cada punto del simplex, desde $\mathbf{x}_l$, el punto generador del *mínimo* valor de la función. Así, $\mathbf{x}_i$ se reemplaza por

$$\mathbf{x}_i = \mathbf{x}_l + \frac{1}{2}\left(\mathbf{x}_i - \mathbf{x}_l\right) \tag{3.11}$$

es decir ;

$$\mathbf{x}_i = \frac{1}{2}\left(\mathbf{x}_i + \mathbf{x}_l\right) \tag{3.12}$$

luego se calcula $f_i = f\left(\mathbf{x}_i\right)$ para $i = 1,2,...,\left(n+1\right)$, se prueba la convergencia y si no se satisface, se regresa al paso 2. En la figura 3.6 se ilustra este proceso.

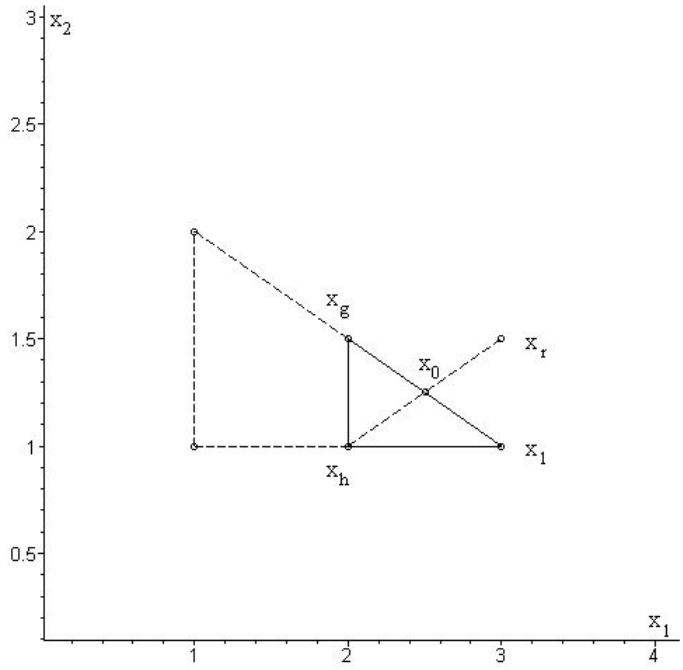


Figura 3.6 *Contracción de todo el simplex y una nueva reflexión.*

8) La prueba de convergencia se basa en la desviación estándar de los $\left(n+1\right)$ valores de la función, de manera que

$$\sigma < \varepsilon \tag{3.13}$$

donde ε es un valor pequeño predeterminado. El valor de σ se determina de:

$$\sigma^2 = \frac{1}{n+1}\sum_{i=1}^{n+1}\left(f_i - \bar{f}\right)^2 \tag{3.14}$$

donde

$$\overline{f} = \frac{1}{n+1}\sum_{i=1}^{n+1} f_i \tag{3.15}$$

Si $\sigma < \varepsilon$, todos los valores de la función están muy cercanos y de esta manera se puede decir con optimismo que todos los puntos están muy cerca del mínimo $\mathbf{x}_l$. Este criterio de convergencia suena razonable, aunque Box (1965) sugiere que éste se tome en consideración como una *prueba de seguridad*. Por último, Nelder y Mead (1964) recomiendan $\alpha = 1.0$, $\beta = 0.5$ y $\gamma = 2.0$ como valores de los factores de reflexión, contracción y expansión respectivamente, recomendación basada en ensayos con muchas combinaciones diferentes y donde parece que el método trabaja con eficiencia. Paviani (1969) recomendó como elección de γ y β los siguientes rangos de valores:

$$2.8 \leq \gamma \leq 3.0$$
$$0.4 \leq \beta \leq 0.6$$

La elección del simplex inicial es arbitraria. Se da un punto $\mathbf{x}_1$ y luego se generan los puntos restantes de acuerdo con

$$\mathbf{x}_{i+1} = \mathbf{x}_1 + h\hat{\mathbf{e}}_i \quad , \quad i = 1,...,n \tag{3.16}$$

donde h es una longitud de paso arbitraria y $\hat{\mathbf{e}}_i$ es un vector unitario en la dirección de la i-ésima coordenada.

Algoritmo 3.1 Método de Nelder y Mead:

Dada una función $f(\mathbf{x})$, un punto inicial $\mathbf{x}_1$, la tolerancia ε, los parámetros de reflexión α, expansión β, contracción δ y una longitud de paso arbitraria h.

1. Generar los puntos del simplex inicial:

$$\mathbf{x}_{i+1} = \mathbf{x}_1 + h\hat{\mathbf{e}}_i \text{ , para } i = 1,...,n \text{ ,}$$

donde $\hat{\mathbf{e}}_i$ son los vectores unitarios a lo largo de los ejes de coordenadas.

2. Calcular:

$$f_i = f(\mathbf{x}_i) \text{ , para } i = 1,...,n+1 .$$

Para k = 1, 2, 3, ..., hasta donde se satisfaga, hacer:

3. Identificar: $\mathbf{x}_l$, $\mathbf{x}_g$ y $\mathbf{x}_h \in \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n+1}\}$ tal que

$$f(\mathbf{x}_l) = \underset{1 \le i \le n+1}{mínimo} f(\mathbf{x}_i) \text{ ,}$$

$$f(\mathbf{x}_g) = \underset{\substack{1 \le i \le n+1 \\ i \ne h}}{máximo} f(\mathbf{x}_i),$$

$$f(\mathbf{x}_h) = \underset{1 \le i \le n+1}{máximo} f(\mathbf{x}_i).$$

4. Calcular el centroide:

$$\mathbf{x}_0 = \frac{1}{n}\sum_{\substack{j=1 \\ j\neq h}}^{n+1} \mathbf{x}_j$$

5. Calcular $\mathbf{x}_r = (1+\alpha)\mathbf{x}_0 - \alpha\mathbf{x}_h;\ f_r = f(\mathbf{x}_r)$.

6. Si $f_r < f_l$, entonces

   Calcular $\mathbf{x}_e = \delta\mathbf{x}_r + (1-\delta)\mathbf{x}_0;\ f_e = f(\mathbf{x}_e)$.

   Si $f_e < f_r$, entonces

   tomar $\mathbf{x}_h = \mathbf{x}_e;\ f_h = f_e;$

   en otro caso,

   tomar $\mathbf{x}_h = \mathbf{x}_r;\ f_h = f_r$.

   Fin del si.

   En otro caso

   Si $f_r \leq f_g$, entonces

   tomar $\mathbf{x}_h = \mathbf{x}_r;\ f_h = f_r$.

   En otro caso

   Si $f_r < f_h$, entonces

   calcular $\mathbf{x}_c = \beta\mathbf{x}_r + (1-\beta)\mathbf{x}_0;\ f_c = f(\mathbf{x}_c)$.

Si $f_r < f_c$, entonces

$$\mathbf{x}_i = \frac{\mathbf{x}_i + \mathbf{x}_l}{2};\ f_i = f\left(\mathbf{x}_i\right).$$

En otro caso,

tomar $\mathbf{x}_h = \mathbf{x}_c;\ f_h = f_c$.

Fin del si.

En otro caso,

calcular $\mathbf{x}_c = \beta\mathbf{x}_h + \left(1-\beta\right)\mathbf{x}_0;\ f_c = f\left(\mathbf{x}_c\right)$.

Si $f_h < f_c$, entonces

$$\mathbf{x}_i = \frac{\mathbf{x}_i + \mathbf{x}_l}{2};\ f_i = f\left(\mathbf{x}_i\right).$$

En otro caso,

tomar $\mathbf{x}_h = \mathbf{x}_c;\ f_h = f_c$.

Fin del si.
Fin del si.
Fin del si.
Fin del si.

7. Calcular $\overline{f} = \frac{1}{n+1}\sum_{i=1}^{n+1} f_i$.

8. Calcular $\sigma^2 = \dfrac{1}{n+1}\sum_{i=1}^{n+1}\left(f_i - \overline{f}\right)^2$.

9. Si $\sigma \leq \varepsilon$, entonces

   tomar $\mathbf{x}_l = \mathbf{x}^*$ y terminar.

   Fin del si.

10. Tomar $k = k+1$ y regresar a 3.

Ejemplo 3.2 Halle el mínimo de $f(\mathbf{x}) = 3(x_1 - 1)^2 + 10(x_2 - 2x_1)^2$ usando el algoritmo 3.1 e iniciando en $\mathbf{x}^T = (0,3)$, con $\varepsilon = 0.01$.

Solución:

Usando el algoritmo 3.1 el mínimo fue $\mathbf{x}^{*T} = (1.00, 2.00)$, donde $f(\mathbf{x}^*) = 0.00$. Este método llegó rápido a la solución; lo hizo en la primera iteración, pero como el simplex era todavía muy grande el criterio de convergencia no se pudo satisfacer sino hasta la décima iteración $k = 10$. La figura 3.7 muestra la evolución del simplex de este ejemplo hasta el punto 7.

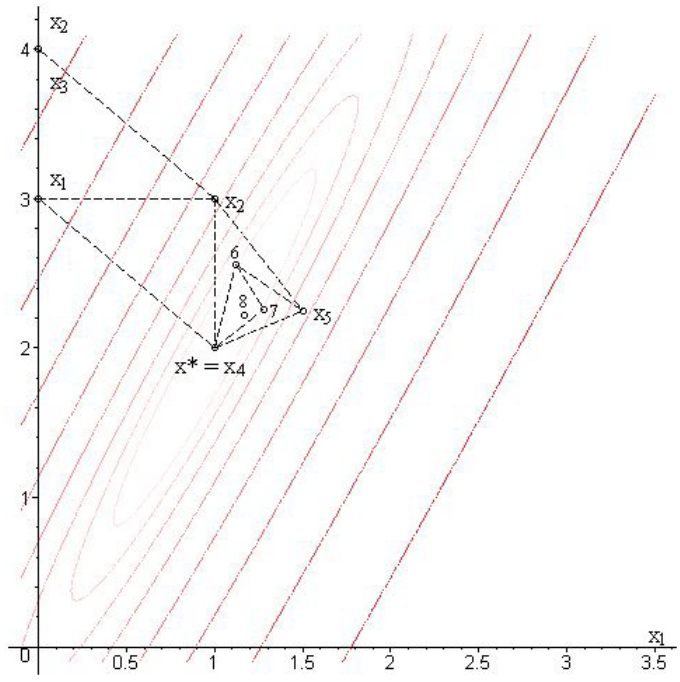


Figura 3.7 *Sucesión de 8 poliedros flexibles obtenidos en la minimización.*

3.3 Métodos con derivadas o gradientes

En contraste con la sección anterior, que describe un método para resolver el problema

$$\begin{aligned} &Minimizar \quad f(\mathbf{x}) \\ &con \qquad\quad \mathbf{x} \in R^n \end{aligned} \tag{3.17}$$

por una estrategia libre de derivadas, es decir, estrategia de búsqueda directa, en lo que sigue de este capítulo se estudiarán métodos que usan gradientes o derivadas de segundo orden, más conocidos como métodos indirectos de optimización. Como una regla general, en la solución de los problemas de programación no lineal sin restricciones los métodos de gradiente o derivadas de segundo

orden convergen más rápido que los métodos de búsqueda directa. Sin embargo, en la práctica los métodos con derivadas tienen dos problemas principales para su realización: primero, en los problemas con un número modestamente grande de variables resulta muy laborioso o imposible dar las funciones analíticas para las derivadas necesarias en un algoritmo de gradiente o de segundas derivadas; aunque la evaluación de las derivadas por esquemas de diferencias finitas puede sustituir el cálculo de las derivadas analíticas, no se evita la generación del error numérico por el uso de estas aproximaciones. Segundo, en principio también es posible usar la manipulación simbólica para desarrollar las derivadas analíticas, pero esto requiere una cantidad relativamente grande de preparación del problema por parte del usuario antes de que pueda desarrollarlo en un algoritmo comparado con las técnicas de búsqueda directa. Dadas las observaciones anteriores, se considerará cómo resolver el problema dado por la ecuación (3.17) por algoritmos que hacen uso de primeras y segundas derivadas parciales de $f(\mathbf{x})$

3.4 Técnicas de diferencias finitas

La disponibilidad de las derivadas de la función objetivo es de gran importancia en los métodos que se van a considerar. Con frecuencia se torna laborioso calcular analíticamente el gradiente o la matriz hessiana de una función compleja. En tales casos, es posible aproximar el gradiente o la matriz hessiana de la función por esquemas de diferencias finitas. Estos procedimientos se describirán brevemente.

Se asume que la función es continua y diferenciable; entonces, de acuerdo con la expansión en serie de Taylor de una función de varias variables, para una perturbación escalar suficientemente pequeña δ_i.

$$\begin{aligned} f(\mathbf{x}+\delta_i\mathbf{e}_i) &= f(\mathbf{x})+\delta_i\mathbf{e}_i^T\nabla f(\mathbf{x})+\frac{1}{2}\delta_i^2\mathbf{e}_i^T\mathbf{H}(\mathbf{x})\mathbf{e}_i+\cdots \\ &= f(\mathbf{x})+\delta_i g_i(\mathbf{x})+\frac{1}{2}\delta_i^2 h_{ii}(\mathbf{x})+\cdots \end{aligned} \tag{3.18}$$

donde $\mathbf{e}_i$ es el vector unitario en la dirección de la i-ésima coordenada, $g_i(\mathbf{x})$ es la i-ésima componente de $\nabla f(\mathbf{x})$ y h_{ii} es el ii-ésimo elemento de $\mathbf{H}(\mathbf{x})$. De este modo, al aproximar $g_i(\mathbf{x})$ por $\tilde{g}_i(\mathbf{x})$ se obtiene la aproximación por *diferencias finitas hacia delante* que es exacta solo para funciones lineales

$$\tilde{g}_i(\mathbf{x}) = \frac{f(\mathbf{x}+\delta_i\mathbf{e}_i) - f(\mathbf{x})}{\delta_i} \tag{3.19}$$

donde se ignoraron los términos de segundo orden y superiores en δ_i de la serie de Taylor. En forma análoga, de

$$f(\mathbf{x}-\delta_i\mathbf{e}_i) = f(\mathbf{x}) - \delta_i\mathbf{e}_i^T\nabla f(\mathbf{x}) + \frac{1}{2}\delta_i^2\mathbf{e}_i^T\boldsymbol{H}(\mathbf{x})\mathbf{e}_i + \cdots$$

$$f(\mathbf{x}-\delta_i\mathbf{e}_i) = f(\mathbf{x}) - \delta_i g_i(\mathbf{x}) + \frac{1}{2}\delta_i^2 h_{ii}(\mathbf{x}) + \cdots \tag{3.20}$$

se obtiene el esquema de *diferencias finitas* hacia atrás dado por

$$\tilde{g}_i(\mathbf{x}) = \frac{f(\mathbf{x}) - f(\mathbf{x}-\delta_i\mathbf{e}_i)}{\delta_i} \tag{3.21}$$

y de esta expresión y la anterior (3.19) se puede obtener el esquema de las *diferencias finitas centrales* dado por

$$\tilde{g}_i(\mathbf{x}) = \frac{f(\mathbf{x}+\delta_i\mathbf{e}_i) - f(\mathbf{x}-\delta_i\mathbf{e}_i)}{2\delta_i} \tag{3.22}$$

En realidad, la aproximación en diferencias centrales ignora sólo términos de tercer orden y superiores en δ_i de la serie de Taylor y es, por lo tanto, exacta para funciones cuadráticas. Por supuesto, el incremento en la precisión es a expensas de realizar el doble de evaluaciones funcionales más que la técnica de las diferencias hacia delante o hacia atrás.

El problema práctico en la aplicación de estas ideas es la elección del valor de δ suficientemente pequeño de manera que los diferentes términos de error queden balanceados. Varias librerías de software comercial proveen la rutina para calcular un valor adecuado de δ. Normalmente se usa un valor pequeño, como $\delta = \sqrt{\varepsilon}$, donde ε es la precisión relativa de la computadora en cuestión. Cuando las primeras derivadas están disponibles en forma analítica, es posible estimar la matriz hessiana completa. La expansión en serie de Taylor para el vector gradiente es

$$\begin{aligned}\nabla f\left(\mathbf{x}+\delta_i\mathbf{e}_i\right) &= \nabla f\left(\mathbf{x}\right)+\delta_i \boldsymbol{H}\left(\mathbf{x}\right)\mathbf{e}_i+\cdots \\ &= \nabla f\left(\mathbf{x}\right)+\delta_i \mathbf{h}_i\left(\mathbf{x}\right)+\cdots\end{aligned} \tag{3.23}$$

donde $\mathbf{h}_i\left(\mathbf{x}\right)$ es la i-ésima columna de $\mathbf{H}\left(\mathbf{x}\right)$. Por lo tanto, la aproximación en diferencias finitas hacia delante $\tilde{\mathbf{h}}_i\left(\mathbf{x}\right)$ de $\mathbf{h}_i\left(\mathbf{x}\right)$ la i-ésima columna de $\mathbf{H}\left(\mathbf{x}\right)$, esta dada por

$$\tilde{\mathbf{h}}_i\left(\mathbf{x}\right)=\frac{\nabla f\left(\mathbf{x}+\delta_i\mathbf{e}_i\right)-\nabla f\left(\mathbf{x}\right)}{\delta_i} \tag{3.24}$$

que es exacta para funciones cuadráticas. Las aproximaciones en diferencias finitas hacia atrás y centrales son

$$\tilde{\mathbf{h}}_i\left(\mathbf{x}\right)=\frac{\nabla f\left(\mathbf{x}\right)-\nabla f\left(\mathbf{x}-\delta_i\mathbf{e}_i\right)}{\delta_i} \tag{3.25}$$

$$\tilde{\mathbf{h}}_i\left(\mathbf{x}\right)=\frac{\nabla f\left(\mathbf{x}+\delta_i\mathbf{e}_i\right)-\nabla f\left(\mathbf{x}-\delta_i\mathbf{e}_i\right)}{2\delta_i} \tag{3.26}$$

respectivamente. La aplicación repetida de la ecuación (3.26) para $i = 1,2,...,n$, donde n es el número de variables, permite construir una aproximación $\tilde{\mathbf{H}}(\mathbf{x})$ de la matriz hessiana $\mathbf{H}(\mathbf{x})$. En general, esta matriz no será exactamente simétrica debido a los errores introducidos en los cálculos, y entonces será mejor usar $1/2\left(\tilde{\mathbf{H}} + \tilde{\mathbf{H}}^T\right)$ en lugar de $\tilde{\mathbf{H}}(\mathbf{x})$.

Ejemplo 3.3 Evaluación numérica del gradiente.

Evalúe el gradiente de la función $f(\mathbf{x}) = x_1^2 + x_2$ en el punto $\mathbf{x}_0^T = (2,1)$ por las aproximaciones en diferencias finitas y compare éstas con el gradiente analítico. Use una perturbación del 1% en las variables independientes.

Solución:

El gradiente analítico de la función es

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 2x_1 \\ 1 \end{pmatrix}$$

evaluado en $\mathbf{x}_0$ da

$$\nabla f(\mathbf{x}_0) = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$$

Al 1% de cambio en las variables se tiene $\delta_1 = 0.02$ y $\delta_2 = 0.01$. Los gradientes numéricos obtenidos aplicando las ecuaciones (3.19), (3.21) y (3.22) son

$$\tilde{\nabla} f(\mathbf{x}_0) = \begin{pmatrix} 4.02 \\ 1.00 \end{pmatrix}, \quad \tilde{\nabla} f(\mathbf{x}_0) = \begin{pmatrix} 3.98 \\ 1.00 \end{pmatrix} \text{ y } \tilde{\nabla} f(\mathbf{x}_0) = \begin{pmatrix} 4.00 \\ 1.00 \end{pmatrix}$$

respectivamente, como puede verificar el lector.

Observe que para la función dada los tres métodos dan una muy buena aproximación al gradiente analítico. El método por diferencias centrales da lo mismo que el gradiente analítico, lo cual comprueba que esta aproximación es exacta para funciones cuadráticas. Observe también que los tres métodos son iguales en $\tilde{g}_2(\mathbf{x}_0)$; esto se debe a que la función es lineal en x_2, o sea que también se comprueba que la diferenciación numérica siempre da gradientes exactos en funciones lineales.

Ejemplo 3.4 Evaluación numérica de la matriz hessiana.

Evalúe la matriz hessiana de la función $f(\mathbf{x}) = x_1^3 + 2x_2^2 - 2x_1 - 3x_2$ en el punto $\mathbf{x}_0^T = (2,1)$ por las aproximaciones en diferencias finitas y compare éstas con la matriz hessiana analítica. Use una perturbación del 1% en las variables independientes.

Solución:

El gradiente y la matriz hessiana analíticos de la función son

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 3x_1^2 - 2 \\ 4x_2 - 3 \end{pmatrix} \text{ y } \mathbf{H}(\mathbf{x}) = \begin{pmatrix} 6x_1 & 0 \\ 0 & 4 \end{pmatrix}$$

evaluada la matriz en $\mathbf{x}_0$ da

$$\mathbf{H}(\mathbf{x}_0) = \begin{pmatrix} 12 & 0 \\ 0 & 4 \end{pmatrix}$$

Al 1% de cambio en las variables se tiene $\delta_1 = 0.02$ y $\delta_2 = 0.01$. Las matrices hessianas numéricas de la función por las tres aproximaciones aplicando las ecuaciones (3.24), (3.25) y (3.26) dan como resultado

$$\tilde{\mathbf{H}}(\mathbf{x}_0) = \begin{pmatrix} 12.06 & 0.00 \\ 0.00 & 4.00 \end{pmatrix}, \quad \tilde{\mathbf{H}}(\mathbf{x}_0) = \begin{pmatrix} 11.96 & 0.00 \\ 0.00 & 4.00 \end{pmatrix} \text{ y } \tilde{\mathbf{H}}(\mathbf{x}_0) = \begin{pmatrix} 12.0 & 0.0 \\ 0.0 & 4.0 \end{pmatrix}$$

respectivamente, como puede verificar el lector.

Observe que para la función dada, los tres métodos dan una muy buena aproximación a la matriz hessiana analítica. De nuevo el método por diferencias centrales da lo mismo que la hessiana analítica.

3.5 Direcciones conjugadas

En el método de Newton será interesante observar que, en contraste con el método de Cauchy (que se verá más adelante), la dirección de búsqueda no es $-\nabla f(\mathbf{x}_k)$ sino $-\mathbf{H}(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$ si se toman en cuenta las segundas derivadas. Los métodos de Fletcher-Reeves y Davidon-Fletcher-Powell (DFP) tratan de obtener la mejor búsqueda investigando en la dirección $\mathbf{d}_k$ y $-\mathbf{G}(\mathbf{x}_k)\nabla f(\mathbf{x}_k)$ respectivamente en la k-ésima etapa, donde $\mathbf{d}_k$ es una combinación lineal de gradientes en el método de Fletcher-Reeves y $\mathbf{G}(\mathbf{x}_k)$ es una matriz simétrica positiva definida que finalmente se iguala a $-\mathbf{H}(\mathbf{x}^*)^{-1}$. De esta manera, ambos métodos evitan la evaluación y la inversión de $\mathbf{H}(\mathbf{x}_k)$ en cada paso. La dirección de búsqueda en cada etapa es, así, un factor crucial en la eficiencia de los métodos de búsqueda iterativos. En cualquier etapa se desea obtener la próxima búsqueda en la *mejor dirección*.

Los métodos de dirección conjugada se pueden considerar como algo intermedio entre el método de Cauchy y el de Newton. Estos métodos se diseñan y analizan sin excepción para el problema de una función cuadrática pura de n variables tal como

$$f(\mathbf{x}) = a + \mathbf{x}^T\mathbf{b} + \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} \tag{3.27}$$

Ahora bien, la mejor dirección en un cierto sentido es en la dirección que es *conjugada* a las direcciones de búsqueda previas. Primero se definirá este concepto y luego se explicará su utilidad.

Definición 3.1 Dada una matriz simétrica $\mathbf{H}$, se dice que dos vectores $\mathbf{p}$ y $\mathbf{q}$ son *ortogonales* respecto a $\mathbf{H}$ o *conjugados* respecto a $\mathbf{H}$ si

$$\mathbf{p}^T\mathbf{H}\mathbf{q} = 0. \tag{3.28}$$

El lector puede mostrar, a partir de esta definición, que si $\mathbf{H}$ es positiva definida y el conjunto de vectores distintos de cero $\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_{n-1}$ son n *direcciones mútuamente conjugadas* en un espacio n dimensional, entonces estos vectores son linealmente independientes. Pero si no, existirán constantes $\alpha_0, \alpha_1, ..., \alpha_{n-1}$, no todas nulas, tales que

$$\alpha_0\mathbf{p}_0 + \alpha_1\mathbf{p}_1 + \cdots + \alpha_{n-1}\mathbf{p}_{n-1} = \mathbf{0}$$

En lugar de analizar el algoritmo de dirección conjugada general, se investigará por qué el concepto de conjugación con respecto a $\mathbf{H}$ es útil para la solución del problema cuadrático

$$f(\mathbf{x}) = a + \mathbf{x}^T\mathbf{b} + \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$$

la cual tiene su punto óptimo en $\mathbf{x}^* = -\mathbf{H}^{-1}\mathbf{b}$.

Es conveniente reescribir la ecuación de $f(\mathbf{x})$ como un desarrollo en serie de Taylor alrededor de $\mathbf{x}^*$ hasta términos de segundo orden

$$f(\mathbf{x}) \approx f(\mathbf{x}^*) + (\mathbf{x}-\mathbf{x}^*)^T \nabla f(\mathbf{x}^*) + \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T \mathbf{H}(\mathbf{x}^*)(\mathbf{x}-\mathbf{x}^*)$$

Como $\nabla f(\mathbf{x}^*) = \mathbf{0}$, entonces

$$f(\mathbf{x}) \approx c + \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)\mathbf{H}(\mathbf{x}^*)(\mathbf{x}-\mathbf{x}^*) \tag{3.29}$$

donde $c = f(\mathbf{x}^*)$ es una constante.

Supóngase que se desea usar una técnica iterativa para hallar el mínimo de la ecuación (3.29). Es claro que no puede decidirse *a priori* cuales son las direcciones de búsqueda por utilizar para resolver el problema, sino que más bien debe admitirse que el conocimiento ganado por búsquedas anteriores ayudará a determinar las direcciones subsecuentes. Entonces, iníciese en $\mathbf{x}_0$ y búsquese en la dirección $\mathbf{p}_0$ para hallar el mínimo

$$\mathbf{x}_1 = \mathbf{x}_0 + \alpha_0 \mathbf{p}_0 \tag{3.30}$$

donde α_0 es algún escalar. Obsérvese que en $\mathbf{x}_1$, $\nabla f(\mathbf{x}_1)$ es ortogonal respecto a $\mathbf{p}_0$, es decir

$$\nabla^T f(\mathbf{x}_1)\mathbf{p}_0 = 0 \tag{3.31}$$

debido a que

$$\frac{d\phi(\alpha_0)}{d\alpha_0} = \frac{df(\mathbf{x}_1)}{d\alpha_0} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_{1i}} \frac{dx_{1i}}{d\alpha_0} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_{1i}} p_{0i} = \nabla^T f(\mathbf{x}_1)\mathbf{p}_0 = 0$$

en el mínimo $\mathbf{x}_1$.

En general, en el k-ésimo paso se inicia desde un punto $\mathbf{x}_k$ en la dirección $\mathbf{p}_k$ para hallar el mínimo en

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k \tag{3.32}$$

donde

$$\nabla^T f\left(\mathbf{x}_{k+1}\right)\mathbf{p}_k = 0 \tag{3.33}$$

y

$$\nabla f\left(\mathbf{x}_k\right) = \mathbf{H}\left(\mathbf{x}_k - \mathbf{x}^*\right) \tag{3.34}$$

para $f\left(\mathbf{x}\right)$ en la ecuación (3.29). Por uso repetido de la ecuación (3.32) después de n pasos, se obtiene

$$\begin{aligned}\mathbf{x}_n &= \mathbf{x}_{n-1} + \alpha_{n-1}\mathbf{p}_{n-1} \\ &= \mathbf{x}_{n-2} + \alpha_{n-2}\mathbf{p}_{n-2} + \alpha_{n-1}\mathbf{p}_{n-1}\end{aligned}$$

y, por lo tanto,

$$\mathbf{x}_n = \mathbf{x}_{j+1} + \sum_{i=j+1}^{n-1} \alpha_i \mathbf{p}_i \tag{3.35}$$

para toda j en $0 \leq j < n-1$. Así, de la ecuación (3.35)

$$\left(\mathbf{x}_n - \mathbf{x}^*\right) = \left(\mathbf{x}_{j+1} - \mathbf{x}^*\right) + \sum_{i=j+1}^{n-1} \alpha_i \mathbf{p}_i \tag{3.36}$$

Multiplicando por $\mathbf{H}$ y tomando el producto escalar con $\mathbf{p}_j$, se tiene

$$\mathbf{p}_j^T \mathbf{H}\left(\mathbf{x}_n - \mathbf{x}^*\right) = \mathbf{p}_j^T \mathbf{H}\left(\mathbf{x}_{j+1} - \mathbf{x}^*\right) + \sum_{i=j+1}^{n-1} \alpha_i \mathbf{p}_j^T \mathbf{H} \mathbf{p}_i \tag{3.37}$$

usando la ecuación (3.34)

$$\mathbf{p}_j^T \nabla f\left(\mathbf{x}_n\right) = \mathbf{p}_j^T \nabla f\left(\mathbf{x}_{j+1}\right) + \sum_{i=j+1}^{n-1} \alpha_i \mathbf{p}_j^T \boldsymbol{H} \mathbf{p}_i \tag{3.38}$$

o bien

$$\mathbf{p}_j^T \nabla f\left(\mathbf{x}_n\right) = \sum_{i=j+1}^{n-1} \alpha_i \mathbf{p}_j^T \boldsymbol{H} \mathbf{p}_i \tag{3.39}$$

debido a la ecuación (3.33). Si todos los vectores $\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_{n-1}$ son mutuamente conjugados, tales que

$$\mathbf{p}_i^T \mathbf{H} \mathbf{p}_j = 0 \text{ , para } i \neq j \tag{3.40}$$

entonces

$$\mathbf{p}_j^T \nabla f\left(\mathbf{x}_n\right) = 0 \text{, para } j = 0,1,...,n-1 \tag{3.41}$$

(También es cierta para $j = n-1$, según la ecuación (3.33)). Pero, ya que en este caso los $\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_{n-1}$ son linealmente independientes y por lo tanto forman una base, se sigue que

$$\nabla f\left(\mathbf{x}_n\right) = \mathbf{0} \tag{3.42}$$

de donde, por la ecuación (3.34)

$$\mathbf{H}\left(\mathbf{x}_n - \mathbf{x}^*\right) = \mathbf{0} \tag{3.43}$$

y, por lo tanto,

$$\mathbf{x}_n = \mathbf{x}^* \tag{3.44}$$

Se sigue de esto que si la búsqueda se lleva a cabo en direcciones mutuamente conjugadas, se hallará el mínimo de una función cuadrática de n variables en n pasos o iteraciones. Sustituyendo la ecuación (3.44) en la ecuación (3.35) con $j = 0$, resulta

$$\mathbf{x}^* = \sum_{i=0}^{n-1} \alpha_i \mathbf{p}_i + \mathbf{x}_0 \tag{3.45}$$

multiplicando por $\mathbf{H}$ y tomando el producto escalar con $\mathbf{p}_j$, se tiene

$$\alpha_i = -\frac{\mathbf{p}_i^T \left(\mathbf{H}\mathbf{x}_0 + \mathbf{b}\right)}{\mathbf{p}_i^T \mathbf{H}\mathbf{p}_i} \tag{3.46}$$

Esto muestra que las α_i y, en consecuencia, la solución $\mathbf{x}^*$ se pueden hallar por la evaluación de productos escalares simples. El resultado final es

$$\mathbf{x}^* = \mathbf{x}_0 - \sum_{i=0}^{n-1} \frac{\mathbf{p}_i^T \left(\mathbf{H}\mathbf{x}_0 + \mathbf{b}\right)}{\mathbf{p}_i^T \mathbf{H}\mathbf{p}_i} \mathbf{p}_i \tag{3.47}$$

Este resultado muestra que la expansión para $\mathbf{x}^*$ se puede considerar como el resultado de un proceso iterativo de n pasos en el que se añade $\alpha_i \mathbf{p}_i$ en el i-ésimo paso. Visto de esta forma el procedimiento y admitiendo un punto inicial arbitrario, se obtiene el método básico de la dirección conjugada. Los métodos de Fletcher-Reeves, DFP y BFGS buscarán explotar esta idea.

Ejemplo 3.5

Minimice $f(\mathbf{x}) = 4(x_1 - 5)^2 + (x_2 - 6)^2$ usando direcciones mutuamente conjugadas. Esta función tiene un mínimo en $\mathbf{x}^* = (5,6)^T$, donde $f(\mathbf{x}^*) = 0.0$. Inicie en $\mathbf{x}_0 = (8,9)^T$, donde $f(\mathbf{x}_0) = 45.0$ y tome como direcciones de búsqueda los vectores $\mathbf{e}_1 = (1,0)^T$ y $\mathbf{e}_2 = (0,1)^T$. ¿Son conjugados los vectores $\mathbf{e}_1$ y $\mathbf{e}_2$ con respecto a la matriz hessiana de $f(\mathbf{x})$?

Solución:

La matriz hessiana de $f(\mathbf{x}) = 4x_1^2 - 40x_1 + x_2^2 - 12x_2 + 136$ es

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} 8 & 0 \\ 0 & 2 \end{pmatrix}$$

Luego,

$$\mathbf{e}_1^T \mathbf{H}\mathbf{e}_2 = 0$$

y por tanto $\mathbf{e}_1$ y $\mathbf{e}_2$ son vectores mutuamente conjugados y linealmente independientes; por otro lado, $\mathbf{b} = \begin{pmatrix} -40 & -12 \end{pmatrix}^T$ es el vector de coeficientes de los términos lineales, usando la ecuación (3.47), se tiene

$$\mathbf{x}^* = \mathbf{x}_0 - \sum_{i=1}^{n} \frac{\mathbf{e}_i^T(\mathbf{H}\mathbf{x}_0 + \mathbf{b})}{\mathbf{e}_i^T \mathbf{H}\mathbf{e}_i}\mathbf{e}_i = \mathbf{x}_0 - \frac{\mathbf{e}_1^T(\mathbf{H}\mathbf{x}_0 + \mathbf{b})}{\mathbf{e}_1^T \mathbf{H}\mathbf{e}_1}\mathbf{e}_1 - \frac{\mathbf{e}_2^T(\mathbf{H}\mathbf{x}_0 + \mathbf{b})}{\mathbf{e}_2^T \mathbf{H}\mathbf{e}_2}\mathbf{e}_2$$

por lo tanto,

$$\mathbf{x}^* = \begin{pmatrix} 8 \\ 9 \end{pmatrix} - \frac{24}{8}\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \frac{6}{2}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 6 \end{pmatrix}$$

que es el mínimo de $f(\mathbf{x})$. Observe que la función es cuadrática en dos variables, por eso el método de direcciones conjugadas alcanzó el mínimo en 2 iteraciones como se esperaba. Para una función no cuadrática se requiere más iteraciones.

3.6 Métodos de primer y segundo orden

Método de Cauchy o de descenso acelerado

En esta ocasión se considerará un método que usa el gradiente de la función, así como los valores de la función misma. El método de direcciones conjugadas que se acaba de describir, consistió en la búsqueda del mínimo a partir de un punto dado hacia una dirección paralela a uno de los ejes en la dirección del mínimo, luego se buscó en otra dirección paralela a otro de los ejes en la dirección del mínimo, y así sucesivamente se procede si la función es de más de dos variables, parece razonable tratar de modificar este método de manera que en cada paso la búsqueda hacia el mínimo se lleve a cabo a lo largo de la *mejor* dirección.

No está claro cuál es la mejor dirección, pero la dirección opuesta a la del gradiente tiene cierto atractivo intuitivo. La dirección del gradiente es la dirección del ascenso acelerado. Así, la dirección opuesta será la dirección del descenso acelerado. Esta propiedad puede probarse como sigue: supóngase que desde un punto $\mathbf{x}$ nos movemos a un punto cercano $\mathbf{x} + \alpha\mathbf{d}$, donde $\mathbf{d}$ es alguna dirección y α alguna longitud de paso. De este modo, nos movemos desde $(x_1, x_2, ..., x_n)$ hasta $(x_1 + \delta x_1, x_2 + \delta x_2, ..., x_n + \delta x_n)$, donde

$$\delta x_i = \alpha d_i \ , \ \ i = 1,2,...,n \tag{3.48}$$

el cambio en el valor de la función esta dado por

$$df = f\left(\mathbf{x}+\delta\mathbf{x}\right) - f\left(\mathbf{x}\right) = \sum_{i=1}^{n} \frac{\partial f\left(\mathbf{x}\right)}{\partial x_i}\delta x_i \tag{3.49}$$

hasta primer orden en las δx_i, donde las derivadas parciales están evaluadas en $\mathbf{x}$. ¿Cómo deben elegirse las d_i sujetas a la ecuación (3.48) de manera que se obtenga el mayor valor posible para df? Otra forma de expresar la ecuación (3.49) es

$$df = \nabla^T f\left(\mathbf{x}\right)\cdot d\mathbf{x} = \left\|\nabla^T f\left(\mathbf{x}\right)\right\|\left\|d\mathbf{x}\right\|\cos\theta \tag{3.50}$$

donde θ es el ángulo entre $\nabla f\left(\mathbf{x}\right)$ y $d\mathbf{x}$. Para una magnitud dada de $d\mathbf{x}$, se minimiza $df\left(\mathbf{x}\right)$ tomando

$$\frac{d\left(df\right)}{d\theta} = -\left\|\nabla^T f\left(\mathbf{x}\right)\right\|\left\|d\mathbf{x}\right\| sen\,\theta = 0 \quad \text{si y sólo si } \theta = 0^\circ \text{ ó } 180^\circ \tag{3.51}$$

ahora bien,

$$\frac{d^2\left(df\right)}{d\theta^2} = -\left\|\nabla^T f\left(\mathbf{x}\right)\right\|\left\|d\mathbf{x}\right\|\cos\theta \text{ implica que } \begin{cases} < 0 & para \quad \theta = 0^\circ \\ > 0 & para \quad \theta = 180^\circ \end{cases} \tag{3.52}$$

de manera que, eligiendo $\theta = 180^\circ$, $d\mathbf{x}$ está en la dirección de $-\nabla f\left(\mathbf{x}\right)$ y $df\left(\mathbf{x}\right)$ alcanza su mínimo valor "local". De este modo, el mayor incremento "local" en la función para un paso "pequeño" α dado ocurre cuando $\mathbf{d}$ está en la dirección del $-\nabla f\left(\mathbf{x}\right)$.

Así, la dirección del descenso acelerado está en la dirección de

$$\mathbf{d} = -\nabla f(\mathbf{x}) \tag{3.53}$$

Recuérdese que la dirección del gradiente es ortogonal al contorno de la función en cualquier punto, porque sobre un contorno el valor de la función no cambia. Así, si $\mathbf{d}$ es un paso pequeño a lo largo del contorno

$$f(\mathbf{x} + \delta\mathbf{x}) = f(\mathbf{x})$$

es decir,

$$df(\mathbf{x}) = 0 = \nabla^T f(\mathbf{x}) \cdot \mathbf{d} \text{ si y sólo si } \nabla^T f(\mathbf{x}) \text{ es ortogonal a } \mathbf{d} \quad . \tag{3.54}$$

El método de descenso acelerado busca explotar esta propiedad de la dirección del gradiente. Por ello, si estamos en cualquier etapa en el punto $\mathbf{x}_k$, en el proceso buscamos el mínimo para la función $f(\mathbf{x})$ a lo largo de la dirección de $-\nabla f(\mathbf{x})$. El método resulta ser un proceso iterativo, y en la etapa k tenemos una aproximación $\mathbf{x}_k$ para el punto mínimo. La siguiente aproximación es

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha^* \nabla f(\mathbf{x}_k) \tag{3.55}$$

donde α^* es el valor que minimiza

$$\phi(\alpha) \equiv f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)) \tag{3.56}$$

Esta α^* puede determinarse usando una de las búsquedas univariadas anteriores (bisección, Newton-Raphson, sección áurea o interpolación cuadrática entre otros métodos univariables).

Recuérdese que este método tiene un atractivo puramente intuitivo y es de interés académico, pero en la práctica es de convergencia muy lenta. El punto es que la propiedad del descenso acelerado es solamente una propiedad "local", de manera que se necesitan cambios frecuentes de dirección, guiando con esto a un procedimiento de cálculo "lento". Por último, se puede decir que el método de Cauchy es simple y robusto, es convergente lentamente y sin embargo, tiene varias desventajas:

1) Aun cuando la convergencia del método esta garantizada, puede requerirse un gran número de iteraciones para la minimización de funciones cuadráticas con matriz definida positiva, o sea que el método puede ser bastante lento en la convergencia hacia el punto mínimo.

2) No se usa la información calculada en iteraciones previas; cada iteración es independiente de las otras y esto hace al método ineficiente.

3) Se usa sólo información de primer orden en cada iteración para establecer la dirección de búsqueda, razón por la cual la convergencia del método es lenta. Otra razón del deterioro de la convergencia se da si se usa búsqueda lineal inexacta. Además, la rapidez de convergencia depende del número de condición de la hessiana de la función objetivo en el punto óptimo; si el número de condición es grande, la rapidez de convergencia del método resulta ser lenta.

4) La experiencia práctica con el método ha mostrado que se obtiene una disminución sustancial en la función objetivo al inicio de las primeras iteraciones, y en las últimas iteraciones la función objetivo disminuye con gran lentitud.

5) La dirección de descenso acelerado resulta ser buena en un sentido local, pero no en un sentido global.

Algoritmo 3.2 Método de Cauchy con búsqueda lineal:

Dada una función $f(\mathbf{x})$ continua y derivable, un punto inicial $\mathbf{x}_0$ y ε.
Para $k = 0,1,2,...$ hasta donde se satisfaga, hacer:

1. $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$.
2. Hallar α^* que minimiza $f(\mathbf{x}_k + \alpha \mathbf{d}_k)$.

3. Calcular $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha^* \mathbf{d}_k$.
4. Si $\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\| \leq \varepsilon$ o $\left\|\nabla f\left(\mathbf{x}_{k+1}\right)\right\| \leq \varepsilon$, entonces

   Tomar $\mathbf{x}_{k+1} \approx \mathbf{x}^*$ y parar.
   Fin del si.
5. Tomar $k = k+1$ y regresar al paso 1.

Ejemplo 3.6 Halle el mínimo de $f\left(\mathbf{x}\right) = 3\left(x_1 - 1\right)^2 + 10\left(x_2 - 2x_1\right)^2$, por el método de Cauchy iniciando en $\mathbf{x}^T = \left(0,3\right)$, con $\varepsilon = 0.01$.

Solución: Usando el algoritmo 3.2 el mínimo hallado con la precisión dada es, $\mathbf{x}^{*T} = \left(1.0002, 2.0005\right)$ donde $f\left(\mathbf{x}^*\right) = 0.0$. Obsérvese que la convergencia del método fue rápida (figura 3.8) debido al uso de la búsqueda lineal inexacta, el número total de iteraciones principales fue de $k = 3$. La figura 3.8 muestra en forma gráfica la búsqueda hacia el óptimo.

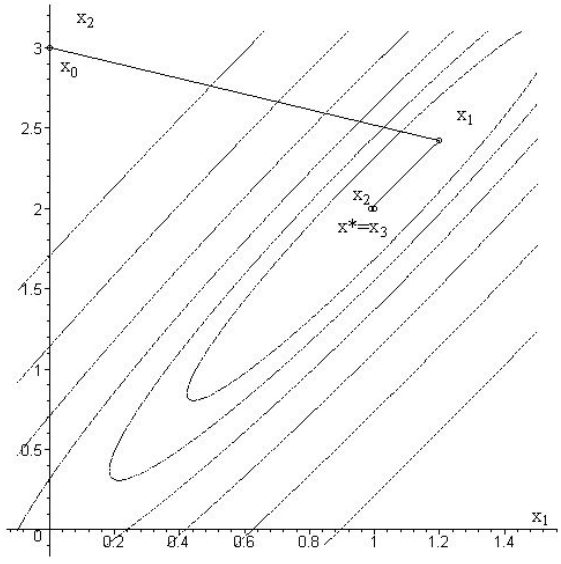


Figura 3.8 *Evolución del método del descenso acelerado en 3 iteraciones.*

Método de Newton

Desde otro punto de vista, el método de Cauchy puede interpretarse como una aproximación lineal de la función $f(\mathbf{x})$. Por tanto, los métodos de segundas derivadas entre los cuales el más conocido es el método de Newton, se originan por una aproximación cuadrática de $f(\mathbf{x})$ dada por

$$f(\mathbf{x}) \approx f(\mathbf{x}_k) + (\mathbf{x}-\mathbf{x}_k)^T \nabla f(\mathbf{x}_k) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_k)^T \mathbf{H}(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k) \quad (3.57)$$

Este método explota la información obtenida de las segundas derivadas de $f(\mathbf{x})$ con respecto a las variables independientes. La dirección de búsqueda $\mathbf{d}$ para el método de Newton se elige como sigue: se reemplaza $\mathbf{x}$ por $\mathbf{x}_{k+1}$ y se define

$$\mathbf{d}_k \equiv \mathbf{x}_{k+1} - \mathbf{x}_k \,, \quad (3.58)$$

por lo que

$$f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k) + \mathbf{d}_k^T \nabla f(\mathbf{x}_k) + \frac{1}{2}\mathbf{d}_k^T \mathbf{H}(\mathbf{x}_k)\mathbf{d}_k \quad (3.59)$$

Ahora, diferenciando $f(\mathbf{x})$ con respecto de $\mathbf{d}_k$ e igualando a cero se obtiene el mínimo de $f(\mathbf{x})$; luego se define $\nabla' \equiv \frac{\partial}{\partial \mathbf{d}_k}$ y se deriva la ecuación (3.59) con respecto a $\mathbf{d}_k$ y entonces

$$\nabla' f(\mathbf{x}_{k+1}) = \nabla' f(\mathbf{x}_k) + \nabla' \left(\mathbf{d}_k^T \nabla f(\mathbf{x}_k) + \frac{1}{2}\mathbf{d}_k^T \mathbf{H}(\mathbf{x}_k)\mathbf{d}_k \right)$$

luego

$$\mathbf{0} = \nabla f(\mathbf{x}_k) + \mathbf{H}(\mathbf{x}_k)\mathbf{d}_k$$

debido a que $\nabla' f(\mathbf{x}_{k+1}) = \nabla' f(\mathbf{x}_k) = \mathbf{0}$, no dependen de $\mathbf{d}_k$; por lo tanto,

$$\mathbf{d}_k = -\mathbf{H}\left(\mathbf{x}_k\right)^{-1} \nabla f\left(\mathbf{x}_k\right) \tag{3.60}$$

y

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}\left(\mathbf{x}_k\right)^{-1} \nabla f\left(\mathbf{x}_k\right) \tag{3.61}$$

Obsérvese que ambas, la dirección y la longitud de paso, están determinadas. Si $f\left(\mathbf{x}\right)$ es una función cuadrática, sólo se requiere de un paso para alcanzar el mínimo de $f\left(\mathbf{x}\right)$, y la longitud de paso si se usa una búsqueda lineal será igual a la unidad. Pero para una función objetivo no lineal general, el mínimo de $f\left(\mathbf{x}\right)$ no se alcanzará en un paso. El criterio para garantizar la convergencia en el método de Newton, suponiendo que la función es doblemente diferenciable, es que la inversa de la matriz hessiana de la función objetivo sea positiva definida para el caso de un mínimo y negativa definida en el de un máximo. El problema con este método es que en cada paso debe evaluarse e invertirse la matriz hessiana, lo cual puede ser muy complicado y laborioso, o bien que la inversa tenga problemas de estabilidad. Pero si el método converge, la rapidez de convergencia que lo caracteriza es de segundo orden.

Las desventajas del método de Newton para aplicaciones generales son las siguientes:

1) Se requiere la evaluación de derivadas de segundo orden en cada iteración, lo que por lo general consume mucho tiempo, y en algunas aplicaciones es posible que no puedan calcularse estas derivadas. Además de la evaluación de las segundas derivadas, también debe resolverse un sistema de ecuaciones lineales simultáneas; por lo tanto, cada iteración del método requiere sustancialmente más cálculos comparado con el método de Cauchy.

2) La matriz hessiana de la función objetivo puede resultar singular en algunas iteraciones y esto implica inestabilidad en el método, ya que no puede usarse para determinar la dirección de búsqueda. Asimismo, a menos que la hessiana sea positiva definida, la dirección de Newton no puede garantizar que sea de descenso para la función objetivo.

3) Como en el método de Cauchy, el método de Newton no usa la información generada del gradiente y la matriz hessiana en las iteraciones anteriores, así que cada iteración es independiente de las anteriores.

4) El método de Newton no es convergente a menos que la hessiana permanezca definida positiva y se use un esquema de búsqueda lineal para la determinación de la longitud de paso; sin embargo, el método tiene una rapidez de convergencia cuadrática y entonces, para una función cuadrática estrictamente convexa, el método convergerá en solo una iteración sin importar el punto inicial elegido.

Algoritmo 3.3 Método de Newton Modificado con búsqueda lineal:

Dada una función $f(\mathbf{x})$ continua y derivable, un punto inicial $\mathbf{x}_0$ y ε. Para $k = 0,1,2,...$ hasta donde se satisfaga, hacer:

1. $\mathbf{d}_k = -\mathbf{H}(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$ o bien resolver $\mathbf{H}(\mathbf{x}_k)\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$ y obtener $\mathrm{d_k}$.
2. Hallar α^* que minimiza $f(\mathbf{x}_k + \alpha\mathbf{d}_k)$.
3. Calcular $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha^*\mathbf{d}_k$.
4. Si $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \varepsilon$ o $\|\nabla f(\mathbf{x}_{k+1})\| \leq \varepsilon$, entonces

   Tomar $\mathbf{x}_{k+1} \approx \mathbf{x}^*$ y parar.
   Fin del si.
5. Tomar $k = k+1$ y regresar al paso 1.

Ejemplo 3.7 Halle el mínimo de $f(\mathbf{x}) = 3(x_1 - 1)^2 + 10(x_2 - 2x_1)^2$, por el método de Newton iniciando en $\mathbf{x}^T = (0,3)$, con $\varepsilon = 0.01$.

Solución:

Usando el algoritmo 3.3 el mínimo hallado es $\mathbf{x}^{*T} = (1.00, 2.00)$ donde $f(\mathbf{x}^*) = 0.0$ Obsérvese que la convergencia del método fue en la primera iteración, y el número total de iteraciones principales fue de $k = 1$. La figura 3.9 muestra en forma gráfica la búsqueda hacia el óptimo.

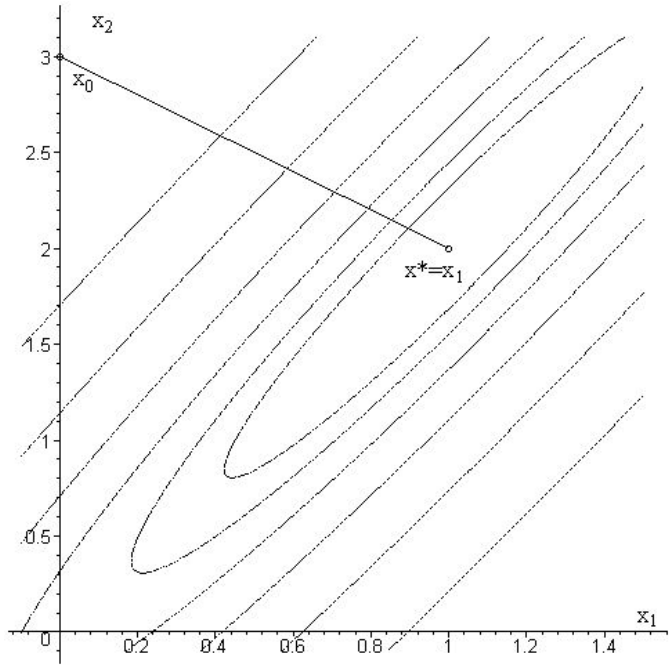


Figura 3.9 *Evolución del método de Newton en 1 iteración.*

Método de Fletcher y Reeves o de gradientes conjugados

El método de Fletcher y Reeves de gradientes conjugados trata de explotar el hecho de que para una función cuadrática de n variables, se requieren de n búsquedas lineales a lo largo de direcciones *mutuamente conjugadas* para localizar el mínimo $\mathbf{x}^*$ de la función; es decir, con esto se evitará el uso directo de la evaluación de la matriz hessiana $\mathbf{H}(\mathbf{x}_k)$. Para esto, en lugar de utilizar una fórmula analítica para α_k como en el método de dirección conjugada, se hallará β_k con una búsqueda lineal que minimice la función objetivo, y después la fórmula obtenida para α_k será independiente de la matriz hessiana, la cual se sustituye por productos entre gradientes.

Considérese el problema general de minimizar la función $f(\mathbf{x})$. Se explorará iniciar una búsqueda a lo largo de direcciones que sean mutuamente conjugadas con respecto a $\mathbf{H}(\mathbf{x})$. La primera dirección de búsqueda desde el primer punto $\mathbf{x}_0$ se hará en la dirección del descenso acelerado que parece razonable

$$\mathbf{d}_0 = -\nabla f_0 \tag{3.62}$$

donde $\nabla f_0 \equiv \nabla f\left(\mathbf{x}_0\right)$. Ahora bien, para hallar el valor de β_0 que minimice

$$\phi\left(\beta_0\right) = f\left(\mathbf{x}_0 + \beta_0 \mathbf{d}_0\right)$$

se usa cualquier técnica de búsqueda unidimensional (Newton-Raphson, sección áurea, etc.). Luego de obtener β_0 por alguna de éstas técnicas, se calcula

$$\mathbf{x}_1 = \mathbf{x}_0 + \beta_0 \mathbf{d}_0 \tag{3.63}$$

y se busca en una dirección $\mathbf{d}_1$ conjugada a $\mathbf{d}_0$, lo cual se hace eligiendo $\mathbf{d}_1$ como una combinación lineal de $\mathbf{d}_0$ y $-\nabla f_1$ de la siguiente manera:

$$\mathbf{d}_1 = -\nabla f_1 + \alpha_0 \mathbf{d}_0$$

Se busca de nuevo β_1 que minimice $\phi\left(\beta_1\right) = f\left(\mathbf{x}_1 + \beta_1 \mathbf{d}_1\right)$ y se calcula

$$\mathbf{x}_2 = \mathbf{x}_1 + \beta_1 \mathbf{d}_1 \tag{3.64}$$

y la dirección de búsqueda $\mathbf{d}_2$ desde $\mathbf{x}_2$ se elige conjugada a $\mathbf{d}_0$ y $\mathbf{d}_1$, etcétera.

En la $(k+1)$-ésima etapa se elige $\mathbf{d}_{k+1}$ como una combinación lineal de $-\nabla f_{k+1}$, $\mathbf{d}_0, \mathbf{d}_1, ..., \mathbf{d}_k$, esto es, conjugada a todas las direcciones $\mathbf{d}_0, \mathbf{d}_1, ..., \mathbf{d}_k$ Así,

$$\mathbf{d}_{k+1} = -\nabla f_{k+1} + \sum_{j=0}^{k} \alpha_j \mathbf{d}_j \quad , \quad k = 0,1,...$$

Este proceso deja ver que todas las α_j son nulas excepto para α_k y que

$$\mathbf{d}_{k+1} = -\nabla f_{k+1} + \alpha_k \mathbf{d}_k \tag{3.65}$$

y

$$\alpha_k = \frac{\nabla^2 f_{k+1}}{\nabla^2 f_k} = \frac{\nabla^T f_{k+1} \nabla f_{k+1}}{\nabla^T f_k \nabla f_k} \tag{3.66}$$

Las direcciones de búsqueda sucesivas en el método de Fletcher-Reeves son conjugadas y el método hallará el mínimo de una función cuadrática de n variables después de n búsquedas. Esto supone que la búsqueda lineal se realiza exactamente e ignora cualquier error de redondeo que se pueda producir. Por supuesto, el método se puede aplicar a funciones no cuadráticas, y se espera que alcance la propiedad de convergencia cuadrática cuando la aproximación cuadrática resulte válida. Fletcher y Reeves (1964) sugieren que en esta situación toda *n-ésima* dirección de búsqueda deberá ser a lo largo de la dirección de descenso acelerado y que deberá reiniciarse la construcción de las direcciones conjugadas. El siguiente algoritmo incluye esta idea. En general, se puede decir que el método es eficiente y robusto. Para funciones no cuadráticas, se obtiene mejor eficiencia usando la fórmula de Polak y Ribiere (1969) para α_k dada por la ecuación:

$$\alpha_k = \frac{\left(\nabla f_{k+1} - \nabla f_k\right)^T \nabla f_{k+1}}{\nabla^T f_k \nabla f_k} \tag{3.67}$$

Algoritmo 3.4 Método de Fletcher Reeves o gradientes conjugados:

Dada una función $f(\mathbf{x})$, un punto inicial $\mathbf{x}_0$ y la tolerancia ε.
Para $j = 0,1,...$ hasta donde se satisfaga, hacer:
Para $k = 0,1,...,n-1$ hacer;

1. $m = nj + k$.

2. Si $k = 0$, entonces

   $\alpha_m = 0$

   En otro caso,

   $$\alpha_m = \frac{\nabla^T f_m \nabla f_m}{\nabla^T f_{m-1} \nabla f_{m-1}}$$

   Fin del si.

3. $\mathbf{d}_m = -\nabla f_m + \alpha_m \mathbf{d}_{m-1}$.

4. Hallar β_m^* que minimiza $f(\mathbf{x}_m + \beta_m \mathbf{d}_m)$.

5. $\mathbf{x}_{m+1} = \mathbf{x}_m + \beta_m^* \mathbf{d}_m$.

6. Si $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \varepsilon$ o $\|\nabla f_{m+1}\| \leq \varepsilon$, entonces

   $\mathbf{x}_{m+1} = \mathbf{x}^*$ y terminar.

   Fin del si.

7. Si $k = n-1$, entonces

   $j = j+1$, $k = 0$ y regresar a 1.

   En otro caso,

   $k = k+1$ y regresar a 1.
   Fin del si.

Ejemplo 3.8 Halle el mínimo de $f(\mathbf{x}) = 3(x_1 - 1)^2 + 10(x_2 - 2x_1)^2$, por el método de Fletcher y Reeves iniciando en $\mathbf{x}^T = (0,3)$, con $\varepsilon = 0.01$.

Solución:

El mínimo hallado con la precisión dada es $\mathbf{x}^{*T} = (1.00, 2.00)$, donde $f(\mathbf{x}^*) = 0.0$. Observe que la convergencia del método fue rápida debido al uso de la búsqueda lineal exacta, el número total de iteraciones menores fue de $k = 2$, como se esperaba pues la función es cuadrática. La figura 3.10 muestra en forma gráfica la búsqueda hacia el óptimo.

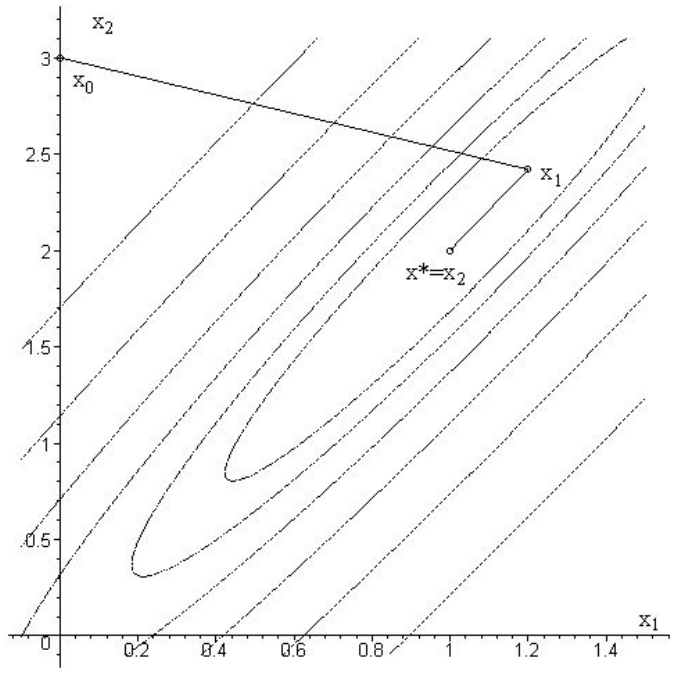


Figura 3.10 *Evolución del método de gradientes conjugados en 2 iteraciones.*

Método de Davidon, Fletcher y Powell (DFP) o de métrica variable

El método DFP está basado en la ecuación de Newton

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}\left(\mathbf{x}_k\right)^{-1} \nabla f\left(\mathbf{x}_k\right)$$

con la diferencia de que el método DFP busca evadir el cálculo de la inversa de la matriz hessiana $\mathbf{H}\left(\mathbf{x}_k\right)^{-1}$ en cada paso, tomando la dirección de búsqueda en cada etapa k como $-\mathbf{G}_k \nabla f\left(\mathbf{x}_k\right)$, donde $\mathbf{G}_k$ es una matriz simétrica positiva definida, la cual se *construye* en cada etapa como será explicado más adelante.

Se inicia con un punto $\mathbf{x}_0$ y una matriz simétrica positiva definida $\mathbf{G}_0$, por lo general la matriz unidad. Por conveniencia se escribirá $\nabla f_i \equiv \nabla f\left(\mathbf{x}_i\right)$, el procedimiento iterativo que describiremos brevemente será formalizado en el siguiente algoritmo más adelante.

1. En la etapa k se tiene un punto $\mathbf{x}_k$ y una matriz simétrica positiva definida $\mathbf{G}_k$.

2. Se toma la dirección de búsqueda como

$$\mathbf{d}_k = -\mathbf{G}_k \nabla f_k \tag{3.68}$$

3. Se realiza una búsqueda lineal sobre la línea $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$ para hallar el valor de α_k^* que minimice $\phi\left(\alpha_k\right) = f\left(\mathbf{x}_k + \alpha_k \mathbf{d}_k\right)$.

4. Se determina

$$\mathbf{v}_k = \alpha_k \mathbf{d}_k \tag{3.69}$$

5. Se calcula

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}_k \tag{3.70}$$

6. Se determina $f_{k+1} = f\left(\mathbf{x}_{k+1}\right)$, ∇f_{k+1}, y se termina el procedimiento si $\left\|\nabla f_{k+1}\right\|$ o $\left\|\mathbf{v}_k\right\|$ son suficientemente pequeños; en caso contrario, se procede con el paso 7. Recuérdese que

$$\nabla^T f_{k+1}\mathbf{v}_k = 0 \tag{3.71}$$

7. Se determina

$$\mathbf{u}_k = \nabla f_{k+1} - \nabla f_k \tag{3.72}$$

8. Se construye

$$\mathbf{G}_{k+1} = \mathbf{G}_k + \boldsymbol{A}_k + \boldsymbol{B}_k \tag{3.73}$$

donde

$$\boldsymbol{A}_k = \frac{\mathbf{v}_k\mathbf{v}_k^T}{\mathbf{v}_k^T\mathbf{v}_k} \tag{3.74}$$

y

$$\boldsymbol{B}_k = -\frac{\mathbf{G}_k\mathbf{u}_k\mathbf{u}_k^T\mathbf{G}_k}{\mathbf{u}_k^T\mathbf{G}_k\mathbf{u}_k} \tag{3.75}$$

9. Se toma $k = k+1$ y se regresa al paso 2.

Se justificará el procedimiento siguiendo los argumentos de Fletcher y Powell. El proceso es estable si $\mathbf{v}_k$ va cuesta abajo para un mínimo y α_k es positiva.

Ya que ∇f_k es la dirección del ascenso acelerado, $\mathbf{v}_k$ irá cuesta abajo si y solo si

$$-\mathbf{v}_k^T \nabla f_k = -\nabla f_k^T \mathbf{v}_k = \alpha_k \nabla f_k^T \mathbf{G}_k \nabla f_k > 0 \tag{3.76}$$

esto será así si $\mathbf{G}_k$ es simétrica positiva definida. La construcción en las ecuaciones (3.73), (3.74) y (3.75) mantiene esta simetría. Si el método DFP se aplica a la función cuadrática $f(\mathbf{x}) = a + \mathbf{x}^T\mathbf{b} + \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$ con $\mathbf{H}$ simétrica positiva definida, entonces $\mathbf{G}_n = \mathbf{H}^{-1}$ y el proceso terminará después de n etapas. Se deja al lector mostrar que $\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_k$ son eigenvectores linealmente independientes de $\mathbf{G}_{k+1}\mathbf{H}$ con eigenvalor 1. De esta manera, $\mathbf{G}_n\mathbf{H}$ debe ser la matriz unitaria. De la ecuación (3.72) observe que

$$\mathbf{u}_k = \nabla f_{k+1} - \nabla f_k = \mathbf{H}\mathbf{x}_{k+1} + \mathbf{b} - \mathbf{H}\mathbf{x}_k - \mathbf{b} = \mathbf{H}\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right)$$

usando la ecuación (3.70) resulta

$$\mathbf{u}_k = \mathbf{H}\mathbf{v}_k \tag{3.77}$$

También $\mathbf{G}_{k+1}\mathbf{H}\mathbf{v}_k = \mathbf{G}_{k+1}\mathbf{u}_k = \mathbf{G}_k\mathbf{u}_k + \mathbf{A}_k\mathbf{u}_k + \mathbf{B}_k\mathbf{u}_k$, luego

$$\mathbf{G}_{k+1}\mathbf{H}\mathbf{v}_k = \mathbf{G}_k\mathbf{u}_k + \frac{\mathbf{v}_k\mathbf{v}_k^T\mathbf{u}_k}{\mathbf{v}_k^T\mathbf{u}_k} - \frac{\mathbf{G}_k\mathbf{u}_k\mathbf{u}_k^T\mathbf{G}_k\mathbf{u}_k}{\mathbf{u}_k^T\mathbf{G}_k\mathbf{u}_k} = \mathbf{G}_k\mathbf{u}_k + \mathbf{v}_k - \mathbf{G}_k\mathbf{u}_k$$

observando que $\mathbf{v}_k^T\mathbf{u}_k$ y $\mathbf{u}_k^T\mathbf{G}_k\mathbf{u}_k$ son escalares que pueden cancelarse. Así

$$\mathbf{G}_{k+1}\mathbf{H}\mathbf{v}_k = \mathbf{v}_k \tag{3.78}$$

Se deja al lector mostrar que para $i = 2,3,...,n$

$$\mathbf{v}_k^T \mathbf{H} \mathbf{v}_j = 0 \quad , \; 0 \leq k < j < i \tag{3.79}$$

que

$$\mathbf{G}_i \mathbf{H} \mathbf{v}_k = \mathbf{v}_k \quad , \quad 0 \leq k < i \tag{3.80}$$

y que

$$\mathbf{v}_k^T \nabla f_i = 0 \; , \; 0 \leq k \leq i \tag{3.81}$$

Esto se hace por inducción sobre i . La ecuación (3.79) muestra que $\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_{n-1}$ son linealmente independientes y que son mutuamente conjugados con respecto a $\mathbf{H}$. Si $i = n$ entonces, de la ecuación (3.80),

$$\mathbf{G}_n \mathbf{H} \mathbf{v}_k = \mathbf{v}_k$$

que muestra que $\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_{n-1}$ son eigenvectores de $\mathbf{G}_n \mathbf{H}$ con eigenvalor 1 y además que $\mathbf{G}_n \mathbf{H}$ debe ser una matriz unitaria, es decir, $\mathbf{G}_n \mathbf{H} = \mathbf{I}$, por tanto

$$\mathbf{G}_n = \mathbf{H}^{-1}$$

Que el mínimo se encuentra en n iteraciones se sigue de la ecuación (3.81) ya que $\mathbf{v}_k^T \nabla f_n = 0$, como $\mathbf{v}_k^T \neq \mathbf{0}$, $\nabla f_n = \mathbf{0}$, por tanto

$$\mathbf{x}_n = \mathbf{x}^* = -\mathbf{H}^{-1}\mathbf{b} = -\mathbf{G}_n \mathbf{b} \tag{3.82}$$

La construcción de la matriz $\mathbf{G}$ se sigue de la ecuación (3.73):

$$\begin{aligned}\mathbf{G}_n &= \mathbf{G}_{n-1} + \mathbf{A}_{n-1} + \mathbf{B}_{n-1} \\ &= \mathbf{G}_{n-2} + \mathbf{A}_{n-2} + \mathbf{B}_{n-2} + \mathbf{A}_{n-1} + \mathbf{B}_{n-1} \\ &= \cdots = \mathbf{G}_0 + \sum_{k=0}^{n-1} \left(\mathbf{A}_k + \mathbf{B}_k \right)\end{aligned}$$

Se deja al lector mostrar que $\mathbf{H}^{-1} = \sum_{k=0}^{n-1} \mathbf{A}_k$. Como la ecuación (3.78) debe ser válida:

$$\mathbf{G}_{k+1}\mathbf{H}\mathbf{v}_k = \mathbf{v}_k \quad \Rightarrow \quad \mathbf{v}_k = \mathbf{G}_k\mathbf{H}\mathbf{v}_k + \mathbf{A}_k\mathbf{H}\mathbf{v}_k + \mathbf{B}_k\mathbf{H}\mathbf{v}_k$$

usando la ecuación (3.77)

$$\mathbf{A}_k\mathbf{H}\mathbf{v}_k = \mathbf{A}_k\mathbf{u}_k = \frac{\mathbf{v}_k\mathbf{v}_k^T\mathbf{u}_k}{\mathbf{v}_k^T\mathbf{u}_k} = \mathbf{v}_k$$

entonces

$$\mathbf{B}_k\mathbf{u}_k = \mathbf{v}_k - \mathbf{v}_k - \mathbf{G}_k\mathbf{u}_k = -\mathbf{G}_k\mathbf{u}_k$$

y, por lo tanto,

$$\mathbf{B}_k\mathbf{u}_k = -\mathbf{G}_k\mathbf{u}_k \tag{3.83}$$

luego, una forma simple (aunque no necesariamente la única) para $\mathbf{B}_k$ es

$$\mathbf{B}_k\mathbf{u}_k\mathbf{z}_k^T\mathbf{z}_k = -\mathbf{G}_k\mathbf{u}_k\mathbf{z}_k^T\mathbf{z}_k \quad \Rightarrow \quad \mathbf{B}_k\mathbf{z}_k\mathbf{u}_k^T\mathbf{z}_k = -\mathbf{G}_k\mathbf{u}_k\mathbf{z}_k^T\mathbf{z}_k$$

$$\mathbf{B}_k\mathbf{z}_k = -\frac{\mathbf{G}_k\mathbf{u}_k\mathbf{z}_k^T}{\mathbf{u}_k^T\mathbf{z}_k}\mathbf{z}_k$$

y entonces

$$\mathbf{B}_k = -\frac{\mathbf{G}_k\mathbf{u}_k\mathbf{z}_k^T}{\mathbf{u}_k^T\mathbf{z}_k}$$

donde $\mathbf{z}$ es un vector arbitrario. Ya que $\mathbf{B}_k$ debe ser simétrica, una buena elección para $\mathbf{z}$ es

$$\mathbf{z} = \mathbf{G}_k\mathbf{u}_k$$

por lo que

$$\mathbf{B}_k = -\frac{\mathbf{G}_k\mathbf{u}_k\mathbf{u}_k^T\mathbf{G}_k}{\mathbf{u}_k^T\mathbf{G}_k\mathbf{u}_k} \tag{3.84}$$

Lo cual termina la teoría del método DFP.

Este método usa las ideas del método de Newton y de las direcciones conjugadas. Cuando se aplica a una función cuadrática de n variables, éste converge en n iteraciones. En general, el procedimiento es robusto y eficiente para funciones cuadráticas o no.

Algoritmo 3.5 Método de Davidon Fletcher Powell (DFP):

Dada una función $f(\mathbf{x})$, un punto inicial $\mathbf{x}_0$, una matriz simétrica positiva definida $\mathbf{G}_0$ y la tolerancia ε.

Para $k = 0,1,...$ hasta donde se satisfaga, hacer:

1. $\mathbf{d}_k = -\mathbf{G}_k \nabla f_k$.
2. Halla α_k^* que minimiza $f(\mathbf{x}_k + \alpha_k \mathbf{d}_k)$.
3. $\mathbf{v}_k = \alpha_k^* \mathbf{d}_k$.
4. $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}_k$.
5. $\mathbf{u}_k = \nabla f_{k+1} - \nabla f_k$.
6. $\mathbf{A}_k = \dfrac{\mathbf{v}_k \mathbf{v}_k^T}{\mathbf{v}_k^T \mathbf{u}_k}$.
7. $\mathbf{B}_k = -\dfrac{\mathbf{G}_k \mathbf{u}_k \mathbf{u}_k^T \mathbf{G}_k}{\mathbf{u}_k^T \mathbf{G}_k \mathbf{u}_k}$.
8. $\mathbf{G}_{k+1} = \mathbf{G}_k + \mathbf{A}_k + \mathbf{B}_k$.
9. Si $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \varepsilon$ o $\|\nabla f_{k+1}\| \leq \varepsilon$, entonces

   $\mathbf{x}_{k+1} = \mathbf{x}^*$ y terminar.

   Fin del si.
10. Tomar $k = k+1$ y regresar a 1.

Ejemplo 3.9 Halle el mínimo de $f(\mathbf{x}) = 3(x_1 - 1)^2 + 10(x_2 - 2x_1)^2$, por el método DFP iniciando en $\mathbf{x}^T = (0,3)$, con $\varepsilon = 0.01$.

Solución:

El mínimo hallado con la precisión dada es $\mathbf{x}^{*T} = (1.00, 2.00)$, donde $f(\mathbf{x}^*) = 0.0$. En este caso también la convergencia del método fue rápida debido al uso de la búsqueda lineal exacta, el número total de iteraciones principales fue de $k = 2$, como se esperaba pues la función es cuadrática. La figura 3.11 muestra en forma gráfica la búsqueda hacia el óptimo.

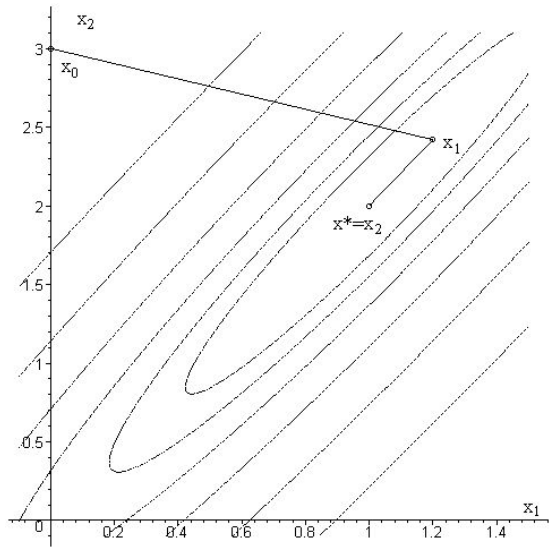


Figura 3.11 *Evolución del método DFP en 2 iteraciones.*

Método de Broyden–Fletcher–Goldfarb-Shanno (BFGS)

Las fórmulas de recurrencia para la inversa de la hessiana consideradas en la sección anterior se basan en la satisfacción de la ecuación (3.80)

$$\mathbf{G}_i \mathbf{u}_k = \mathbf{v}_k, \ 0 \leq k < i \tag{3.85}$$

deducida de la ecuación (3.77)

$$\mathbf{u}_k = \mathbf{H}\mathbf{v}_k \ , \ \ 0 \leq k < i \tag{3.86}$$

que se cumpliría sólo en el caso cuadrático puro. También se pueden actualizar aproximaciones a la propia hessiana $\mathbf{H}$, en lugar de su inversa. De manera análoga, se buscaría satisfacer

$$\mathbf{v}_k = \mathbf{G}_k^{'}\mathbf{u}_k \ , \ 0 \leq k < i \tag{3.87}$$

La ecuación tiene exactamente la misma forma que la ecuación (3.85), excepto que $\mathbf{u}_k$ y $\mathbf{v}_k$ se intercambian y $\mathbf{G}_i$ se sustituye por $\mathbf{G}_i$. Esta última actualización se conoce como método de Broyden–Fletcher–Goldfarb-Shanno (BFGS).

Entonces la fórmula complementaria correspondiente al método DFP es

$$\mathbf{G}_{k+1} = \mathbf{G}_k + \frac{\mathbf{u}_k\mathbf{u}_k^T}{\mathbf{u}_k^T\mathbf{v}_k} - \frac{\mathbf{G}_k\mathbf{v}_k\mathbf{v}_k^T\mathbf{G}_k}{\mathbf{v}_k^T\mathbf{G}_k\mathbf{v}_k} \quad \text{(BFGS).} \tag{3.88}$$

Los experimentos numéricos han indicado que este método es superior al DFP, razón por la cual en la actualidad se le prefiere.

Algoritmo 3.6 Método de Broyden Fletcher Goldfarb Shanno (BFGS):

Dada una función $f(\mathbf{x})$, un punto inicial $\mathbf{x}_0$, una matriz simétrica positiva definida $\mathbf{G}_0$ y la tolerancia ε.

Para $k = 0,1,...$ hasta donde se satisfaga, hacer:

1. Resolver $\mathbf{G}_k\mathbf{d}_k = -\nabla f_k$ para obtener $\mathbf{d}_k$.
2. Hallar α_k^* que minimiza $f(\mathbf{x}_k + \alpha_k\mathbf{d}_k)$.
3. $\mathbf{v}_k = \alpha_k^*\mathbf{d}_k$.
4. $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}_k$.
5. $\mathbf{u}_k = \nabla f_{k+1} - \nabla f_k$.
6. $\mathbf{A}_k = \dfrac{\mathbf{u}_k\mathbf{u}_k^T}{\mathbf{u}_k^T\mathbf{v}_k}$.
7. $\mathbf{B}_k = -\dfrac{\mathbf{G}_k\mathbf{v}_k\mathbf{v}_k^T\mathbf{G}_k}{\mathbf{v}_k^T\mathbf{G}_k\mathbf{v}_k}$.
8. $\mathbf{G}_{k+1} = \mathbf{G}_k + \mathbf{A}_k + \mathbf{B}_k$.
9. Si $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \le \varepsilon$ o $\|\nabla f_{k+1}\| \le \varepsilon$, entonces

   $\mathbf{x}_{k+1} = \mathbf{x}^*$ y terminar.

   Fin del si.
10. Tomar $k = k+1$ y regresar a 1.

Ejemplo 3.10 Halle el mínimo de $f(\mathbf{x}) = 3(x_1 - 1)^2 + 10(x_2 - 2x_1)^2$, por el método BFGS iniciando en $\mathbf{x}^T = (0,3)$, con $\varepsilon = 0.01$.

Solución:

El mínimo hallado con la precisión dada es $\mathbf{x}^{*T} = (1.00, 2.00)$, donde $f(\mathbf{x}^*) = 0.0$. Este método también tuvo una convergencia rápida debido al uso de la búsqueda lineal exacta, el número total de iteraciones principales fue de $k = 2$, igual a los dos métodos anteriores como era de esperarse. La figura 3.12 muestra en forma gráfica la búsqueda hacia el óptimo.

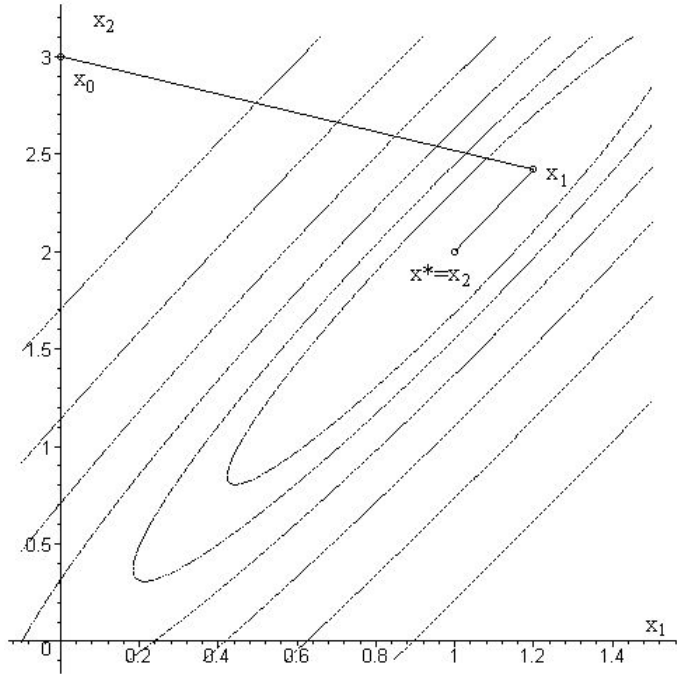


Figura 3.12 *Evolución del método BFGS en 2 iteraciones.*

Problemas

Para cada uno de los siguientes problemas encuentre la solución numéricamente con cualquiera de los métodos analizados y compárela con el valor de la solución exacta (analítica si existe).

3.1 Optimizar la ecuación siguiente, la cual se obtiene de un modelo de elemento finito para una viga volada sujeta a cargas y momentos (figura 1).

Figura 1

$$f(x,y) = 5x^2 - 5xy + 2.5y^2 - x - 1.5y$$

donde x = desplazamiento final, y y = momento final. Calcule los valores de x y y que minimizan $f(x,y)$.

3.2 Un recipiente cilíndrico abierto se utilizará para almacenar 10 m^3 de líquido. La función objetivo para la suma de los costos de capital y operación del recipiente es

$$f(h,r) = \frac{1}{\pi r^2 h} 2\pi rh + 10\pi r^2$$

¿Puede usarse el método de Newton para minimizar esta función? La solución es $[r^*, h^*]^T = [0.22, 2.16]^T$.

3.3 El costo de operación anual f para un sistema de línea eléctrica está dado por la expresión siguiente

$$f(V,C) = \frac{\left(21.9\times10^7\right)}{V^2 C} + \left(3.9\times10^6\right)C + \left(1.0\times10^3\right)V$$

donde V = voltaje en kilovolts (kV) y C = la conductancia en Siemens (S). Encuentre los puntos estacionarios para la función, y determine V y C para minimizar los costos de operación.

3.4 Una inyección de x_1 miligramos de cierto medicamento A y x_2 miligramos del medicamento B produce una respuesta de f unidades, y

$$f\left(x_1, x_2\right) = x_1^2 x_2^3 \left(c - x_1 - x_2\right)$$

donde $c > 0$ es una constante. ¿Qué dosis de cada medicamento ocasionarán la respuesta máxima?

3.5 Se construye una caja rectangular cerrada con un volumen de 16 m^3 empleando tres tipos de materiales. El costo del material para el fondo y la tapa es de \$0.18 por m^2, el costo del material para el frente y la parte trasera es de \$0.16 por m^2, y el costo del material para los otros dos lados es de \$0.12 por m^2. Calcule las dimensiones de la caja de modo que el costo de los materiales sea un mínimo.

3.6 Suponga que T grados es la temperatura en cualquier punto (x_1, x_2, x_3) de la esfera

$$x_1^2 + x_2^2 + x_3^2 = 4$$

y

$$T\left(\mathbf{x}\right) = 100 x_1 x_2^2 x_3$$

Obtenga los puntos de la esfera donde la temperatura es la máxima y también los puntos donde es mínima. Además, calcule la temperatura en estos puntos.

3.7 Suponga que t horas después de la inyección de x miligramos de adrenalina la respuesta es de f unidades, y

$$f\left(t, x\right) = te^{-t}\left(c - x\right)x$$

donde $c > 0$ es una constante.¿Qué valores de t y x producirán la respuesta máxima?

3.8 El potencial de una partícula en el plano xy se da por

$$V\left(x, y\right) = 2x^2 - 5xy + 3y^2 + 6x - 7y$$

Muestre que existe un punto y sólo uno en el cual la partícula permanece en equilibrio. Encuentre las coordenadas de ese punto.

Capítulo 4

Mínimos cuadrados

4.1 Introducción

Los datos que se obtienen mediante mediciones fluctúan, esto se debe a errores aleatorios del sistema de medición aplicado al comportamiento intrínsecamente estocástico del sistema en observación. Cualquiera que sea la razón, es frecuente que surja la necesidad de ajustar una función a los datos de la medición. Por ejemplo, un investigador podría intentar desarrollar una fórmula empírica para el sistema en observación, o bien un economista desearía ajustar una curva a una tendencia económica actual para poder predecir el futuro. Esto se hace para determinar el comportamiento general de los datos, por ejemplo para ver si un crecimiento es exponencial o para evaluar la función en puntos diferentes de los datos.

La curva de ajuste puede ser una línea, un polinomio de grado n o una función logarítmica, exponencial, cosenoidal, o de algún otro tipo. La curva adecuada se escoge dependiendo de la distribución de los datos experimentales, de manera tal que se minimice la suma de los cuadrados de los errores (método conocido como mínimos cuadrados).

4.2 Formulación del problema de regresión lineal o múltiple

Considérese que los datos experimentales consisten de n observaciones en una variable de respuesta o dependiente $\mathbf{y}$ y en k variables explicativas o independientes $\mathbf{x}$.

Las observaciones suelen registrarse como sigue:

Datos	y	x_1	x_2	...	x_{k-1}	x_k
1	y_1	x_{11}	x_{12}	...	x_{1k-1}	x_{1k}
2	y_2	x_{21}	x_{22}	...	x_{2k-1}	x_{2k}
3	y_3	x_{31}	x_{32}	...	x_{3k-1}	x_{3k}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	y_n	x_{n1}	x_{n2}	...	x_{nk-1}	x_{nk}

Como hay n observaciones, el modelo de regresión se puede resumir escribiendo una serie de n ecuaciones, como:

$$\begin{aligned}
y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\
y_3 &= \beta_1 x_{31} + \beta_2 x_{32} + \beta_3 x_{33} + \cdots + \beta_k x_{3k} + \varepsilon_3 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \cdots + \beta_k x_{nk} + \varepsilon_n
\end{aligned} \qquad (4.1)$$

donde $x_{i1} = 1$ para todo i, es decir, $\mathbf{x}_1^T = (x_{11}, x_{21}, \ldots, x_{n1}) = (1,1,\ldots,1)$.

La formulación matricial correspondiente del modelo es

$$\mathbf{y}=\mathbf{X}\,\boldsymbol{\beta}+\boldsymbol{\varepsilon} \qquad (4.2)$$

donde $\mathbf{y}$ es el vector de observaciones de orden n x1, $\mathbf{X}$ es la matriz de variables independientes de orden nxk, β es el vector de parámetros por estimar de orden n x1, y ε es el vector de errores de orden nx1.

En nuestra representación de la matriz $\mathbf{X}$, cada componente x_{ij} tiene dos subíndices: el primero denota el renglón apropiado de la observación y el segundo la columna apropiada de la variable independiente. Cada columna de $\mathbf{X}$ representa un vector x_j de n observaciones en una variable dada (j=1,2,3,…,k), con todas las observaciones asociadas con la intercepción igual a x_1^T=(1,1,…,1) .

A continuación vamos a establecer las suposiciones fundamentales acerca del modelo de regresión lineal general como sigue:

1) La descripción del modelo está dada por la ecuación (4.1).

2) Los elementos de $\mathbf{X}$ son variables fijas (no aleatorias) y tienen varianza finita.

3) $\mathbf{X}$ tiene rango $k \leq n$.

4) $\mathbf{y}$ y $\boldsymbol{\varepsilon}$ son variables aleatorias, $\boldsymbol{\varepsilon}$ está normalmente distribuida con

i) $E(\boldsymbol{\varepsilon})=\mathbf{0}$

ii) $Var(\boldsymbol{\varepsilon}) = E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right) = \sigma^2\mathbf{I}$

donde $\mathbf{I}$ es la matriz identidad de orden $n\,x\,n$

La primera suposición significa que el modelo explica todas las observaciones hasta cierta precisión. La segunda suposición implica que las variables independientes están libres de errores. La tercera suposición de que $\mathbf{X}$ tiene rango k garantiza que no está presente la colinealidad perfecta, lo que implica que una de las columnas de $\mathbf{X}$ debería ser una combinación lineal de las restantes columnas y el rango de $\mathbf{X}$ debe ser menor que k .

Las suposiciones acerca del error son muy sólidas, ya que garantizan las propiedades aritméticas y estadísticas del proceso de estimación ordinario de los mínimos cuadrados. Además de la normalidad, se supone que cada término de error tiene media cero,

$$E(\varepsilon) = E\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix} = \begin{pmatrix}E(\varepsilon_1)\\ E(\varepsilon_2)\\ \vdots\\ E(\varepsilon_n)\end{pmatrix} = \begin{pmatrix}0\\ 0\\ \vdots\\ 0\end{pmatrix} = \mathbf{0}$$

La matriz de varianza-covarianza $\sigma^2\mathbf{I}$ se expresa como sigue;

$$E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right) = Var(\boldsymbol{\varepsilon}) = \begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1,\varepsilon_2) & \cdots & Cov(\varepsilon_1,\varepsilon_n)\\ Cov(\varepsilon_2,\varepsilon_1) & Var(\varepsilon_2) & \cdots & Cov(\varepsilon_2,\varepsilon_n)\\ \vdots & \vdots & \ddots & \vdots\\ Cov(\varepsilon_n,\varepsilon_1) & Cov(\varepsilon_n,\varepsilon_2) & \cdots & Var(\varepsilon_n)\end{bmatrix}$$

es decir,

$$E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right) = Var(\boldsymbol{\varepsilon}) = \begin{bmatrix}\sigma^2 & 0 & \cdots & 0\\ 0 & \sigma^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma^2\end{bmatrix} = \sigma^2\begin{bmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{bmatrix} = \sigma^2\mathbf{I}$$

$$E\left(\varepsilon\,\varepsilon^{\mathrm{T}}\right) = Var(\varepsilon) = \sigma^2\mathbf{I} \tag{4.3}$$

todas las varianzas son constantes, $Var(\varepsilon_i) = \sigma^2$, y todas las covarianzas son nulas, $Cov(\varepsilon_i, \varepsilon_j) = Cov(\varepsilon_j, \varepsilon_i) = 0$, es decir, no hay correlación entre residuales diferentes.

4.3 Estimación por mínimos cuadrados lineales

El objetivo es hallar un vector de parámetros $\hat{\boldsymbol{\beta}}$ que minimice la Suma de Cuadrados de los Errores (SCE) ≡S($\hat{\beta}$) ;

es decir, la función objetivo por minimizar es

$$S(\hat{\beta}) = \hat{\varepsilon}^T \hat{\varepsilon} = \sum_{i=1}^{n} \varepsilon_i^2 \tag{4.4}$$

donde

$$\hat{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} \tag{4.5}$$

y

$$\hat{\mathrm{y}} = \mathbf{X}\hat{\beta} \tag{4.6}$$

$\hat{\varepsilon}$ es el vector de residuales de $n x 1$, mientras $\hat{\mathbf{y}}$ es el vector de valores ajustados o estimados de $\mathbf{y}$ de $n x 1$. Sustituyendo las ecuaciones (4.5) y (4.6) en (4.4), se obtiene

$$\hat{\varepsilon}^T \hat{\varepsilon} = \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)^T \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right) = \mathbf{y}^T\mathbf{y} - 2\hat{\beta}^T\mathbf{X}^T\mathbf{y} + \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta}$$

debido a que

$$\hat{\beta}^T\mathbf{X}^T\mathbf{y} = \mathbf{y}^T\mathbf{X}\hat{\beta}$$

por lo tanto,

$$\hat{\varepsilon}^T\hat{\varepsilon} = \mathbf{y}^T\mathbf{y} - 2\hat{\beta}^T\mathbf{X}^T\mathbf{y} + \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} \tag{4.7}$$

Para determinar los estimadores por mínimos cuadrados, se deriva $S\left(\hat{\beta}\right)$ con respecto a $\hat{\beta}$ y se iguala a cero como sigue:

$$\nabla_{\hat{\beta}} S\left(\hat{\beta}\right) = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{0} \tag{4.8}$$

por lo que,

$$\boxed{\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}} \tag{4.9}$$

La matriz simétrica $\mathbf{X}^T\mathbf{X}$ se llama *matriz de sumas de cuadrados y productos cruzados* de las variables independientes $\mathbf{x}$, y está garantizado que va a tener inversa debido a la suposición de que $\mathbf{X}$ tiene rango k; el vector $\mathbf{X}^T\mathbf{y}$ se llama *vector de sumas de productos cruzados* de las $\mathbf{x}$ con las $\mathbf{y}$. Trasponiendo la ecuación (4.8), se tiene

$$\nabla_{\hat{\beta}}^T S\left(\hat{\beta}\right) = -2\mathbf{y}^T\mathbf{X} + 2\hat{\beta}^T\mathbf{X}^T\mathbf{X}$$

derivando de nuevo, se tiene

$$\nabla_{\hat{\beta}}\left(\nabla_{\hat{\beta}}^T S\left(\hat{\beta}\right)\right) \equiv \mathbf{H}\left(\hat{\beta}\right) = 2\mathbf{X}^T\mathbf{X}$$

y, por lo tanto,

$$\mathbf{H}(\hat{\beta}) = 2\mathbf{X}^T\mathbf{X} \tag{4.10}$$

es la matriz hessiana de $S\left(\hat{\beta}\right)$, de manera que

1) Si $\mathbf{H}$ es positiva definida, $\mathbf{X}^T\mathbf{X}$ también lo es y $\hat{\beta}$ minimiza a $S\left(\hat{\beta}\right)$.

2) Si $\mathbf{H}$ es negativa definida $\mathbf{X}^T\mathbf{X}$ también lo es y $\hat{\beta}$ maximiza a $S\left(\hat{\beta}\right)$.

En este caso, $\mathbf{H}$ resulta ser positiva definida y, por lo tanto, $\hat{\beta}$ minimiza a $S\left(\hat{\beta}\right)$

Estimación de σ^2

Para calcular la matriz de varianza-covarianza de los parámetros estimados, se necesita determinar una estimación para el escalar σ^2. Una elección natural es

$$s^2 \equiv \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-k} \tag{4.11}$$

donde s^2 es un estimador insesgado de σ^2.

Estimación de R^2

La variación total en $\mathbf{y}$ puede partirse en dos porciones: una representa la variación explicada y la segunda, la no explicada. Primero, se asumirá que la variable $\mathbf{y}$ tiene media nula ($\overline{y} = 0$). En notación matricial, la derivación se sigue del hecho de que el vector $\mathbf{y}$ puede escribirse como la suma de sus valores predichos

$$\hat{\mathrm{y}} = \mathbf{X}\hat{\beta}$$

y el vector residual $\hat{\varepsilon}$:

$$\mathrm{y} = \mathbf{X}\hat{\beta} + \hat{\varepsilon} \tag{4.12}$$

luego,

$$\mathbf{y}^T\mathbf{y} = \left(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}\right)^T \left(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}\right) = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} \tag{4.13}$$

ya que $\hat{\boldsymbol{\varepsilon}}^T\mathbf{X} = \mathbf{X}^T\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$. Si tiene media no nula ($\overline{y} \neq 0$), entonces la ecuación anterior se puede modificar por la siguiente expresión:

$$\mathbf{y}^T\mathbf{y} - n\overline{y}^2 = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - n\overline{y}^2 + \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} \tag{4.14}$$

o bien

$$SCT \equiv SCR + SCE \tag{4.15}$$

donde $SCT =$ suma de cuadrados total = $\mathbf{y}^T\mathbf{y} - n\overline{y}^2$, $SCR =$ suma de cuadrados explicada por la regresión = $\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - n\overline{y}^2$ y $SCE =$ suma de cuadrados (no explicada) de los errores = $\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}$. El *coeficiente de correlación* $\overline{R}^2$ se define como

$$\overline{R}^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT} \tag{4.16}$$

En este sentido, $\overline{R}^2$ mide la proporción de la variación en $\mathbf{y}$ que es explicada por la ecuación de regresión lineal general; por lo tanto,

$$\overline{R}^2 = \frac{\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - n\overline{y}^2}{\mathbf{y}^T\mathbf{y} - n\overline{y}^2} = 1 - \frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{\mathbf{y}^T\mathbf{y} - n\overline{y}^2} \tag{4.17}$$

Este coeficiente de correlación es sesgado; por tanto, la corrección para que sea insesgado debe tomar en cuenta el número de grados de libertad de la ecuación de regresión, y entonces

$$R^2 = 1 - \frac{\dfrac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{n-k}}{\dfrac{\mathbf{y}^T\mathbf{y} - n\bar{y}^2}{n-1}} = 1 - \frac{n-1}{n-k}\frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{\mathbf{y}^T\mathbf{y} - n\bar{y}^2} \tag{4.18}$$

Ejemplo 4.1 Ajuste el modelo $y = \beta_1 + \beta_2 x$ a los datos dados en la tabla 4.1

n	y	x
1	0	0
2	2	1
3	4	2
4	6	3
5	8	4
6	10	5

Tabla 4.1

Solución:

Se identifican los vectores y las matrices necesarias:

$$\mathbf{y} = \begin{pmatrix} 0 \\ 2 \\ 4 \\ 6 \\ 8 \\ 10 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{pmatrix}, \quad \mathbf{X}^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

aplicando la ecuación (4.9), resulta

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$$

y el modelo que describe al conjunto de datos está dado por

$$y = 2x$$

como era de esperar con sólo observar la tabla de datos. La matriz hessiana esta dada por la ecuación (4.10):

$$\mathbf{H}\left(\hat{\boldsymbol{\beta}}\right) = 2\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 12 & 30 \\ 30 & 110 \end{pmatrix}$$

cuyos determinantes de los menores principales son

$$D_1 = 12 > 0 \quad , \quad .D_2 = 420 > 0$$

Como ambos son positivos, la matriz hessiana es positiva definida y $\hat{\boldsymbol{\beta}}$ minimiza la suma de cuadrados de los residuales. Con la información obtenida hasta ahora, se pueden obtener los vectores $\hat{\mathbf{y}}$ de valores ajustados y $\hat{\boldsymbol{\varepsilon}}$ de residuales dados por las ecuaciones (4.6) y (4.5) respectivamente:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} 0 & 2 & 4 & 6 & 8 & 10 \end{pmatrix}^T, \ \hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T$$

de manera que la función suma de cuadrados de los errores, ecuación (4.4), da

$$S\left(\hat{\boldsymbol{\beta}}\right)=\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}=\sum_{i=1}^{n}\varepsilon_i^2=0$$

La varianza s^2 dada por la ecuación (4.11) es

$$s^2=\frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{n-k}=\frac{0}{6-2}=0$$

y el resto de estimadores en los que interviene s^2 (o σ^2) se anulan, por lo que ya no se continuarán los cálculos. Finalmente, el valor del coeficiente de correlación, ecuación (4.18), da

$$R^2=1-\frac{n-1}{n-k}\frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{\mathbf{y}^T\mathbf{y}-n\overline{y}^2}=1-\frac{5}{4}\frac{0}{70}=1$$

indicando con este valor una correlación perfecta entre los datos y el modelo ajustado, como era de esperar para un caso ideal.

Ejemplo 4.2 Se tomaron diez datos de un experimento en el que la variable independiente representa el porcentaje molar de un reactante y la variable dependiente a la producción porcentual, dados en la tabla 4.2.

n	y	x
1	73	20
2	78	20
3	85	30
4	90	40
5	91	40
6	87	50
7	86	50
8	91	50
9	75	60
10	65	70

Tabla 4.2

Ajuste un modelo cuadrático con estos datos y determine el valor de x que maximiza la producción.

Solución:

El modelo puede representarse como cuadrático de una sola variable o como de regresión lineal múltiple de dos variables, es decir;

$$y = \beta_1 + \beta_2 x + \beta_3 x^2$$

o bien

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3$$

donde $x_2 = x$ y $x_3 = x^2$. Aplicando la ecuación (4.9) resulta

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 35.6574 \\ 2.6314 \\ -0.0319 \end{pmatrix}$$

y el modelo cuadrático queda como

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 = 35.6574 + 2.6314x - 0.0319x^2$$

Para obtener la máxima producción, se deriva esta ecuación y se iguala a cero para determinar el valor de x que optimiza a y,

$$x^* = -\frac{\beta_2}{2\beta_3} = -\frac{2.6314}{2(-0.0319)} = 41.2445$$

la producción predicha en el óptimo será

$$y_{máx} = 89.9228$$

y el resultado esta de acuerdo con el valor estimado de una gráfica de los datos mostrada en la figura 4.1.

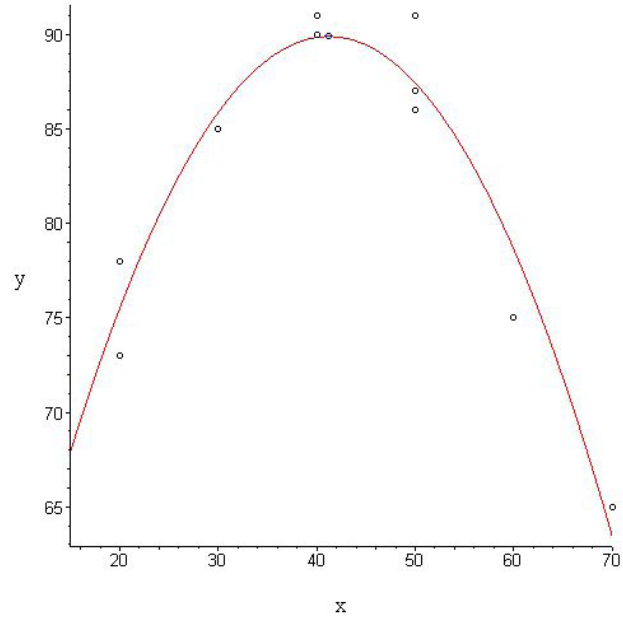


Figura 4.1 *Puntos experimentales y curva ajustada.*

El objetivo se alcanzó pero se pueden obtener más resultados acerca del problema. En este caso, la matriz hessiana es

$$\mathbf{H}\left(\hat{\boldsymbol{\beta}}\right) = 2\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 20 & 860 & 41800 \\ 860 & 41800 & 2210000 \\ 41800 & 2210000 & 12394000 \end{pmatrix}$$

por la ecuación (4.10), los determinantes de los menores principales son

$$D_1 = 20 > 0 \quad , \quad D_2 = 96400 > 0 \quad \text{y} \quad D_3 = 121344000000 > 0$$

por tanto, la matriz es positiva definida, indicando que $\hat{\boldsymbol{\beta}}$ es un mínimo, debido a que $\beta_3 = -0.0319$ es negativo x^* representa un máximo cuya producción da el valor de $y_{máx} = 89.9$

Los vectores de valores ajustados y residuales son

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} 75.51 \\ 75.51 \\ 85.87 \\ 89.84 \\ 89.84 \\ 87.43 \\ 87.43 \\ 87.43 \\ 78.63 \\ 63.45 \end{pmatrix} \quad , \quad \hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} -2.51 \\ 2.48 \\ -0.87 \\ 0.15 \\ 1.15 \\ -0.43 \\ -1.43 \\ 3.56 \\ -3.63 \\ 1.54 \end{pmatrix}$$

por las ecuaciones (4.6) y (4.5), observe que $\sum_i \varepsilon_i = 0.01$ con la precisión mostrada pero en realidad se anula con mayor precisión; la suma de cuadrados de los errores dada por la ecuación (4.4) resulta en

$$S\left(\hat{\boldsymbol{\beta}}\right) = \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} \varepsilon_i^2 = 45.19$$

y la varianza s^2 dada por la ecuación (4.11) es

$$s^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}}{n-k} = \frac{45.19}{10-3} = \frac{45.19}{7} = 6.45$$

Por otro lado,

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 82.1 \quad , \quad n\overline{y} = 821$$

$$\mathbf{y}^T\mathbf{y} = \sum_{i=1}^{n} y_i^2 = 68115 \quad , \quad \mathbf{y}^T\mathbf{y} - n\overline{y} = 67294$$

por tanto,

$$R^2 = 1 - \frac{n-1}{n-k}\frac{\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}}{\mathbf{y}^T\mathbf{y} - n\overline{y}^2} = 1 - \frac{9}{7}\frac{45.19}{67294} = 0.9991$$

lo que indica un buen acuerdo entre los datos y el modelo propuesto.

4.4 Aproximación por diferencias finitas de la matriz jacobiana

En esta sección se darán las fórmulas que pueden usarse en problemas en donde no esta disponible la información de las derivadas. La primera aproximación que debe considerarse es reemplazar la matriz jacobiana por una aproximación en diferencias finitas $\tilde{\mathbf{J}}$ igual que como se hizo en el capítulo 3 para el gradiente y la matriz hessiana. Usando diferencias hacia adelante se tiene

$$\tilde{\mathbf{j}}_i(\mathbf{x}) = \frac{\boldsymbol{f}(\mathbf{x} + \delta_i \mathbf{e}_i) - \boldsymbol{f}(\mathbf{x})}{\delta_i} \tag{4.19}$$

para diferencias hacia atrás se tiene

$$\tilde{\mathbf{j}}_i(\mathbf{x}) = \frac{\boldsymbol{f}(\mathbf{x}) - \boldsymbol{f}(\mathbf{x} - \delta_i \mathbf{e}_i)}{\delta_i} \tag{4.20}$$

para las diferencias centrales es

$$\tilde{\mathbf{j}}_i(\mathbf{x}) = \frac{\boldsymbol{f}(\mathbf{x} + \delta_i \mathbf{e}_i) - \boldsymbol{f}(\mathbf{x} - \delta_i \mathbf{e}_i)}{2\delta_i} \tag{4.21}$$

para $i = 1,2,...,n$, donde $\tilde{\mathbf{j}}_i$ es la i-ésima columna de $\tilde{\mathbf{J}}$.

Ejemplo 4.3 Para la función

$$\boldsymbol{f}(\mathbf{x}) = \begin{pmatrix} x_1^2 + x_2 - 11 \\ x_1 + x_2^2 - 7 \end{pmatrix}$$

evalúe la matriz jacobiana de la función en el punto $\mathbf{x}_0^T = (2,1)$ por las aproximaciones en diferencias finitas y compare éstas con la matriz jacobiana analítica. Use una perturbación del 1% en las variables independientes.

Solución:

La matriz jacobiana analítica de la función es

$$\mathbf{J}(\mathbf{x}) = \begin{pmatrix} 2x_1 & 1 \\ 1 & 2x_2 \end{pmatrix}$$

que, evaluada en $\mathbf{x}_0$, da

$$\mathbf{J}(\mathbf{x}_0) = \begin{pmatrix} 2(2) & 1 \\ 1 & 2(1) \end{pmatrix} = \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix}$$

Al 1% de cambio en las variables se tiene $\delta_1 = 0.02$ y $\delta_2 = 0.01$. Por las ecuaciones (4.19) a (4.21) las matrices jacobianas numéricas son

$$\tilde{\mathbf{J}}(\mathbf{x}_0) = \begin{pmatrix} 4.02 & 1.00 \\ 1.00 & 2.01 \end{pmatrix}, \quad \tilde{\mathbf{J}}(\mathbf{x}_0) = \begin{pmatrix} 3.98 & 1.00 \\ 1.00 & 1.99 \end{pmatrix} \text{ y } \tilde{\mathbf{J}}(\mathbf{x}_0) = \begin{pmatrix} 4.00 & 1.00 \\ 1.00 & 2.00 \end{pmatrix}$$

respectivamente, como puede verificar el lector.

Observe que para la función $\boldsymbol{f}(\mathbf{x})$ dada, los tres métodos dan una muy buena aproximación a la matriz jacobiana analítica. El método por diferencias centrales da igual a la matriz analítica, lo cual comprueba que esta aproximación es exacta para funciones cuadráticas.

4.5 Formulación del problema de regresión no lineal y de sistemas de ecuaciones no lineales

El *análisis de regresión* es la aplicación de métodos matemáticos y estadísticos para el análisis de datos experimentales y el ajuste de modelos matemáticos a estos datos a través de la estimación de los parámetros desconocidos del modelo. Para el análisis de regresión los modelos matemáticos son clasificados como *lineales* o *no lineales* con respecto a los parámetros desconocidos. Por ejemplo, para los modelos siguientes:

$$y = \beta_0 + \beta_1 x$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{n-1} x_{n-1} + \beta_n x_n$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{n-1} x^{n-1} + \beta_n x^n$$

la primera y segunda ecuaciones son lineales tanto en los coeficientes como en las variables. Éstas corresponden a un modelo lineal simple y a un modelo lineal múltiple respectivamente; la tercera ecuación es lineal en los coeficientes y no lineal en la variable independiente y corresponde a un modelo polinomial. Para los modelos siguientes:

$$\frac{1}{y} = \beta_0 \pm \beta_1 x$$

$$y = \beta_0 \pm \frac{\beta_1}{x}$$

$$\frac{x}{y} = \beta_0 \pm \beta_1 x$$

$$y = \beta_0 x^{\beta_1}$$

$$y = \beta_0 \beta_1^{\,x}$$

los coeficientes son lineales en las primeras tres ecuaciones y no lineales en las dos últimas, y todas las ecuaciones son no lineales entre las variables. La ventaja de estos modelos es que pueden transformarse a una relación lineal entre las variables mediante una transformación de variables y coeficientes adecuada para poder abordarse como un problema semejante a la primera ecuación del primer conjunto de modelos analizados. Sin embargo, siempre habrá modelos no lineales que no puedan transformarse como los anteriores, y entonces habrá que desarrollar métodos matemáticos para estimar los coeficientes asociados a estos modelos; tal es el objetivo de esta sección.

La regresión no lineal es una extensión de los métodos de regresión lineal usados de manera iterativa para llegar a los valores de los parámetros de los modelos no lineales. El análisis estadístico de los resultados de la regresión no lineal también es una extensión del aplicado en regresión lineal, pero no posee las bases teóricas rigurosas del último. Los métodos de los capítulos 2 y 3 se diseñaron para ser efectivos sobre todas las funciones objetivo suficientemente suaves y en especial sobre las funciones cuadráticas. Sin embargo, a veces es posible descubrir métodos que resultan ser más eficaces cuando la función objetivo tiene una forma especial; por ejemplo, una de las formas más importantes que se ha encontrado en la práctica es la suma de cuadrados de otras funciones no lineales, como

$$F(\mathbf{x}) = \sum_{i=1}^{m} f_i^2(\mathbf{x}) \tag{4.22}$$

donde $f_i(\mathbf{x})$ para $i = 1,2,...,m$, son funciones no lineales de varias variables. Ésta es parecida a la suma de cuadrados de los errores en el caso de la regresión lineal, según puede observarse.

La minimización de funciones de este tipo se llama *mínimos cuadrados no lineales*. La función objetivo con esta forma nunca puede tener valores negativos en los problemas donde surge ésta. Como notación es conveniente agrupar las funciones $f_i(\mathbf{x})$ en un vector, de la siguiente manera:

$$\boldsymbol{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix} \tag{4.23}$$

y entonces podemos escribir la ecuación (4.22) como

$$F(\mathbf{x}) = \boldsymbol{f}^T(\mathbf{x})\boldsymbol{f}(\mathbf{x}) \tag{4.24}$$

Ahora veamos dos problemas comunes en los que se dan tales funciones en la práctica.

Regresión no lineal

Un problema común es enfrentarse al ajuste de una función matemática a datos experimentales variando los parámetros de la función. Por ejemplo, sobre bases teóricas o empíricas se selecciona una forma funcional específica $\Phi(\boldsymbol{\beta},\mathbf{x})$, donde los elementos de $\mathbf{x}$ son las variables independientes o de control. Éstas son puestas experimentalmente en algún valor x_i y se mide alguna propiedad y_i del sistema que se esta estudiando. $\boldsymbol{\beta}$ es un vector de n parámetros que tienen que ajustarse hasta que se obtenga el mejor ajuste de $\Phi(\boldsymbol{\beta},\mathbf{x})$; por tanto, éstas son las variables de optimización de nuestro problema. Cuando $\Phi(\boldsymbol{\beta},\mathbf{x})$ es no lineal en $\boldsymbol{\beta}$ este tipo de problema se conoce como *regresión no lineal*. Cuando $m > n$ se dice que el problema esta *sobredeterminado* y cuando $m < n$, el problema se caracteriza como un *sistema subdeterminado*. Para la mayoría de los problemas sobredeterminados no será posible hallar $\boldsymbol{\beta}$ de manera que $\Phi(\boldsymbol{\beta},\mathbf{x})$ pase a través de todos los datos experimentales.

La diferencia entre y_i y el valor predicho de $\hat{y}_i = \Phi(\boldsymbol{\beta},x_i)$ se llama *residual* $\hat{\varepsilon}_i$ y se expresa como

$$\hat{\varepsilon}_i = y_i - \Phi(\boldsymbol{\beta},x_i) \text{ para } i = 1,...,m \tag{4.25}$$

El mejor ajuste por mínimos cuadrados se obtiene minimizando la función $S(\boldsymbol{\beta})$, que es la suma de cuadrados de los residuales

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{m} \hat{\varepsilon}_i^2 \tag{4.26}$$

con respecto a $\boldsymbol{\beta}$. Se toman los cuadrados para evitar la cancelación entre residuales de signo opuesto como en el caso de la regresión lineal. Observe que $S\left(\boldsymbol{\beta}^*\right)=0$ si se obtiene un ajuste perfecto de $\Phi\left(\boldsymbol{\beta},\mathbf{x}\right)$ a los datos experimentales.

Sistemas de ecuaciones no lineales

Otro problema común es hallar la solución de un sistema de n ecuaciones no lineales

$$\begin{aligned} f_1\left(\mathbf{x}\right)&=0\\ &\vdots\\ f_n\left(\mathbf{x}\right)&=0 \end{aligned} \tag{4.27}$$

en n incógnitas $\mathbf{x}$. En notación vectorial, la ecuación (4.27) puede exponerse como

$$\boldsymbol{f}\left(\mathbf{x}\right)=\mathbf{0} \tag{4.28}$$

El mínimo $\mathbf{x}^*$ de la función

$$F\left(\mathbf{x}\right)=\sum_{i=1}^{n} f_i^2\left(\mathbf{x}\right)=\mathbf{f}^T\left(\mathbf{x}\right)\mathbf{f}\left(\mathbf{x}\right) \tag{4.29}$$

para la cual $F\left(\mathbf{x}^*\right)=0$ es la solución deseada debido a que esta puede darse solo si $\mathbf{x}^*$ satisface cada una de las ecuaciones (4.27). Debe tenerse en mente que tal punto no necesariamente existe cuando se dice que el sistema es *inconsistente* y por lo tanto no tiene solución. Asimismo, la función $F\left(\mathbf{x}\right)$ puede tener cualquier número de mínimos locales con $F\left(\mathbf{x}^*\right)>0$, lo cual no tiene mucha relevancia en este problema excepto que no sean los mínimos deseados.

4.6 Derivadas de la suma de cuadrados de $F(\mathbf{x})$

Para obtener una expresión para el vector gradiente de una función de la forma de la ecuación (4.24), derivamos con respecto a $\mathbf{x}$

$$\nabla F(\mathbf{x}) = \sum_{j=1}^{n} \frac{\partial}{\partial x_j} \left(\sum_{i=1}^{n} f_i^2 \right) \hat{\mathbf{e}}_j = \sum_{j=1}^{n} \sum_{i=1}^{n} 2 f_i \frac{\partial f_i}{\partial x_j} \hat{\mathbf{e}}_j \tag{4.30}$$

De la definición 1.12 o la ecuación (1.4a) acerca de la matriz jacobiana, se tiene

$$\mathbf{J}(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial f_1(\mathbf{x})}{\partial x_1} & \dfrac{\partial f_1(\mathbf{x})}{\partial x_2} & \cdots & \dfrac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \dfrac{\partial f_2(\mathbf{x})}{\partial x_1} & \dfrac{\partial f_2(\mathbf{x})}{\partial x_2} & \cdots & \dfrac{\partial f_2(\mathbf{x})}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_n(\mathbf{x})}{\partial x_1} & \dfrac{\partial f_n(\mathbf{x})}{\partial x_2} & \cdots & \dfrac{\partial f_n(\mathbf{x})}{\partial x_n} \end{pmatrix} \tag{4.31}$$

y como la ecuación (4.30) expresa a un vector columna producto de una matriz y un vector columna, entonces el vector gradiente puede escribirse como

$$\nabla F(\mathbf{x}) = 2\mathbf{J}^T(\mathbf{x})\mathbf{f}(\mathbf{x}) \tag{4.32}$$

Derivando la ecuación (4.30) con respecto a $\mathbf{x}$ se obtiene la matriz hessiana de $F(\mathbf{x})$:

$$\nabla\left(\nabla^T F(\mathbf{x})\right) = \mathbf{H}(\mathbf{x}) = \sum_{k=1}^{n} \frac{\partial}{\partial x_k} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} 2 f_i \frac{\partial f_i}{\partial x_j} \right) \hat{\mathbf{e}}_k \hat{\mathbf{e}}_j^T$$

$$\mathbf{H}(\mathbf{x}) = 2\sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\frac{\partial f_i}{\partial x_k}\frac{\partial f_i}{\partial x_j} + f_i\frac{\partial^2 f_i}{\partial x_k \partial x_j}\right)\hat{\mathbf{e}}_k\hat{\mathbf{e}}_j^T$$

y el resultado es una matriz, como se esperaba; el primer término del lado derecho corresponde a un producto entre dos matrices jacobianas, mientras que el segundo término es un producto entre cada función f_i y la matriz hessiana de cada función f_i. Es decir, si se define $\mathbf{G}_i(\mathbf{x})$ como la matriz hessiana de $f_i(\mathbf{x})$, entonces la matriz hessiana completa de $F(\mathbf{x})$ puede escribirse como

$$\mathbf{H}(\mathbf{x}) = 2\mathbf{J}^T(\mathbf{x})\mathbf{J}(\mathbf{x}) + 2\sum_{i=1}^{n} f_i(\mathbf{x})\mathbf{G}_i(\mathbf{x}) \tag{4.33}$$

Además, definiendo

$$\mathbf{S}(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})\mathbf{G}_i(\mathbf{x}) \tag{4.34}$$

se tiene

$$\mathbf{H}(\mathbf{x}) = 2\mathbf{J}^T(\mathbf{x})\mathbf{J}(\mathbf{x}) + 2\mathbf{S}(\mathbf{x}) \tag{4.35}$$

Método de Newton

El punto de inicio para los algoritmos de mínimos cuadrados no lineales especializados es el método de Newton, la iteración básica de dicho método está dada por

$$\mathbf{H}_k \mathbf{d}_k = -\nabla F_k$$

es decir,

$$\left(\mathbf{J}_k^T \mathbf{J}_k + \mathbf{S}_k\right)\mathbf{d}_k = -\mathbf{J}_k^T \mathbf{f}_k \tag{4.36}$$

y

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k \tag{4.37}$$

El problema principal en la aplicación del método de Newton o métodos de Newton modificados a la suma de cuadrados de las funciones es el cálculo del término $\mathbf{S}_k$, que si bien permite la simetría, implica la determinación y evaluación de $mn(n+1)/2$ términos de segundas derivadas. Sin embargo, el resto de la matriz hessiana completa está expresada sólo en términos de primeras derivadas.

Esta observación da lugar a dos amplias clases de algoritmos especializados para mínimos cuadrados no lineales: aquellos que ignoran el término en $\mathbf{S}_k$, que suelen denominarse *algoritmos de residuales pequeños*, y aquellos que aproximan este término en alguna forma y que se denominan *algoritmos de residuales grandes*. Queda fuera del alcance de este texto el análisis de los algoritmos de residuales grandes, por lo que sólo se analizarán los primeros.

4.7 Estimación por mínimos cuadrados no lineales con algoritmos de residuales pequeños

En esta sección se analizarán métodos que intentan resolver el problema expresado por la ecuación (4.29) ignorando el término de evaluar la matriz $\mathbf{S}_k$ en la ecuación (4.35).

Método de Gauss-Newton

Ignorando el término en $\mathbf{S}_k$, la ecuación (4.36) del método de Newton resulta en

$$\mathbf{J}_k^T \mathbf{J}_k \mathbf{d}_k = -\mathbf{J}_k^T \mathbf{f}_k \tag{4.38}$$

que, junto con la ecuación

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k \tag{4.39}$$

definen el *método de Gauss-Newton*. El sistema de ecuaciones simultáneas (4.38) se conocen como *las ecuaciones normales de los mínimos cuadrados*. Es importante saber cuándo puede justificarse hacer esta aproximación, y la respuesta es hasta donde lo permita la generación de direcciones de descenso, pues la ecuación (4.38) es menos complicada que la ecuación (4.36) del método de Newton, debido a que la matriz $\mathbf{J}_k^T \mathbf{J}_k$ al menos siempre es positiva semidefinida. Para ver esto, tómese un vector arbitrario $\mathbf{z} \neq \mathbf{0}$ y sea $\mathbf{y} = \mathbf{J}_k \mathbf{z}$; entonces

$$\mathbf{z}^T \mathbf{J}_k^T \mathbf{J}_k \mathbf{z} = \mathbf{y}^T \mathbf{y} \geq 0$$

El único problema que puede darse con respecto a esta simplificación es cuando $\mathbf{J}_k$ sea de rango deficiente y $\mathbf{J}_k^T \mathbf{J}_k$ resulte singular. Sin embargo, aún si $\mathbf{d}_k$ es una dirección de descenso, esto no garantiza que el valor de la función objetivo decrezca, es decir, que $F_{k+1} < F_k$. El paso expresado por la ecuación (4.39) podría se muy largo, localizando a $\mathbf{x}_{k+1}$ en un punto muy alejado del punto mínimo lineal. Por estas razones, se requiere de un buen punto inicial para que exista una oportunidad de convergencia hacia el mínimo.

Refiriéndose a las ecuaciones (4.33), (4.34) y (4.36), está claro que si $\boldsymbol{f}(\mathbf{x}) \to \mathbf{0}$ como $\mathbf{x} \to \mathbf{x}^*$ también $\mathbf{S}(\mathbf{x}) \to \mathbf{0}$, y el método de Gauss-Newton tiende al método de Newton como se vaya acercando al mínimo, lo cual tiene consecuencias favorables para la rapidez de convergencia. Esta situación ocurre en regresión no lineal cuando es posible un buen ajuste o en la solución de un conjunto consistente y bien condicionado de ecuaciones simultáneas no lineales. Las consecuencias también son las mismas cuando las funciones f_i son casi lineales de manera que $\mathbf{G}_i \simeq \mathbf{0}$ para $i = 1,...,m$, o cuando los términos sucesivos $f_i \mathbf{G}_i$ se cancelen a través de diferencias de signo. Finalmente, las propiedades de convergencia del método de Gauss-Newton pueden mejorarse mucho si éste se usa junto con una búsqueda lineal en la ecuación (4.39).

Algoritmo 4.1 Método de Gauss-Newton con búsqueda lineal:

Dados el conjunto de funciones $\boldsymbol{f}(\mathbf{x})$ y la matriz jacobiana $\mathbf{J}(\mathbf{x})$, un punto inicial $\mathbf{x}_0$, $\alpha = 1$ y ε.

Para $k = 0,1,2,...$ hasta donde se satisfaga, hacer:

1. $\mathbf{d}_k = -\left(\mathbf{J}_k^T \mathbf{J}_k\right)^{-1} \mathbf{J}_k^T \mathbf{f}_k$ o bien resolver $\mathbf{J}_k^T \mathbf{J}_k \mathbf{d}_k = -\mathbf{J}_k^T \mathbf{f}_k$ y obtener $\mathbf{d}_k$.

   Para $j = 1,2,3,...$ hasta donde se satisfaga, hacer;

2. Calcular $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha \mathbf{d}_k$.

   Si $\left\| \boldsymbol{f}(\mathbf{x}_{k+1}) \right\| < (1 - \alpha/2) \left\| \boldsymbol{f}(\mathbf{x}_k) \right\|$, entonces

   Ir al paso 5.
   Fin del si.

3. $\alpha = \dfrac{\alpha}{2}$.
4. Tomar $j = j + 1$ e ir al paso 2.

5. Si $\left\| \mathbf{x}_{k+1} - \mathbf{x}_k \right\| \leq \varepsilon$ o $\left\| 2\mathbf{J}_{k+1}^T \mathbf{f}_{k+1} \right\| \leq \varepsilon$, entonces

   Tomar $\mathbf{x}_{k+1} \approx \mathbf{x}^*$ y parar.

Fin del si.

6. Tomar $k = k + 1$ y regresar al paso 1.

Ejemplo 4.4 Aplicando el método de Gauss-Newton estime los coeficientes en la correlación de la función:

$$y = e^{\beta_1 + \beta_2 x}$$

usando los siguientes datos experimentales y minimizando la suma de los cuadrados de las desviaciones entre los datos experimentales y los valores predichos de y.

n	y_{exp}	x
1	0.50	-2.00
2	1.00	-1.00
3	2.00	0.00
4	4.00	1.00

Solución:

Iniciando con $\boldsymbol{\beta}_0 = (1,1)^T$ y $\varepsilon = 0.001$, el mínimo hallado con la precisión dada es $\boldsymbol{\beta}^{*T} = (0.6931, 0.6931)$, donde $F\left(\boldsymbol{\beta}^*, x\right) = 0.0$, el número total de iteraciones principales fue de $k = 4$.

n	x	y_{exp}	y_{cal}	$error$
1	-2.00	0.5000	0.5000	0.0000
2	-1.00	1.0000	1.0000	0.0000
3	0.00	2.0000	2.0000	-0.0000
4	1.00	4.0000	4.0000	-0.0000

Tabla 4.3 Tabla de datos observados y calculados por la correlación.

Método de Levenberg-Marquardt

El método de *Levenberg-Marquardt* incorpora una técnica *ad-hoc* para tratar los problemas relacionados con la singularidad en la matriz $\mathbf{J}_k^T\mathbf{J}_k$ y es un algoritmo efectivo para problemas de residuales pequeños. La ecuación (4.38) se modifica a

$$\left(\mathbf{J}_k^T\mathbf{J}_k + \mu_k\mathbf{I}\right)\mathbf{d}_k = -\mathbf{J}_k^T\mathbf{f}_k \tag{4.40}$$

donde $\mu_k \geq 0$ es un escalar e $\mathbf{I}$ es la matriz unidad de orden n. Entonces se usa la ecuación (4.39) para obtener un punto con el que se inicia la próxima iteración. Para un valor suficientemente grande de μ_k, la matriz $\mathbf{J}_k^T\mathbf{J}_k + \mu_k\mathbf{I}$ es positiva definida y $\mathbf{d}_k$ es una dirección de descenso. Sin embargo, se requiere que $\mu_k \to 0$ como $\mathbf{x}_k \to \mathbf{x}^*$, así que el método adquiere la rapidez de convergencia asintótica del método de Gauss-Newton. En el método originalmente inventado por Levenberg (1944), μ_k se elige para minimizar $F(\mathbf{x}_k + \mathbf{d}_k)$ con $\mathbf{d}_k$ dada por la ecuación (4.40), mientras todo lo demás se mantiene constante. Como un paso infinitesimal en la dirección de descenso, por definición siempre reducirá el valor de la función en puntos no estacionarios, un valor suficientemente grande de μ_k siempre será exitoso y, como consecuencia, el método puede hacerse globalmente convergente. La técnica para hallar el valor de μ_k esta muy relacionado con la búsqueda lineal, pero no es estrictamente el caso.

Estrategia de Marquardt

Marquardt (1963) mejoró la eficiencia del algoritmo inventando una mejor estrategia para seleccionar μ_k. Este se pone inicialmente en algún valor positivo (por ejemplo, 0.01) y se utiliza un factor $\nu > 1$ (por ejemplo, 10) con el cual incrementar o disminuir μ_k. Al principio de cada iteración μ_k se reduce por el factor ν, en un intento por acercar el algoritmo al método de Gauss-Newton. Si éste no reduce el valor de la función objetivo, se incrementa repetidamente el nuevo valor de μ_k por el factor ν hasta que se obtenga una reducción en el valor de la función.

Algoritmo 4.2 Método de Levenberg-Marquardt (por Marquardt):

Dados el conjunto de funciones $\boldsymbol{f}(\mathbf{x})$ y la matriz jacobiana $\mathbf{J}(\mathbf{x})$, un punto inicial $\mathbf{x}_0$ y ε.

Sea $\mu_0 = 0.01$ y $\nu = 10$.

Para $k = 0,1,2,...$ hasta donde se satisfaga, hacer:

1. $\mu_k = \mu_k / \nu$.
2. Para $j = 0,1,2,...$ hasta donde se satisfaga, hacer;
3. $\mathbf{d}_k = -\left(\mathbf{J}_k^T\mathbf{J}_k + \mu_k\mathbf{I}\right)^{-1}\mathbf{J}_k^T\mathbf{f}_k$ o bien resolver $\left(\mathbf{J}_k^T\mathbf{J}_k + \mu_k\mathbf{I}\right)\mathbf{d}_k = -\mathbf{J}_k^T\mathbf{f}_k$

   y obtener $\mathbf{d}_k$.
4. Calcular $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$.
5. Si $F_{k+1} > F_k$, entonces $\mu_k = \mu_k\nu$.
   Fin del si.
6. Si $F_{k+1} < F_k$, entonces
   Ir al paso 8.
   Fin del si.
7. Tomar $j = j + 1$ regresar a 3.

8. $\mu_{k+1} = \mu_k$.

9. Si $\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\| \leq \varepsilon$ o $\left\|2\mathbf{J}_{k+1}^T \mathbf{f}_{k+1}\right\| \leq \varepsilon$, entonces

   Tomar $\mathbf{x}_{k+1} \approx \mathbf{x}^*$ y parar.
   Fin del si.

10. Tomar $k = k+1$ y regresar al paso 1.

En este algoritmo no se desperdicia tiempo refinando el valor de μ_k , ya que es mejor proceder en la próxima iteración desde el punto que se inicia más cerca de la solución. Marquardt sugiere que podría ser ventajoso reemplazar la matriz unidad $\mathbf{I}$ por una matriz diagonal $\mathbf{D}_k$ cuyos elementos diagonales no negativos se eligen de manera que reflejen el escalamiento de las variables.

Ejemplo 4.5 Aplicando el método de Levenberg-Marquardt estime los coeficientes en la correlación de la función:

$$y = \beta_1 + \beta_2 sen(\beta_3 x)$$

usando los siguientes datos experimentales y minimizando la suma de los cuadrados de las desviaciones entre los datos experimentales y los valores predichos de y.

n	y_{exp}	x
1	10.5244	1
2	10.7278	2
3	8.4233	3
4	5.7295	4
5	5.1232	5
6	7.1617	6
7	9.9709	7
8	10.9680	8
9	9.2363	9
10	6.3679	10
11	5.0000	11
12	6.3902	12
13	9.2605	13
14	10.9718	14
15	9.9508	15
16	7.1362	16
17	5.1158	17
18	5.7470	18
19	8.4496	19
20	10.7388	20

Solución:

Con $\boldsymbol{\beta}_0 = (1.,-1.1)^T$ y $\varepsilon = 0.001$, el mínimo es $\boldsymbol{\beta}^{*T} = (8.00,3.00,1.00)$ donde $F(\boldsymbol{\beta}^*, x) = 0.0$, el número total de iteraciones principales fue de $k = 2$, la tabla 4.4 muestra la comparación de datos observados contra los calculados.

n	x	y_{exp}	y_{cal}	$error$
1	1	10.5244	10.5244	0.0000
2	2	10.7278	10.7278	0.0000
3	3	8.4233	8.4233	0.0000
4	4	5.7295	5.7295	-0.0000
5	5	5.1232	5.1232	-0.0000
6	6	7.1617	7.1617	-0.0000
7	7	9.9709	9.9709	-0.0000
8	8	10.9680	10.9680	0.0000
9	9	9.2363	9.2363	0.0000
10	10	6.3679	6.3679	0.0000
11	11	5.0000	5.0000	-0.0000
12	12	6.3902	6.3902	-0.0000
13	13	9.2605	9.2605	-0.0000
14	14	10.9718	10.9718	0.0000
15	15	9.9508	9.9508	0.0000
16	16	7.1362	7.1362	0.0000
17	17	5.1158	5.1158	0.0000
18	18	5.7470	5.7470	-0.0000
19	19	8.4496	8.4496	-0.0000
20	20	10.7388	10.7388	-0.0000

Tabla 4.4 Tabla de datos observados y calculados por la correlación.

Métodos Cuasi-Newton

En el campo de los mínimos cuadrados no lineales, los métodos cuasi-Newton se aplican cuando la matriz jacobiana no esta disponible analíticamente y la aproximación por diferencias finitas resulta muy laboriosa. En la k-ésima iteración se obtiene una aproximación $\mathbf{B}_k$ a $\mathbf{J}_k$ desde el punto actual $\mathbf{x}_k$; luego se determina el vector de búsqueda $\mathbf{d}_k$ por alguno de los métodos ya estudiados usando $\mathbf{B}_k$ en lugar de $\mathbf{J}_k$, y se realiza una búsqueda lineal para determinar

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \tag{4.41}$$

de manera que $\left\|\mathbf{f}_{k+1}\right\|$ sea suficientemente menor que $\left\|\mathbf{f}_{k}\right\|$. La expansión en serie de Taylor hasta términos de primer orden en $\left(\mathbf{x}_{k+1}-\mathbf{x}_{k}\right)$ muestra que

$$\mathbf{f}_{k+1}=\mathbf{f}_{k}+\mathbf{J}_{k}\left(\mathbf{x}_{k+1}-\mathbf{x}_{k}\right) \tag{4.42}$$

o bien

$$\mathbf{J}_{k}\left(\mathbf{x}_{k+1}-\mathbf{x}_{k}\right)=\mathbf{f}_{k+1}-\mathbf{f}_{k} \tag{4.43}$$

por tanto, $\mathbf{B}_{k}$ se actualiza en la forma usual de acuerdo con

$$\mathbf{B}_{k+1}=\mathbf{B}_{k}+\mathbf{A}_{k} \tag{4.44}$$

donde $\mathbf{A}_{k}$ es una matriz de actualización de manera que

$$\mathbf{B}_{k}\left(\mathbf{x}_{k+1}-\mathbf{x}_{k}\right)=\mathbf{f}_{k+1}-\mathbf{f}_{k} \tag{4.45}$$

esta es la *condición cuasi-Newton* para mínimos cuadrados no lineales. Para explicarse de una manera sencilla, supóngase que $\boldsymbol{f}\left(\mathbf{x}\right)=\mathbf{0}$ es un sistema de ecuaciones lineales, por ejemplo $\mathbf{A}\mathbf{x}-\mathbf{b}=\mathbf{0}$. Si se restan los valores de $\boldsymbol{f}\left(\mathbf{x}\right)$ en dos puntos sucesivos del proceso iterativo, k y $\mathrm{k}+1$, se tiene

$$\boldsymbol{f}\left(\mathbf{x}_{k+1}\right)-\boldsymbol{f}\left(\mathbf{x}_{k}\right)=\mathbf{A}\mathbf{x}_{k+1}-\mathbf{b}-\mathbf{A}\mathbf{x}_{k}+\mathbf{b}=\mathbf{A}\left(\mathbf{x}_{k+1}-\mathbf{x}_{k}\right)$$

en el caso no lineal esta igualdad no se cumple aunque puede hacerse eligiendo $\mathbf{A}_{k+1}$ adecuadamente:

$$\mathbf{A}_{k+1}\left(\mathbf{x}_{k+1}-\mathbf{x}_{k}\right)\approx\boldsymbol{f}\left(\mathbf{x}_{k+1}\right)-\boldsymbol{f}\left(\mathbf{x}_{k}\right) \tag{4.46}$$

los métodos cuasi-Newton construyen una sucesión $\{\mathbf{A}_k\}$ de tal forma que $\mathbf{A}_k$ aproxime lo mejor posible a la matriz jacobiana $\mathbf{J}(\mathbf{x}_k)$.

Método de Broyden de rango uno

Broyden (1965) utilizó una idea muy simple para obtener una aproximación satisfactoria de $\mathbf{A}_{k+1}$; escogerla de tal forma que se minimice el valor de la función que se obtendría en un mismo punto mediante las aproximaciones $\mathbf{A}_{k+1}$ y $\mathbf{A}_k$ y que se cumpla a la vez la condición cuasi-Newton, ecuación (4.46), dada como

$$\mathbf{A}_{k+1}(\mathbf{x}_{k+1}-\mathbf{x}_k)=\boldsymbol{f}(\mathbf{x}_{k+1})-\boldsymbol{f}(\mathbf{x}_k)$$

Con la aproximación mencionada partiendo de $\mathbf{x}_{k+1}$ y $\mathbf{x}_k$, la diferencia de los valores de la función en un punto $\mathbf{x}$ que habría que minimizar sería

$$\boldsymbol{f}(\mathbf{x}_{k+1})+\mathbf{A}_{k+1}(\mathbf{x}-\mathbf{x}_{k+1})-\boldsymbol{f}(\mathbf{x}_k)-\mathbf{A}_k(\mathbf{x}-\mathbf{x}_k)$$

Desarrollándola, queda

$$\boldsymbol{f}(\mathbf{x}_{k+1})-\boldsymbol{f}(\mathbf{x}_k)-\mathbf{A}_{k+1}(\mathbf{x}_{k+1}-\mathbf{x}_k)+(\mathbf{A}_{k+1}-\mathbf{A}_k)(\mathbf{x}-\mathbf{x}_k)$$

y sustituyendo la ecuación (4.46) en ésta última expresión, la diferencia por minimizar resulta

$$(\mathbf{A}_{k+1}-\mathbf{A}_k)(\mathbf{x}-\mathbf{x}_k) \tag{4.47}$$

Si para todo $\mathbf{x}$ la diferencia $\mathbf{x}-\mathbf{x}_k$ se expresa como

$$\mathbf{x}-\mathbf{x}_k=\alpha\Delta\mathbf{x}_k+\mathbf{s}$$

donde $\Delta\mathbf{x}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ y se cumple que $\mathbf{s}^T \Delta\mathbf{x} = 0$; la expresión por minimizar queda

$$\alpha\left(\mathbf{A}_{k+1} - \mathbf{A}_k\right)\Delta\mathbf{x}_k + \left(\mathbf{A}_{k+1} - \mathbf{A}_k\right)\mathbf{s}$$

Sobre el primer término no se puede actuar puesto que, según la ecuación (4.46),

$$\left(\mathbf{A}_{k+1} - \mathbf{A}_k\right)\Delta\mathbf{x}_k = \Delta\mathbf{f}_k - \mathbf{A}_k\Delta\mathbf{x}_k \tag{4.48}$$

donde $\Delta\mathbf{f}_k = \boldsymbol{f}\left(\mathbf{x}_{k+1}\right) - \boldsymbol{f}\left(\mathbf{x}_k\right)$. El segundo término se puede hacer cero para todo $\mathbf{x}$ escogiendo $\mathbf{A}_{k+1}$ de tal manera que

$$\left(\mathbf{A}_{k+1} - \mathbf{A}_k\right)\mathbf{s} = \mathbf{0} \tag{4.49}$$

para todo $\mathbf{s}$ ortogonal a $\Delta\mathbf{x}_k$. Esto requiere que la matriz $\mathbf{A}_{k+1} - \mathbf{A}_k$ sea de rango uno, es decir, de la forma, $\mathbf{u}\Delta\mathbf{x}_k^T$, con $\mathbf{u} \in R^n$. Ahora bien, para que se cumpla la condición cuasi-Newton, ecuación (4.46), lo que equivale a la ecuación (4.48) como se acaba de ver, es decir, a $\Delta\mathbf{f}_k - \mathbf{A}_k\Delta\mathbf{x}_k$, el vector $\mathbf{u}$ debe expresarse como

$$\mathbf{u} \equiv \frac{\Delta\mathbf{f}_k - \mathbf{A}_k\Delta\mathbf{x}_k}{\Delta\mathbf{x}_k^T\Delta\mathbf{x}_k} \tag{4.50}$$

La matriz, haciendo $\mathbf{B} = \mathbf{A}$

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\left(\Delta\mathbf{f}_k - \mathbf{B}_k\Delta\mathbf{x}_k\right)\Delta\mathbf{x}_k^T}{\Delta\mathbf{x}_k^T\Delta\mathbf{x}_k} \tag{4.51}$$

es, por lo tanto, la que cumple ese propósito de minimizar la diferencia indicada, verificándose además la condición cuasi-Newton, ecuación (4.46). La fórmula dada por la ecuación (4.51) es la que propuso Broyden para aproximar la matriz jacobiana en cada iteración del método de Newton. Para estos métodos,

$\mathbf{B}_{k+1}$ mantiene la información derivativa exacta a lo largo del vector $\mathbf{d}_k$ cuando $\mathbf{f}$ es lineal. Broyden (1965) establece la fórmula de actualización de manera única dada por

$$\mathbf{A}_k = \frac{\left(\mathbf{f}_{k+1} - \mathbf{f}_k - \mathbf{B}_k\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right)\right)\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right)^T}{\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right)^T\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right)} \tag{4.52}$$

que es la base del método de rango uno de Broyden. Esta ecuación se incorpora dentro del método de Gauss-Newton para establecer el siguiente algoritmo.

Algoritmo 4.3 Método de Broyden de rango uno con búsqueda lineal:

Dados el conjunto de funciones $\boldsymbol{f}\left(\mathbf{x}\right)$, un punto inicial $\mathbf{x}_0$ y ε .

Calcular $\mathbf{B}_0 = \tilde{\mathbf{J}}\left(\mathbf{x}_0\right)$ por diferencias finitas.

Para $k = 0,1,2,...$ hasta donde se satisfaga, hacer:

1. $\mathbf{d}_k = -\left(\mathbf{B}_k^T\mathbf{B}_k\right)^{-1}\mathbf{B}_k^T\mathbf{f}_k$ o bien resolver $\mathbf{B}_k^T\mathbf{B}_k\mathbf{d}_k = -\mathbf{B}_k^T\mathbf{f}_k$ y obtener $\mathbf{d}_k$.
   Para $j = 1,2,3,...$ hasta donde se satisfaga, hacer;
2. Calcular $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha\mathbf{d}_k$.
   Si $\left\|\boldsymbol{f}\left(\mathbf{x}_{k+1}\right)\right\| < \left(1-\alpha/2\right)\left\|\boldsymbol{f}\left(\mathbf{x}_k\right)\right\|$, entonces
   Ir al paso 5.
   Fin del si.
3. $\alpha = \frac{\alpha}{2}$.
4. Tomar $j = j+1$ e ir al paso 2.

5. Si $\left\|\mathbf{x}_{k+1}-\mathbf{x}_k\right\|\leq\varepsilon$ o $\left|\dfrac{F_{k+1}-F_k}{F_k}\right|\leq\varepsilon$, entonces
   Tomar $\mathbf{x}_{k+1}\approx\mathbf{x}^*$ y parar.
   Fin del si.
6. Calcular $\mathbf{A}_k=\dfrac{\left(\mathbf{f}_{k+1}-\mathbf{f}_k-\mathbf{B}_k\left(\mathbf{x}_{k+1}-\mathbf{x}_k\right)\right)\left(\mathbf{x}_{k+1}-\mathbf{x}_k\right)^T}{\left(\mathbf{x}_{k+1}-\mathbf{x}_k\right)^T\left(\mathbf{x}_{k+1}-\mathbf{x}_k\right)}$.
7. Calcular $\mathbf{B}_{k+1}=\mathbf{B}_k+\mathbf{A}_k$.
8. Tomar $k=k+1$ y regresar al paso 1.

Ejemplo 4.6 Usando el método de Broyden estime los coeficientes en la correlación

$$y=\beta_1x_1^2+\beta_2x_2^2+\beta_3x_1x_2$$

con los siguientes datos experimentales, minimizando la suma de los cuadrados de las desviaciones entre los datos experimentales y los valores predichos de y .

n	y_{exp}	x_1	x_2
1	162	274	2450
2	120	180	3254
3	223	375	3802
4	131	205	2838
5	67	86	2347
6	169	265	3782
7	81	98	3008
8	192	330	2450
9	116	195	2137
10	55	53	2560
11	252	430	4020
12	232	372	4427
13	144	236	2660
14	103	157	2088
15	212	370	2605

Solución:

Con $\boldsymbol{\beta}_0 = (1.0, 1.0, 1.0)^T$ y $\varepsilon = 10^{-6}$, el mínimo hallado con la precisión dada, es $\boldsymbol{\beta}^{*T} = (0.002055, 0.000014, -0.000175)$ donde $F(\boldsymbol{\beta}^*, \mathbf{x}) = 21652.75$, La convergencia del método se obtuvo en $k = 21$ iteraciones principales.

n	x_1	x_2	y_{exp}	y_{cal}	$error$
1	274	2450	162.0	118.8	43.2
2	180	3254	120.0	108.6	11.4
3	375	3802	223.0	236.9	-13.9
4	205	2838	131.0	94.5	36.5
5	86	2347	67.0	55.0	12.0
6	265	3782	169.0	164.2	4.8
7	98	3008	81.0	91.6	10.6
8	330	2450	192.0	164.3	27.7
9	195	2137	116.0	67.6	48.4
10	53	2560	55.0	71.4	-16.4
11	430	4020	252.0	298.2	-46.2
12	372	4427	232.0	263.8	-31.8
13	236	2660	144.0	101.2	42.8
14	157	2088	103.0	52.8	50.2
15	370	2605	212.0	205.4	6.6

Tabla 4.5 Tabla de datos observados y calculados por la correlación.

4.8 Solución de sistemas de ecuaciones no lineales

El método de Newton-Raphson para sistemas de ecuaciones no lineales se define de una manera semejante a como se hizo en el caso de una sola variable. Se estudiará el caso de resolver el problema dado por la ecuación (4.28)

$$\boldsymbol{f}\left(\mathbf{x}\right)=\mathbf{0}$$

Método de Newton

El enfoque que se considerará es el similar a la formulación de un problema de mínimos cuadrados no lineales de residuales pequeños con una matriz jacobiana de tamaño n por n. Cuando la matriz jacobiana es cuadrada y no singular las ecuaciones normales (4.38)

$$\mathbf{J}_k^T\mathbf{J}_k\mathbf{d}_k=-\mathbf{J}_k^T\mathbf{f}_k$$

pueden reducirse premultiplicando por $\mathbf{J}_k^{-T}$ para obtener

$$\mathbf{J}_k\mathbf{d}_k=-\mathbf{f}_k \tag{4.53}$$

Esta formulación junto con la ecuación (4.37) dada por

$$\mathbf{x}_{k+1}=\mathbf{x}_k+\mathbf{d}_k$$

definen al método de Newton para sistemas de ecuaciones no lineales que tiene propiedades de convergencia semejantes al método de Gauss-Newton. De nuevo, la estabilidad del método puede ampliarse realizando búsquedas lineales

$$\mathbf{x}_{k+1}=\mathbf{x}_k+\alpha_k\mathbf{d}_k$$

hasta reducir $\left\|\mathbf{f}_k\right\|$.

Algoritmo 4.4 Método de Newton:

Dados el conjunto de funciones $\boldsymbol{f}(\mathbf{x})$ y la matriz jacobiana $\mathbf{J}(\mathbf{x})$, un punto inicial $\mathbf{x}_0$ y ε .

Para $k = 0,1,2,...$ hasta donde se satisfaga, hacer.

1. $\mathbf{d}_k = -\mathbf{J}_k^{-1}\mathbf{f}_k$ o bien resolver $\mathbf{J}_k\mathbf{d}_k = -\mathbf{f}_k$ y obtener $\mathbf{d}_k$.
2. Calcular $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$.
3. Si $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \varepsilon$ o $\|\mathbf{f}_{k+1}\| \leq \varepsilon$, entonces

   Tomar $\mathbf{x}_{k+1} \approx \mathbf{x}^*$ y parar.
   Fin del si.
4. Tomar $k = k+1$ y regresar al paso 1.

Ejemplo 4.7 Resuelva utilizando el método de Newton y partiendo del punto $\mathbf{x}_0 = (1.0,1.0,1.0)^T$ y $\varepsilon = 0.01$, el sistema siguiente de ecuaciones no lineales.

$$f_1(\mathbf{x}) = 3x_1 - \cos(x_2x_3) - \frac{1}{2} = 0$$

$$f_2(\mathbf{x}) = x_1^2 - 81\left(x_2 + \frac{1}{10}\right)^2 + \text{sen}(x_3) + 1.06 = 0$$

$$f_3(\mathbf{x}) = e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0$$

Solución:

El mínimo hallado con la precisión dada es $\boldsymbol{\beta}^{*T} = (0.50,0.00,-0.52)$, donde $F(\mathbf{x}) = 0.0$, la convergencia del método se obtuvo $k = 6$ iteraciones principales.

Algoritmo 4.5 Método de Newton con búsqueda lineal:

Dados el conjunto de funciones $\boldsymbol{f}(\mathbf{x})$ y la matriz jacobiana $\mathbf{J}(\mathbf{x})$, un punto inicial $\mathbf{x}_0$ y ε .

Para $k = 0,1,2,...$ hasta donde se satisfaga, hacer:

1. $\mathbf{d}_k = -\mathbf{J}_k^{-1}\mathbf{f}_k$ o bien resolver $\mathbf{J}_k\mathbf{d}_k = -\mathbf{f}_k$ y obtener $\mathbf{d}_k$.
   Para $j = 1,2,3,...$ hasta donde se satisfaga, hacer;
2. Calcular $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha\mathbf{d}_k$.
   Si $\left\|\boldsymbol{f}(\mathbf{x}_{k+1})\right\| < (1-\alpha/2)\left\|\boldsymbol{f}(\mathbf{x}_k)\right\|$, entonces
   Ir al paso 5.
   Fin del si.
3. $\alpha = \dfrac{\alpha}{2}$.
4. Tomar $j = j+1$ e ir al paso 2.
5. Si $\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\| \leq \varepsilon$ o $\left\|\mathbf{f}_{k+1}\right\| \leq \varepsilon$, entonces
   Tomar $\mathbf{x}_{k+1} \approx \mathbf{x}^*$ y parar.
   Fin del si.
6. Tomar $k = k+1$ y regresar al paso 1.

Ejemplo 4.8

Resuelva el sistema de ecuaciones no lineales del ejemplo 4.7 utilizando el método de Newton con búsqueda lineal y partiendo del punto $\mathbf{x}_0 = (1.0,1.0,1.0)^T$ y $\varepsilon = 0.01$.

Solución:

El mínimo es $\boldsymbol{\beta}^{*T} = (0.50, 0.00, -0.52)$ con $F(\mathbf{x}) = 0.0$, la convergencia del método se obtuvo en $k = 6$ iteraciones principales.

Métodos Cuasi-Newton

El objetivo de estos métodos consiste en aproximar la matriz jacobiana en cada iteración del método de Newton mediante relaciones de recurrencia que la relacionen con el valor que toma en anteriores iteraciones. Los métodos cuasi-Newton son muy adecuados para la solución de sistemas de ecuaciones no lineales.

Método de Broyden de rango uno

Para llevar numéricamente a la práctica el método de Broyden, son varios los aspectos importantes que hay que considerar. El primero consiste en determinar una buena aproximación inicial de $\mathbf{B}_0$; lo que suele hacerse en este sentido es utilizar $\tilde{\mathbf{J}}_0$ la aproximación por diferencias finitas de $\mathbf{J}_0$. Otra cuestión importante que debe tenerse en cuenta nace del hecho de que el método de Broyden adapta de iteración en iteración la matriz $\mathbf{B}$ y no la $\mathbf{B}^{-1}$, con la que se operaría mucho más eficazmente. En este sentido y sin tener en cuenta otras consideraciones, ¿por qué no partir de una $\mathbf{B}_0^{-1}$ y readaptar $\mathbf{B}^{-1}$? Se reduciría el número global de operaciones necesarias para resolver el sistema de ecuaciones.

A tal efecto, la actualización dada por la ecuación (4.51) de rango uno de Broyden puede desarrollarse usando la fórmula de Householder del siguiente lema:

Lema 4.1 *(Sherman-Morrison-Woodbury)*.

(i) Si $\mathbf{A}$ es una matriz no singular de orden $n\,x\,n$ y $\mathbf{r}$ y $\mathbf{s}$ dos vectores cualesquiera de R^n, $\mathbf{A} + \mathbf{rs}^T$ es no singular, si y sólo si $u = 1 + \mathbf{s}^T \mathbf{A}^{-1} \mathbf{r} \neq 0$.

(ii) También

$$\left(\mathbf{A} + \mathbf{rs}^T\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{rs}^T\mathbf{A}^{-1}}{u} \qquad (4.54)$$

Como

$$\mathbf{A}+\mathbf{rs}^T=\left(\mathbf{I}+\mathbf{rs}^T\boldsymbol{A}^{-1}\right)\mathbf{A}$$

y $\mathbf{rs}^T$ es de rango uno, el punto (i) resulta del hecho de que la matriz

$$\mathbf{I}+\mathbf{rs}^T\mathbf{A}^{-1}$$

tiene $n-1$ valores propios iguales a la unidad y el restante es u . La fórmula del punto (ii) se comprueba fácilmente multiplicándola por $\mathbf{A}+\mathbf{rs}^T$. Entonces,

$$\left(\mathbf{A}+\mathbf{rs}^T\right)\left(\mathbf{A}+\mathbf{rs}^T\right)^{-1}=\mathbf{AA}^{-1}-\frac{\mathbf{AA}^{-1}\mathbf{rs}^T\mathbf{A}^{-1}}{u}$$

$$+\quad\mathbf{rs}^T\mathbf{A}^{-1}-\frac{\mathbf{rs}^T\mathbf{A}^{-1}\mathbf{rs}^T\mathbf{A}^{-1}}{u}$$

$$\left(\mathbf{A}+\mathbf{rs}^T\right)\left(\mathbf{A}+\mathbf{rs}^T\right)^{-1}=\mathbf{I}-\frac{\mathbf{rs}^T\mathbf{A}^{-1}}{u}+\mathbf{rs}^T\mathbf{A}^{-1}-\frac{(u-1)\mathbf{rs}^T\mathbf{A}^{-1}}{u}$$

y, por consiguiente,

$$\left(\mathbf{A}+\mathbf{rs}^T\right)\left(\mathbf{A}+\mathbf{rs}^T\right)^{-1}=\mathbf{I}$$

La aplicación inmediata de este lema lleva a deducir que si se conoce $\mathbf{B}_0^{-1}\left(=\tilde{\mathbf{J}}_0^{-1}\right)$ la fórmula de actualización de Broyden es

$$\mathbf{B}_{k+1}^{-1}=\mathbf{B}_k^{-1}-\frac{\left(\mathbf{B}_k^{-1}\Delta\mathbf{f}_k-\Delta\mathbf{x}_k\right)\Delta\mathbf{x}_k^T\mathbf{B}_k^{-1}}{\Delta\mathbf{x}_k^T\mathbf{B}_k^{-1}\Delta\mathbf{f}_k} \tag{4.55}$$

Entonces, las matrices $\mathbf{B}^{-1}$ reemplazan a las matrices $\mathbf{B}$ en la formulación básica y $\mathbf{d}_k$ se obtiene por analogía con la ecuación (4.53) como

$$\mathbf{d}_k=-\mathbf{B}_k^{-1}\mathbf{f}_k \tag{4.56}$$

sin necesidad de resolver un sistema de ecuaciones lineales.

Algoritmo 4.6 Método de Broyden de rango uno con búsqueda lineal:

Dados el conjunto de funciones $\boldsymbol{f}\left(\mathbf{x}\right)$, un punto inicial $\mathbf{x}_0$ y ε.

Calcular $\mathbf{B}_0^{-1}=\tilde{\mathbf{J}}_0^{-1}$ por diferencias finitas.

Para $k=0,1,2,...$ hasta donde se satisfaga, hacer:

1. $\mathbf{d}_k=-\mathbf{B}_k^{-1}\mathbf{f}_k$.
   Para $j=1,2,3,...$ hasta donde se satisfaga, hacer;
2. Calcular $\mathbf{x}_{k+1}=\mathbf{x}_k+\alpha\mathbf{d}_k$.
   Si $\left\|\boldsymbol{f}\left(\mathbf{x}_{k+1}\right)\right\|<\left(1-\alpha/2\right)\left\|\boldsymbol{f}\left(\mathbf{x}_k\right)\right\|$, entonces
   Ir al paso 5.

Fin del si.

3. $\alpha = \frac{\alpha}{2}$.
4. Tomar $j = j + 1$ e ir al paso 2.
5. Si $\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\| \leq \varepsilon$ o $\left\|\mathbf{f}_{k+1}\right\| \leq \varepsilon$, entonces

   Tomar $\mathbf{x}_{k+1} \approx \mathbf{x}^*$ y parar.
   Fin del si.

6. Calcular $\mathbf{A}_k = \frac{\left(\mathbf{B}_k^{-1}\Delta\mathbf{f}_k - \Delta\mathbf{x}_k\right)\Delta\mathbf{x}_k^T\mathbf{B}_k^{-1}}{\Delta\mathbf{x}_k^T\mathbf{B}_k^{-1}\Delta\mathbf{f}_k}$.
7. Calcular $\mathbf{B}_{k+1} = \mathbf{B}_k + \mathbf{A}_k$.
8. Tomar $k = k + 1$ y regresar al paso 1.

Ejemplo 4.9
Resuelva el sistema de ecuaciones no lineales del ejemplo 4.7 utilizando el método de Broyden con búsqueda lineal y partiendo del punto $\mathbf{x}_0 = (1.0, 1.0, 1.0)^T$ y $\varepsilon = 0.01$.

Solución:

El mínimo es $\boldsymbol{\beta}^{*T} = (0.49, 0.00, -0.52)$ con $F(\mathbf{x}) = 0.000001$, la convergencia el método se obtuvo en $k = 13$ iteraciones principales.

Problemas

Para los siguientes 4 problemas utilice el método de mínimos cuadrados lineales.

4.1 Se realizó un experimento para medir la impedancia de un circuito $R-L$ en serie. La impedancia Z se da en función de la resistencia R, la frecuencia de la fuente f y la inductancia L como

$$Z^2 = R^2 + 4\pi^2 f^2 L^2$$

En el experimento se midió Z como función de f, las lecturas obtenidas se dan en la tabla.

$f(Hz)$	$Z(\Omega)$
123	7.4
158	8.4
194	9.1
200	9.6
229	10.3
245	10.5
269	11.4
292	11.9
296	12.2

a) Verifique si las observaciones se pueden interpretar en términos de una recta para todo el rango o parte de él.
b) Estime los coeficientes en la correlación

$$y = \beta_1 + \beta_2 x$$

donde $y = Z^2$ y $x = f^2$.

c) Calcule los mejores valores de L y de R .
d) Evalúe el coeficiente de correlación.

4.2 Se ha llevado a cabo un experimento para investigar la dependencia con la temperatura de la resistencia de un alambre de cobre. Un modelo común se representa por la ecuación

$$R = R_0\left(1+\alpha T\right)$$

donde R es la resistencia a la temperatura $T°C$, R_0 es el valor de la resistencia a $0°C$, y α es el coeficiente de temperatura de la resistencia. Las mediciones de R y T obtenidas se dan en la tabla de abajo.

R(Ω)	***T(°C)***
12.3	10
12.9	20
13.6	30
13.8	40
14.5	50
15.1	60
15.2	70
15.9	80

a) obtenga el mejor valor para los coeficientes.

b) obtenga el mejor valor de R_0 y α .

c) obtenga el coeficiente de correlación.

4.3 Las siguientes mediciones se efectuaron durante una investigación de fenómenos para los cuales no hay modelos disponibles. Identifique en cada caso una función adecuada y evalúe sus constantes.

a)

y	x
0.61	0.1
0.75	0.2
0.91	0.3
1.11	0.4
1.36	0.5
1.66	0.6
2.03	0.7
2.48	0.8
3.03	0.9

b)

y	x
94.8	2.0
87.9	5.0
81.3	8.0
74.9	11.0
68.7	14.0
64.0	17.0

4.4 Se obtuvieron los siguientes datos:

y	x
1.00	10.0
1.26	20.0
1.86	30.0
3.31	40.0
7.08	50.0

¿Cuál de los tres modelos siguientes representa mejor la relación entre las variables?.

a) $y = e^{\beta_1 + \beta_2 x}$ b) $y = e^{\beta_1 + \beta_2 x + \beta_3 x^2}$ c) $y = \beta_1 x^{\beta_2}$

En los 3 problemas que siguen, enuncie las variables o combinaciones de variables que deben graficarse para verificar la variación sugerida, y diga cómo puede encontrarse la incógnita o parámetros desconocidos del modelo (pendiente, ordenada al origen, etc.).

4.5 La velocidad de flujo de salida de un fluido ideal por un orificio en el lado de un tanque está dada por

$$v = \sqrt{\frac{2P}{\rho}}$$

donde v y P son las variables medidas. Determine ρ.

4.6 La ley de los gases para un gas ideal es

$$PV = nRT$$

donde P y T son las variables medidas; V y n son fijas y conocidas. Determine R.

4.7 La variación relativista de la masa con la velocidad es

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}}$$

donde m y v son variables medidas. Determine m_0 y c.

Resuelva el siguiente conjunto de 3 problemas de mínimos cuadrados no lineales. Proponga en cada caso el vector inicial y la tolerancia. Emplee en cada caso el método que juzgue más conveniente. Grafique las curvas o superficies con el software de que disponga.

4.8 Estime los valores de los parámetros β_1 y β_2 minimizando la suma de cuadrados de los residuales

$$F(\boldsymbol{\beta},t)=\sum_{i=1}^{n}\left(y_i-y(t_i)\right)^2$$

donde

$$y(t)=\frac{\beta_1}{\beta_1-\beta_2}\left(e^{-\beta_2 t}-e^{-\beta_1 t}\right)$$

para el siguiente conjunto de datos:

t	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
y	0.273	0.446	0.560	0.593	0.648	0.640	0.615	0.602	0.579

4.9 Estime los coeficientes en la correlación

$$y=\beta_1 x_1^{\beta_2} x_2^{\beta_3}$$

del siguiente conjunto de datos minimizando la suma de cuadrados de los residuales entre los valores y de datos experimentales y los ajustados.

x_1	2.0	6.0	9.0	2.5	4.5	9.5	8.0	4.0	3.0	7.0	6.5
x_2	36.0	8.0	3.0	6.25	7.84	1.44	4.0	7.0	9.0	2.0	5.0
y	46.5	591	1285	36.8	241	1075	1024	151	80	485	632

4.10 Estime los coeficientes para la siguiente correlación y conjunto de datos:

$$y = \beta_1 e^{\left(-\frac{1}{2}\left(\frac{x-\beta_2}{\beta_3}\right)^2\right)} + \beta_4 + \beta_5 x + \beta_6 x^2$$

n	y	x	n	y	x
1	185.1	1.176	16	218.0	52.44
2	176.0	4.436	17	291.1	55.22
3	164.2	10.23	18	277.4	57.48
4	155.8	13.56	19	205.3	59.74
5	145.3	18.92	20	179.6	60.46
6	135.0	24.98	21	91.37	64.42
7	128.0	27.42	22	83.39	65.45
8	118.0	34.56	23	76.71	67.48
9	110.3	39.00	24	70.53	70.06
10	106.4	42.25	25	72.68	70.65
11	105.8	44.19	26	67.90	74.46
12	106.2	46.83	27	60.31	82.14
13	105.3	46.26	28	58.96	85.18
14	173.7	51.12	29	56.63	89.38
15	195.0	51.84	30	54.10	94.55
			31	50.10	100.8

Resuelva los siguientes 3 sistemas de ecuaciones no lineales. Proponga en cada caso el vector inicial y la tolerancia. Emplee en cada caso el método que juzgue más conveniente. Grafique las superficies con el software de que disponga.

4.13 $\boldsymbol{f}(\mathbf{x}) = \begin{pmatrix} x_1 + x_2 - 3 \\ x_1^2 + x_2^2 - 9 \end{pmatrix}$, tiene más de una solución.

4.14 $\boldsymbol{f}(\mathbf{x}) = \begin{pmatrix} x_1^2 + x_2^2 - x_1 \\ x_1^2 - x_2^2 - x_2 \end{pmatrix}$, tiene más de una solución.

4.15 $\boldsymbol{f}(\mathbf{x}) = \begin{pmatrix} x_1 \\ x_2^2 + x_2 \\ e^{x_3} - 1 \end{pmatrix}$, tiene más de una solución.

CAPÍTULO 5

Fundamentos de optimización restringida

5.1 Introducción

Esta sección describe los conceptos relacionados a problemas de optimización restringidos. Las condiciones necesarias de optimalidad se explican e ilustran con ejemplos. Todos los puntos óptimos deben satisfacer estas condiciones.

El modelo de optimización de diseño general trata de hallar un vector

$$\mathbf{x}^T = \left(x_1, x_2, ..., x_n\right) \tag{5.1}$$

de variables independientes para minimizar un función objetivo

$$f\left(\mathbf{x}\right) = f\left(x_1, x_2, ..., x_n\right) \tag{5.2}$$

sujeta a las restricciones de igualdad

$$h_j\left(\mathbf{x}\right) = 0; \quad j = 1, ..., m \tag{5.3}$$

y las restricciones de desigualdad

$$g_k\left(\mathbf{x}\right) \leq 0; \quad k = 1, ..., l \tag{5.4}$$

las restricciones de desigualdad serán ignoradas inicialmente para discutir el Teorema de Lagrange dado en libros de cálculo. Luego se extenderá el teorema para las restricciones de desigualdad para obtener las condiciones necesarias de Karush-Kuhn-Tucker para el modelo general definido en las ecuaciones (5.2) a (5.4)

Basándose en la discusión de problemas de optimización no restringidos se podría concluir que sólo la naturaleza de la función objetivo $f(\mathbf{x})$ para los problemas restringidos, determinará la localización del punto mínimo. Sin embargo, esto no es del todo cierto, las funciones de restricción juegan un papel importante en la determinación de la solución óptima. Los siguientes ejemplos ilustran estas situaciones.

Ejemplo 5.1

$$\text{Minimizar} \quad f(\mathbf{x}) = (x_1 - 2)^2 + (x_2 - 2)^2$$

$$\begin{aligned} & g_1(\mathbf{x}) = x_1 + x_2 - 2 \leq 0 \\ \text{sujeta a:} \quad & g_2(\mathbf{x}) = -x_1 \leq 0 \\ & g_3(\mathbf{x}) = -x_2 \leq 0 \end{aligned}$$

Solución:

El conjunto de soluciones para el problema es la región triangular de color mostrada en la figura 5.1.

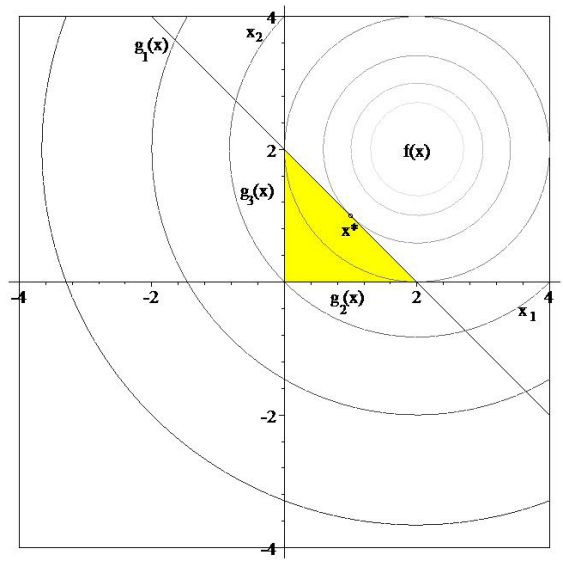


Figura 5.1 *Punto óptimo restringido.*

Si se ignoran las restricciones, $f(\mathbf{x})$ tiene un mínimo en el punto (2,2) que viola la restricción g_1. Obsérvese que los contornos de $f(\mathbf{x})$ son círculos concéntricos, ellos crecen en diámetro como el valor de la función $f(\mathbf{x})$ crece. Es claro que el valor mínimo para $f(\mathbf{x})$, corresponde a un círculo con el radio más pequeño intersectando la región factible (conjunto de restricciones), este es el punto (1,1) en el cual $f(\mathbf{x}) = 2$. El punto está en la frontera de la región factible. De esta manera, la localización del punto óptimo esta gobernada por las restricciones del problema planteado.

Ejemplo 5.2

$$\text{Minimizar } f(\mathbf{x}) = \left(x_1 - \frac{1}{2}\right)^2 + \left(x_2 - \frac{1}{2}\right)^2$$

$$\begin{aligned} g_1(\mathbf{x}) &= x_1 + x_2 - 2 \leq 0 \\ \text{sujeto a: } g_2(\mathbf{x}) &= -x_1 \leq 0 \\ g_3(\mathbf{x}) &= -x_2 \leq 0 \end{aligned}$$

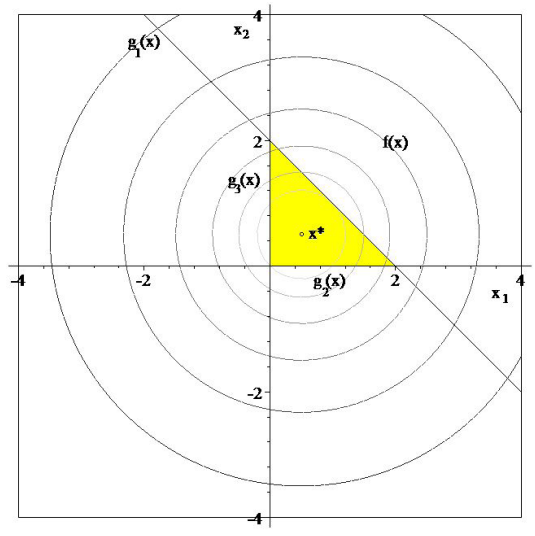


Figura 5.2 *Punto óptimo no restringido para un problema restringido.*

Solución:

El conjunto de restricciones es el mismo que en el ejemplo anterior. La función objetivo ha sido modificada, sin embargo, si se ignoran las restricciones, $f(\mathbf{x})$ tiene un mínimo en $(1/2,1/2)$ .Ya que el punto también satisface todas las restricciones, éste es la solución óptima. La solución para este problema ocurre en el interior de la región factible y las restricciones no juegan un papel importante en su localización.

Obsérvese que una solución a un problema de optimización restringido puede no existir. Esto sucede si se sobre restringe el sistema. Los requerimientos pueden ser conflictivos de manera que sea imposible construir un sistema para satisfacerlos. En tal caso, se debe reexaminar la formulación del problema y relajar las restricciones. El siguiente ejemplo ilustra la situación.

Ejemplo 5.3

$$\text{Minimizar } f(\mathbf{x}) = (x_1 - 2)^2 + (x_2 - 2)^2$$

$$\begin{aligned} &g_1(\mathbf{x}) = x_1 + x_2 - 2 \leq 0 \\ \text{sujeta a: } &g_2(\mathbf{x}) = -x_1 \leq 0 \\ &g_3(\mathbf{x}) = -x_2 \leq 0 \\ &g_4(\mathbf{x}) = -x_1 + x_2 + 3 \leq 0 \end{aligned}$$

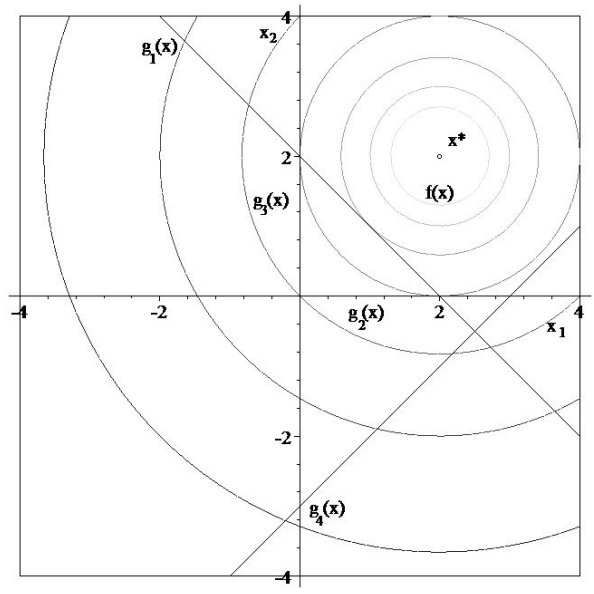


Figura 5.3 *Problema infactible.*

Solución:

La figura 5.3 muestra una gráfica de las restricciones para el problema. Puede verse que no hay un punto que satisfaga todas las restricciones en el primer cuadrante y por lo tanto no hay solución, es decir, no hay punto factible. $\mathbf{x}^*$ es el óptimo no restringido.

5.2 Condiciones para la minimización restringida

En esta ocasión, se inicia con el estudio de las condiciones necesarias y suficientes que se satisfacen en puntos solución. Estas condiciones además de su valor intrínseco para caracterizar soluciones, definen a los multiplicadores de Lagrange y a cierta matriz Hessiana que, considerados juntos, forman la base del desarrollo y el análisis de los algoritmos que se utilizan para resolver problemas de optimización restringidos.

Restricciones

Los problemas que se tratan son los del tipo general de programación no lineal de la forma

$$\begin{aligned} &\text{Minimizar } f(\mathbf{x}) \\ &\text{sujeta a: } \begin{aligned} h_j(\mathbf{x}) &= 0; \quad j = 1,...,m \\ g_k(\mathbf{x}) &\leq 0; \quad k = 1,...,l \\ \mathbf{x} &\in E^n \end{aligned} \end{aligned} \tag{5.5}$$

donde $m \leq n$, y las funciones f, h_j y g_k son continuas y en general, se supone que tienen segundas derivadas parciales continuas. Por simplicidad de notación, se introducen las funciones con valores vectoriales $\mathbf{h}^T = (h_1,...,h_m)$ y $\mathbf{g}^T = (g_1,...,g_l)$ y se vuelve a escribir (5.5) como

$$\begin{aligned} &\text{Minimizar } f(\mathbf{x}) \\ &\text{sujeta a: } \begin{aligned} \mathbf{h}(\mathbf{x}) &= \mathbf{0} \\ \mathbf{g}(\mathbf{x}) &\leq \mathbf{0} \\ \mathbf{x} &\in E^n \end{aligned} \end{aligned} \tag{5.6}$$

El Plano Tangente

Un conjunto de restricciones de igualdad en E^n

$$\mathbf{h}(\mathbf{x}) = \mathbf{0} \tag{5.7}$$

define un subconjunto de E^n, que se considera mejor tratándolo como una hipersuperficie. Si las restricciones son regulares en todas partes, de acuerdo como lo que se expone más adelante, esta hipersuperficie tiene dimensión $n - m$. Si como se supone en este momento, las funciones h_j, $j = 1,...,m$ tienen primeras derivadas parciales continuas, la superficie definida por ellas se denomina uniforme.

Asociado a un punto de una superficie uniforme está el plano tangente en ese punto, término que en dos o tres dimensiones tiene un significado evidente. Para formalizar la noción general, se comienza definiendo curvas en una superficie.

Definición 5.1

Una curva en una superficie S es una familia de puntos $\mathbf{x}(t) \in S$ continuamente parametrizados por t para $a \leq t \leq b$.

La curva es diferenciable si existe $\dot{\mathbf{x}}(t)$, y es doblemente diferenciable si existe $\ddot{\mathbf{x}}(t)$. Se dice que una curva $\mathbf{x}(t)$ pasa por el punto $\mathbf{x}^*$ si $\mathbf{x}^* = \mathbf{x}(t^*)$ para alguna t^*, $a \leq t^* \leq b$, la derivada de la curva en $\mathbf{x}^*$ está lógicamente definida como $\dot{\mathbf{x}}(t^*)$, que es un vector de E^n.

Considérense ahora todas las curvas diferenciables en S que pasan por un punto $\mathbf{x}^*$. El plano tangente en $\mathbf{x}^*$ se define como el conjunto de las derivadas de todas las curvas diferenciables en $\mathbf{x}^*$. El plano tangente es un subespacio de E^n. Para las superficies definidas por un conjunto de relaciones de restricción como (5.7), el problema de obtener una representación explícita para el plano tangente es un problema fundamental que se estudia a continuación. Idealmente, sería deseable expresar este plano tangente desde el punto de vista de las derivadas de las funciones h_j que definen la superficie.

Definición 5.2 Un punto $\mathbf{x}^*$ que satisfaga la restricción $\mathbf{h}\left(\mathbf{x}^*\right)=\mathbf{0}$, se denomina punto regular de la restricción si los vectores gradientes $\nabla h_1\left(\mathbf{x}^*\right),...,\nabla h_m\left(\mathbf{x}^*\right)$ son linealmente independientes.

La independencia lineal significa que no hay dos gradientes que sean paralelos entre sí, y ningún gradiente puede expresarse como una combinación lineal de los otros.

Teorema 5.1 En un punto regular $\mathbf{x}^*$ de la superficie S definida por $\mathbf{h}\left(\mathbf{x}\right)=\mathbf{0}$, el plano tangente es igual a

$$T=\left\{\mathbf{y}:\nabla\mathbf{h}^T\left(\mathbf{x}^*\right)\mathbf{y}=\mathbf{0}\right\}$$

Los multiplicadores de Lagrange tienen un significado geométrico así como físico. Sus valores dependen de la forma de la función objetivo y de las restricciones, si estas funciones cambian, el valor de los multiplicadores de Lagrange también cambia.

Para introducir la idea de los multiplicadores de Lagrange, se considerará el siguiente ejemplo de minimizar una función objetivo de dos variables con una ecuación de restricción de igualdad.

Ejemplo 5.4 Hallar $\mathbf{x}^*$ para

$$\text{minimizar } f\left(\mathbf{x}\right)=\left(x_1-2\right)^2+\left(x_2-2\right)^2 \tag{5.8}$$

$$\text{sujeta a } \ h\left(\mathbf{x}\right)=x_1+x_2-2=0 \tag{5.9}$$

Solución:

Resolviendo el problema despejando una de las variables y optimizando con respecto a la variables independiente resulta

$$\mathbf{x}^* = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \text{ y } f\left(\mathbf{x}^*\right) = 2$$

La línea recta $h\left(\mathbf{x}\right)$ representa la restricción de igualdad y la región factible para el problema. Por lo tanto la solución óptima debe caer sobre la línea. La función objetivo es la ecuación de un círculo con su centro en el punto (2,2). Los contornos de la función dependen del valor de $f\left(\mathbf{x}\right)$. Puede verse que el punto $\mathbf{x}^*$ con coordenadas (1,1) da la solución óptima para el problema. El contorno de valor 2 de la función objetivo toca justo a la línea $h\left(\mathbf{x}\right)$, así, este es el valor mínimo para la función objetivo.

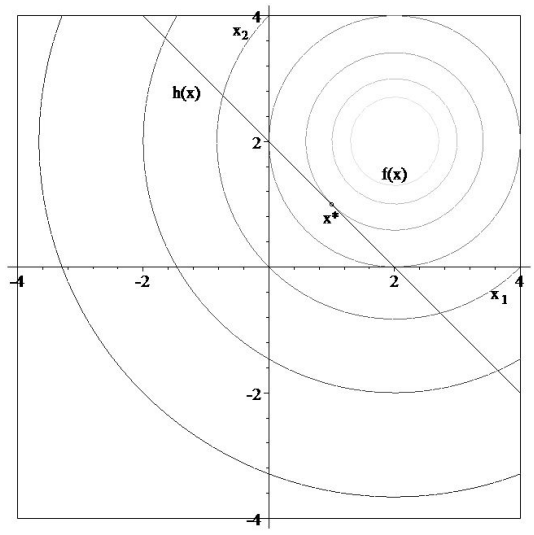


Figura 5.4 *Problema con una restricción de igualdad.*

Reducción de Variables e Introducción a los Multiplicadores de Lagrange

Ahora se verá que condiciones matemáticas se satisfacen en el punto mínimo $\mathbf{x}^*$ del ejemplo anterior. Sea el punto óptimo representado como $\mathbf{x}^* = \left(x_1^*, x_2^*\right)$. Para derivar las condiciones e introducir el multiplicador de Lagrange, primero se supondrá que la restricción de igualdad puede usarse para resolver para una variable en términos de la otra (al menos simbólicamente), es decir, se asume que se puede escribir

$$x_2 = \Phi\left(x_1\right) \tag{5.10}$$

donde Φ es una función apropiada de x_1. En muchos problemas, puede que no sea posible escribir explícitamente la función $\Phi\left(x_1\right)$, pero para propósitos de derivación, se asumirá su existencia. Será visto posteriormente que la forma explícita de la función no es necesaria. Para el ejemplo 5.4, $\Phi\left(x_1\right)$ de la ecuación (5.9) está dada como

$$\Phi\left(x_1\right) = -x_1 + 2 \tag{5.11}$$

Sustituyendo la ecuación (5.10) en la ecuación (5.8), se elimina x_2 de la función objetivo y se obtiene el problema de minimización no restringido en términos de x_1, solamente:

$$\text{Minimizar } f\left(x_1, \Phi\left(x_1\right)\right) \tag{5.12}$$

para el ejemplo 5.4, sustituyendo la ecuación (5.11) en la ecuación (5.8) se tiene

$$f\left(x_1\right) = \left(x_1 - 2\right)^2 + \left(-x_1\right)^2 = 2x_1^2 - 4x_1 + 4$$

la condición necesaria

$$\frac{df}{dx_1} = 2\left(x_1 - 2\right) + 2x_1 = 4x_1 - 4 = 0$$

da $x_1^* = 1$. Luego la ecuación (5.11) da $x_2^* = 1$ y la función objetivo en el punto (1,1) es 2. Puede verificarse que la condición suficiente

$$\frac{d^2 f}{dx_1^2} = 4 > 0$$

se satisface. Así el punto $\mathbf{x}^* = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ es verdaderamente un mínimo. Si se supone que la forma explícita de la función $\Phi(x_1)$ no puede obtenerse (que es el caso generalmente), entonces, algún procedimiento alternativo debe desarrollarse para obtener la solución óptima. Se derivará tal procedimiento y se verá que el *multiplicador de Lagrange* para la restricción se define naturalmente en el proceso. Usando la regla de la cadena de la diferenciación, se escribe la condición necesaria

$$\frac{df}{dx_1} = 0$$

para el problema definido en la ecuación (5.12) como

$$\frac{df}{dx_1} = \frac{\partial f(x_1, x_2)}{\partial x_1} + \frac{\partial f(x_1, x_2)}{\partial x_2} \frac{dx_2}{dx_1}$$

sustituyendo la ecuación (5.10), la ecuación anterior puede escribirse en el punto óptimo (x_1^*, x_2^*) como

$$\frac{\partial f(x_1^*, x_2^*)}{\partial x_1} + \frac{\partial f(x_1^*, x_2^*)}{\partial x_2} \frac{d\Phi}{dx_1} = 0 \qquad (5.13)$$

ya que Φ no se conoce, se necesita eliminar $d\Phi/dx_1$ de la ecuación (5.13). Para realizar esto, se deriva la ecuación de restricción $h(x_1, x_2) = 0$ en el punto (x_1^*, x_2^*) como

$$\frac{dh}{dx_1} = \frac{\partial h(x_1^*, x_2^*)}{\partial x_1} + \frac{\partial h(x_1^*, x_2^*)}{\partial x_2}\frac{d\Phi}{dx_1} = 0$$

y despejando $\frac{d\Phi}{dx_1}$, se obtiene

$$\frac{d\Phi}{dx_1} = -\frac{\dfrac{\partial h(x_1^*, x_2^*)}{\partial x_1}}{\dfrac{\partial h(x_1^*, x_2^*)}{\partial x_2}} \tag{5.14}$$

ahora, sustituyendo $\frac{d\Phi}{dx_1}$, de la ecuación (5.14) en (5.13), se obtiene

$$\frac{\partial f}{\partial x_1} - \frac{\partial f}{\partial x_2}\left(\frac{\dfrac{\partial h}{\partial x_1}}{\dfrac{\partial h}{\partial x_2}}\right) = 0 \tag{5.15}$$

si se define una cantidad λ como

$$\lambda \equiv -\frac{\dfrac{\partial f}{\partial x_2}}{\dfrac{\partial h}{\partial x_2}} \tag{5.16}$$

y sustituyendo ésta en la ecuación (5.15), se tiene

$$\frac{\partial f}{\partial x_1} + \lambda \frac{\partial h}{\partial x_1} = 0 \tag{5.17}$$

También, rearreglando la ecuación (5.16) que define λ, se obtiene

$$\frac{\partial f}{\partial x_2} + \lambda \frac{\partial h}{\partial x_2} = 0 \tag{5.18}$$

Las ecuaciones (5.17) y (5.18) junto con la restricción de igualdad $h(\mathbf{x}) = 0$, son las *condiciones necesarias de optimalidad.* Cualquier punto que viole estas condiciones no puede ser un punto mínimo para el problema. Para el ejemplo 5.4 estas condiciones dan

$$\begin{aligned} 2(x_1 - 2) + \lambda &= 0 \\ 2(x_2 - 2) + \lambda &= 0 \\ x_1 + x_2 - 2 &= 0 \end{aligned}$$

la solución a este conjunto de ecuaciones es $\mathbf{x}^{*T} = (1,1)$ y $\lambda^* = 2$.

Significado Geométrico de los Multiplicadores de Lagrange

Se acostumbra usar lo que se conoce como la *función de Lagrange* para escribir las condiciones necesarias. La función de Lagrange se denota con la letra $\boldsymbol{L}$ y se define usando las funciones objetivo y de restricciones como

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda h(\mathbf{x}) \tag{5.19}$$

se ha observado que las condiciones necesarias de las ecs. (5.17) y (5.18) están dadas en términos de $\boldsymbol{L}$ como

$$\nabla L\left(\mathbf{x}^*\right)=\mathbf{0} \tag{5.20}$$

escribiendo esta condición usando la ecuación (5.19) se obtiene

$$\nabla f\left(\mathbf{x}^*\right)+\lambda\nabla h\left(\mathbf{x}^*\right)=\mathbf{0} \tag{5.21}$$

que puede reescribirse de la siguiente manera

$$\nabla f\left(\mathbf{x}^*\right)=-\lambda\nabla h\left(\mathbf{x}^*\right) \tag{5.22}$$

Esta ecuación muestra que en el punto mínimo candidato, *los gradientes de las funciones objetivo y de restricciones están a lo largo de la misma línea y son proporcionales.* (Obsérvese (5.22) con el ejemplo dado). El concepto de los multiplicadores de Lagrange es bastante general. El multiplicador de Lagrange puede interpretarse para una restricción como una fuerza requerida para imponer la restricción.

Condiciones Necesarias: Restricciones de Desigualdad

Si ahora se incluyen restricciones de desigualdad de la forma

$$g_k\left(\mathbf{x}\right)\leq 0 \qquad k=1,...,l$$

en el problema de optimización general, se puede transformar una restricción de desigualdad a una igualdad adicionando una nueva variable a ésta, llamada la *variable de holgura*, debido a que la restricción es de la forma "≤", esto es, su valor es negativo o nulo. De esta manera, la variable de holgura siempre deberá ser no negativa, es decir, nula o positiva respectivamente, para hacer de la desigualdad una igualdad. Una restricción de desigualdad

$$g_k(\mathbf{x}) \leq 0$$

es equivalente a la restricción de igualdad

$$g_k(\mathbf{x}) + s_k = 0 \ , \ s_k \geq 0$$

Las variables s_k son tratadas como incógnitas del problema de optimización junto con las variables originales. Sus valores deben determinarse como una parte de la solución. Cuando la variable s_k tiene valor nulo ($s_k = 0$), la correspondiente restricción de desigualdad se satisface como igualdad. Tal desigualdad es llamada una *restricción activa* (o *ajustada*), es decir, no hay *holgura* en la restricción. Para cualquier $s_k > 0$, la correspondiente restricción es una desigualdad estricta. Esta es llamada una *restricción inactiva* (o *pasiva*), tiene holgura dada por s_k.

Obsérvese que en el procedimiento anterior, debe introducirse una variable adicional s_k y una restricción adicional $s_k \geq 0$ para tratar cada restricción de desigualdad. Esto incrementa la dimensión del problema de optimización. La restricción ($s_k \geq 0$) puede evitarse si se usa s_k^2 como la variable de holgura en lugar de s_k. Por tanto, la desigualdad $g_k(\mathbf{x}) \leq 0$ es convertida a igualdad como

$$g_k(\mathbf{x}) + s_k^2 = 0 \tag{5.23}$$

donde s_k puede tener cualquier valor real. Esta forma puede usarse en el Teorema de Multiplicadores de Lagrange para tratar restricciones de desigualdad y derivar las condiciones necesarias correspondientes.

Las l nuevas ecuaciones necesarias para determinar las variables de holgura, son obtenidas requiriendo que el Lagrangiano sea estacionario con respecto a las variables de holgura

$$\left(\frac{\partial L}{\partial \mathbf{s}} = \mathbf{0} \right)$$

Obsérvese que una vez que un punto óptimo se determina, la ecuación (5.23) se utiliza para calcular la variable de holgura s_k^2. Si la restricción se satisface en el punto, es decir, $g_k(\mathbf{x}) \leq 0$, entonces $s_k^2 \geq 0$. Si esta se viola, entonces $s_k^2 < 0$, lo cual no es aceptable, es decir, el punto no es punto mínimo candidato.

Hay una condición necesaria adicional para los multiplicadores de Lagrange de las restricciones del tipo "≤" dadas como

$$\mu_k^* \geq 0\ ,\ k = 1,...,l$$

donde μ_k^* es el multiplicador de Lagrange para la k-ésima restricción de igualdad. De esta manera, el multiplicador de Lagrange para cada restricción de desigualdad del tipo "≤" debe ser no negativa. Si la restricción es inactiva en el óptimo, su multiplicador de Lagrange asociado es cero. Si ésta es activa ($g_k(\mathbf{x}) = 0$), entonces el multiplicador asociado debe ser no negativo.

Las condiciones necesarias para las restricciones de igualdad y desigualdad pueden resumirse en lo que son comúnmente conocidas como las *condiciones necesarias de Kuhn-Tucker* (K-T). Aunque estas condiciones pueden expresarse en diferentes formas, discutiremos solamente la que se va a dar, es decir, aquélla que define a las variables de holgura.

Teorema 5.2 Condiciones Necesarias de Kuhn-Tucker (K-T)

Sea $\mathbf{x}^*$ un punto regular del conjunto de restricciones, es decir, un mínimo local para $f(\mathbf{x})$ sujeta a las restricciones

$$h_j(\mathbf{x}) = 0\ ,\ j = 1,...,m$$
$$g_k(\mathbf{x}) \leq 0\ ,\ k = 1,...,l$$

se define la función de Lagrange para el problema de optimización como

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \mathbf{s}) = f(\mathbf{x}) + \sum_{j=1}^{m} \lambda_j h_j(\mathbf{x}) + \sum_{k=1}^{l} \mu_k \left(g_k(\mathbf{x}) + s_k^2 \right)$$

o bien

$$L(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu},\mathbf{s}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x}) + \boldsymbol{\mu}^T \left(\mathbf{g}(\mathbf{x}) + \mathbf{s}^2\right) \tag{5.24}$$

luego existen multiplicadores de Lagrange $\boldsymbol{\lambda}^*$ y $\boldsymbol{\mu}^*$ tales que el Lagrangiano es estacionario con respecto a x_i, λ_j, μ_k y s_k, es decir

$$\frac{\partial L}{\partial x_i} = \frac{\partial f}{\partial x_i} + \sum_{j=1}^{m} \lambda_j^* \frac{\partial h_j}{\partial x_i} + \sum_{k=1}^{l} \mu_k^* \frac{\partial g_k}{\partial x_i} = 0 \ , \ i = 1,...,n \tag{5.25}$$

$$\frac{\partial L}{\partial \lambda_j} = h_j\left(\mathbf{x}^*\right) = 0 \ , \ j = 1,...,m \tag{5.26}$$

$$\frac{\partial L}{\partial \mu_k} = g_k\left(\mathbf{x}^*\right) + s_k^2 = 0 \ , \ k = 1,...,l \tag{5.27}$$

$$\frac{\partial L}{\partial s_k} = 2\mu_k^* s_k = 0 \ , \ k = 1,...,l \tag{5.28}$$

$$\mu_k^* \geq 0 \ , \ k = 1,...,l \tag{5.29}$$

y donde todas las derivadas se evalúan en el punto $\mathbf{x}^*$.

Las condiciones anteriores son llamadas algunas veces las *condiciones necesarias de primer orden*. Es importante entender su uso para: 1) verificar la posible optimalidad de un punto dado y, 2) determinar los puntos estacionarios.

Obsérvese primero de las ecuaciones (5.26) y (5.27) que *el punto estacionario debe ser factible*, así deben tenerse en cuenta todas las restricciones para asegurar su satisfacción. Las condiciones de los gradientes de la ecuación (5.25) también deben satisfacerse simultáneamente. Estas condiciones tienen un significado geométrico. Para ver esto, reescribimos la ecuación (5.25) como

$$-\frac{\partial f}{\partial x_i} = \sum_{j=1}^{m} \lambda_j^* \frac{\partial h_j}{\partial x_i} + \sum_{k=1}^{l} \mu_k^* \frac{\partial g_k}{\partial x_i} \ , \ i = 1,...,n \qquad (5.30)$$

que muestra que en el punto estacionario, la dirección del gradiente negativo para la función objetivo, es una combinación lineal de los gradientes de las restricciones con multiplicadores de Lagrange como los parámetros escalares de la combinación lineal.

Las l condiciones en la ecuación (5.28) se conocen como *las condiciones de cambio o las condiciones de relajamiento complementarias*. Ellas pueden satisfacerse imponiendo ya sea $s_k = 0$ (cero holgura implica desigualdades activas, es decir $g_k\left(\mathbf{x}^*\right) = 0$), o $\mu_k = 0$ (en este caso g_k debe ser ≤ 0 para satisfacer la factibilidad). Estas condiciones establecen varios casos en cálculos reales y su uso debe entenderse claramente. El número de condiciones de cambio es igual al número de restricciones de desigualdad para el problema. Diferentes combinaciones de estas condiciones pueden dar muchos casos de solución. En general, con l restricciones de desigualdad, las condiciones de cambio guían a 2^l casos de solución normales distintos (caso anormal es aquél en donde $\mu_k = 0$ y $s_k = 0$).

Para cada caso, se requiere resolver las condiciones necesarias restantes para puntos estacionarios. Dependiendo de las funciones del problema, puede ser o no posible resolver analíticamente las condiciones necesarias de cada caso. Si las funciones son no lineales, se tendrá que usar los métodos numéricos para hallar sus raíces. En resumen, cada caso puede dar puntos estacionarios.

Para problemas generales, las incógnitas son $\mathbf{x}$,$\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ y $\mathbf{s}$. Estos son vectores n-, m- , l- y l- dimensionales respectivamente. De este modo, hay $(n+m+2l)$ variables desconocidas y necesitamos $(n+m+2l)$ ecuaciones para determinarlas. Las ecuaciones necesarias de Kuhn-Tucker. Estas ecuaciones deben satisfacerse simultáneamente para puntos estacionarios. Después de que las soluciones se han hallado, las condiciones necesarias restantes de las ecuaciones (5.29) deberán verificarse.

Condiciones de Segundo Orden para Optimización Restringida

Las soluciones de las condiciones necesarias son puntos estacionarios.

Las condiciones de suficiencia determinan si un punto estacionario es verdaderamente un mínimo local o no. En esta sección, se discutirán las condiciones necesarias de segundo orden y de suficiencia para problemas de optimización restringidos. Primero se abordarán las condiciones suficientes para problemas de programación convexa y luego para problemas de optimización más generales.

Condiciones Suficientes para Problemas Convexos

Para problemas de programación convexa, las condiciones necesarias de primer orden de Kuhn-Tucker también resultan ser suficientes. De esta manera, si se muestra la convexidad de un problema cualquier solución de las condiciones necesarias, satisfacerán automáticamente las condiciones suficientes. En adición, la solución será un mínimo global.

Teorema 5.3 Condición Suficiente para un Problema Convexo

Si $f(\mathbf{x})$ es una función objetivo convexa definida sobre una región factible convexa (conjunto de restricciones), entonces las condiciones de primer orden de Kuhn-Tucker son necesarias así como suficientes para un mínimo global.

Condiciones de Segundo Orden para Problemas Generales

Como en el caso no restringido, se puede usar la *información de segundo orden* acerca de las funciones (es decir, la curvatura) en el punto estacionario $\mathbf{x}^*$ para determinar si este es verdaderamente un mínimo local. Recuérdese que la suficiencia local para el problema no restringido requiere que la parte cuadrática de la expansión en serie de Taylor para la función en $\mathbf{x}^*$ sea positiva para todos los cambios $\mathbf{d}$ no nulos. En el caso restringido, también deben considerarse restricciones activas en $\mathbf{x}^*$ para determinar cambios factibles $\mathbf{d}$. Se considerarán sólo los puntos $\mathbf{x}=\mathbf{x}^*+\mathbf{d}$ en la vecindad de $\mathbf{x}^*$ que satisfacen las ecuaciones de restricción activas. Cualquier $\mathbf{d} \neq \mathbf{0}$ satisfaciendo restricciones activas de primer orden debe estar en el plano tangente a la restricción (véase la figura 5.5)

Tales vectores $\mathbf{d}$ son entonces ortogonales a los gradientes de las restricciones activas (los gradientes de las restricciones activas son normales al plano tangente de la restricción). Por lo tanto, el producto punto de $\mathbf{d}$ con cada uno de los gradientes de las restricciones ∇h_j y ∇g_k deben ser cero, es decir,

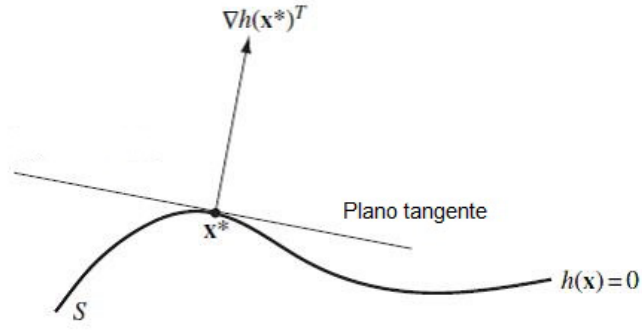
$$\nabla h_j^T \cdot \mathbf{d} = 0$$
$$\nabla g_k^T \cdot \mathbf{d} = 0$$

De este modo, las direcciones $\mathbf{d}$ son determinadas para definir una región factible alrededor del punto $\mathbf{x}^*$. Obsérvese que sólo las restricciones de desigualdad activas son usadas en determinar $\mathbf{d}$.

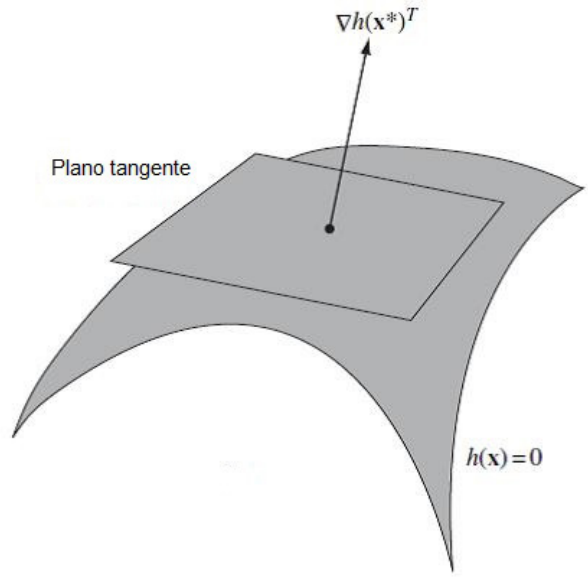
Para derivar las condiciones de segundo orden, se escribirá la expansión en serie de Taylor de la función de Lagrange y se considerará solo aquella $\mathbf{d}$ que satisfaga las condiciones anteriores, $\mathbf{x}^*$ es entonces un punto mínimo local si el término de segundo orden de la expansión en serie de Taylor es positiva para todo $\mathbf{d}$ en el plano tangente de la restricción. Esta es entonces la condición suficiente. Como una condición necesaria, el término de segundo orden debe ser no negativo.

Se resumen estos resultados en los siguientes teoremas:

a)



b)



c)

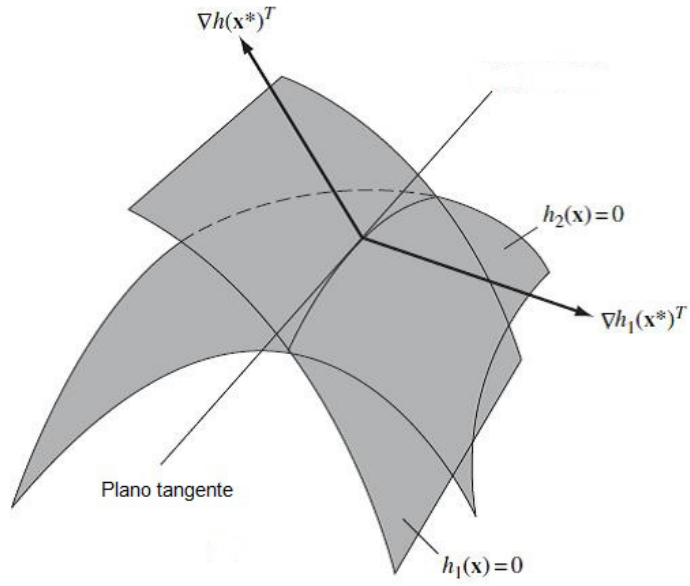


Figura 5.5 *Planos tangentes (a), (b) y (c) a las restricciones en* x^* .

Teorema 5.4 Condiciones Necesarias de Segundo Orden

Sea $\mathbf{x}^*$ un punto que satisface las condiciones necesarias de K-T.

Se define la matriz Hessiana de la función de Lagrange L en $\mathbf{x}^*$ como

$$\nabla^2 L = \nabla^2 f + \sum_{j=1}^{m} \lambda_j^* \nabla^2 h_j + \sum_{k=1}^{l} \mu_k^* \nabla^2 g_k \qquad (5.31)$$

Sea $\mathbf{d}$ una dirección factible no nula ($\mathbf{d} \neq \mathbf{0}$) que satisface al siguiente sistema lineal en el punto $\mathbf{x}^*$:

$$\nabla h_j^T \cdot \mathbf{d} = 0 \ , \ j = 1,...,m \qquad (5.32)$$

y

$$\nabla g_k^T \cdot \mathbf{d} = 0 \ , \ \forall \ g_k(\mathbf{x}) \ \text{ activa} \qquad (5.33)$$

entonces, si $\mathbf{x}^*$ es un punto mínimo local para el problema de optimización debe satisfacerse que

$$\mathbf{d}^T \cdot \nabla^2 L \cdot \mathbf{d} \geq 0 \qquad (5.34)$$

Obsérvese que cualquier punto que no satisface las condiciones necesarias de segundo orden, no puede ser un punto mínimo local.

Teorema 5.5 Condiciones Suficientes de Segundo Orden

Sea $\mathbf{x}^*$ un punto que satisface las condiciones necesarias de primer orden de K-T. Sea la matriz Hessiana de la función Lagrangiana en $\mathbf{x}^*$ como

$$\nabla^2 L = \nabla^2 f + \sum_{j=1}^{m} \lambda_j^* \nabla^2 h_j + \sum_{k=1}^{l} \mu_k^* \nabla^2 g_k \qquad (5.35)$$

Definimos direcciones factibles no nulas ($\mathbf{d} \neq \mathbf{0}$) como soluciones de los sistemas lineales

$$\nabla h_j^T \cdot \mathbf{d} = 0\,,\ j = 1,...,m \qquad (5.36)$$

y

$$\nabla g_k^T \cdot \mathbf{d} = 0\,,\ k = 1,...,l \qquad (5.37)$$

para desigualdades activas con $\mu_k > 0$.

También sea $\nabla g_k{}^T \cdot \mathbf{d} \leq 0$ para aquellas restricciones con $\mu_k = 0$. Si

$$\mathbf{d}^T \cdot \nabla^2 L\left(\mathbf{x}^*\right) \cdot \mathbf{d} > 0 \qquad (5.38)$$

entonces, $\mathbf{x}^*$ es un mínimo local aislado (aislado significa que no hay otros puntos mínimos locales en la cercanía o vecindad de $\mathbf{x}^*$).

Obsérvese primero la diferencia en las condiciones para las direcciones $\mathbf{d}$ en (5.33) para la condición necesaria y (5.37) para la condición suficiente. En (5.33) todas las desigualdades activas con multiplicadores no negativos están incluidas, mientras que en (5.37) sólo aquellas desigualdades activas con un multiplicador positivo son incluidas.

Condiciones Necesarias de Primer Orden: Restricciones de Igualdad

La deducción de condiciones necesarias y suficientes para que un punto sea un punto mínimo local sujeto a las restricciones de igualdad es bastante sencilla, ahora que se conoce la representación del plano tangente. Se comienza deduciendo las condiciones necesarias de primer orden.

Lema 5.1 Sea $\mathbf{x}^*$ un punto regular de las restricciones $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ y un punto extremo local (mínimo o máximo) de $f(\mathbf{x})$ sujeto a estas restricciones. Entonces todo $\mathbf{y} \in E^n$ que cumpla

$$\nabla \mathbf{h}^T(\mathbf{x}^*)\mathbf{y} = \mathbf{0} \tag{5.39}$$

debe cumplir también

$$\nabla^T f(\mathbf{x}^*)\mathbf{y} = 0 \tag{5.40}$$

El lema anterior expresa que $\nabla f(\mathbf{x}^*)$ es ortogonal al plano tangente. A continuación, se concluye que esto implica que $\nabla f(\mathbf{x}^*)$ es una combinación lineal de los gradientes de $\mathbf{h}$ en $\mathbf{x}^*$, una relación que da lugar a la introducción de los multiplicadores de Lagrange.

Teorema 5.6

Sea $\mathbf{x}^*$ un punto extremo local de $f(\mathbf{x})$ sujeto a las restricciones $\mathbf{h}(\mathbf{x}) = \mathbf{0}$. Supóngase, además que $\mathbf{x}^*$ es un punto regular de estas restricciones. Entonces, existe un $\boldsymbol{\lambda} \in E^m$ tal que

$$\nabla f(\mathbf{x}^*) + \boldsymbol{\lambda}^T \nabla \mathbf{h}^T(\mathbf{x}^*) = \mathbf{0} \tag{5.41}$$

Obsérvese que las condiciones necesarias de primer orden (5.41) junto con las restricciones

$$\mathbf{h}\left(\mathbf{x}^*\right)=\mathbf{0}$$

proporcionan un total de $n+m$ ecuaciones (en general, no lineales) en las $n+m$ variables que comprenden $\mathbf{x}^*$, $\boldsymbol{\lambda}$. Así, las condiciones necesarias son un conjunto completo, pues, al menos localmente, determinan una solución única.

Conviene introducir el Lagrangiano asociado al problema con restricciones definido como

$$l\left(\mathbf{x},\boldsymbol{\lambda}\right)\equiv f\left(\mathbf{x}\right)+\boldsymbol{\lambda}^T\mathbf{h}\left(\mathbf{x}\right) \tag{5.42}$$

entonces, las condiciones necesarias se pueden expresar en las formas

$$\nabla_{\mathbf{x}} l\left(\mathbf{x},\boldsymbol{\lambda}\right)=\mathbf{0} \tag{5.43}$$

$$\nabla_{\boldsymbol{\lambda}} l\left(\mathbf{x},\boldsymbol{\lambda}\right)=\mathbf{0} \tag{5.44}$$

la segunda no es más que un nuevo planteamiento de las restricciones. Se dejará, por un momento, el desarrollo matemático para considerar algunos ejemplos de problemas de optimización con restricciones.

Ejemplo 5.5: Considérese el problema

Minimizar $f\left(\mathbf{x}\right)=x_1x_2+x_2x_3+x_1x_3$

Sujeta a: $h\left(\mathbf{x}\right)=x_1+x_2+x_3-3=0$

Solución:

Las condiciones necesarias se convierten en

$$x_2 + x_3 + \lambda = 0$$

$$x_1 + x_3 + \lambda = 0$$

$$x_1 + x_2 + \lambda = 0$$

estas tres ecuaciones, junto con la restricción, proporcionan cuatro ecuaciones que se pueden resolver para las cuatro incógnitas x_1, x_2, x_3 y λ. La solución da

$$\mathbf{x}^{*T} = (1,1,1) \text{ y } \lambda^* = -2$$

Condiciones de Segundo Orden

Con un argumento análogo al utilizado para el caso sin restricciones, también se pueden deducir las correspondientes condiciones de segundo orden para problemas con restricciones. En esta sección, se supone que $f(\mathbf{x}), \mathbf{h}(\mathbf{x}) \in C^2$.

Teorema 5.7 Condiciones Necesarias de Segundo Orden

Supóngase que $\mathbf{x}^*$ es un mínimo local de $f(\mathbf{x})$ sujeto a $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ y que $\mathbf{x}^*$ es un punto regular de estas restricciones. Entonces, existe un $\boldsymbol{\lambda} \in E^m$ tal que

$$\nabla f(\mathbf{x}) + \boldsymbol{\lambda}^T \nabla \mathbf{h}^T(\mathbf{x}) = \mathbf{0} \qquad (5.45)$$

Si se representa por T el plano tangente, $T = \left\{ \mathbf{y} : \nabla \mathbf{h}^T \left(\mathbf{x} \right) \mathbf{y} = \mathbf{0} \right\}$ entonces la matriz

$$\mathbf{L}\left(\mathbf{x}^* \right) = \mathbf{F}\left(\mathbf{x}^* \right) + \boldsymbol{\lambda}^T \mathbf{H}\left(\mathbf{x}^* \right) \tag{5.46}$$

es semidefinida positiva en T, esto es, $\mathbf{y}^T \mathbf{L}\left(\mathbf{x}^* \right) \mathbf{y} \geq 0$ para todo $\mathbf{y} \in T$.

El teorema anterior es el primer encuentro con la matriz $\mathbf{L}=\mathbf{F}+\boldsymbol{\lambda}^T\mathbf{H}$ que es la matriz de las segundas derivadas parciales con respecto a $\mathbf{x}$, del Lagrangiano $l(\mathbf{x})$. Esta matriz es la base de la teoría de algoritmos para problemas con restricciones y se encontrará a menudo en las secciones posteriores. A continuación, se enuncia el conjunto correspondiente de condiciones suficientes.

Teorema 5.8 Condiciones de Suficiencia de Segundo Orden

Supóngase que hay un punto $\mathbf{x}^*$ que satisface $\mathbf{h}\left(\mathbf{x}^* \right) = \mathbf{0}$ y un $\boldsymbol{\lambda} \in E^m$ tal que

$$\nabla f\left(\mathbf{x}^* \right) + \boldsymbol{\lambda}^T \nabla \mathbf{h}^T \left(\mathbf{x}^* \right) = \mathbf{0} \tag{5.47}$$

Supóngase también que la matriz $\mathbf{L}\left(\mathbf{x}^* \right) = \mathbf{F}\left(\mathbf{x}^* \right) + \boldsymbol{\lambda}^T \mathbf{H}\left(\mathbf{x}^* \right)$ es definida positiva en $T = \left\{ \mathbf{y} : \nabla \mathbf{h}^T \left(\mathbf{x}^* \right) \mathbf{y} = \mathbf{0} \right\}$, esto es, para $\mathbf{y} \in T$, $\mathbf{y} \neq \mathbf{0}$ se cumple que

$$\mathbf{y}^T \mathbf{L}\left(\mathbf{x}^* \right) \mathbf{y} \geq 0 \tag{5.48}$$

entonces, $\mathbf{x}^*$ es un mínimo local estricto de $f\left(\mathbf{x} \right)$ sujeta a $\mathbf{h}\left(\mathbf{x} \right) = \mathbf{0}$.

Ejemplo 5.6: Considérese el problema

$$\text{Maximizar} \quad f(\mathbf{x}) = x_1x_2 + x_2x_3 + x_1x_3$$
$$\text{Sujeta a:} \quad h(\mathbf{x}) = x_1 + x_2 + x_3 - 3 = 0$$

Solución:

En el ejemplo 5.5 anterior resultó que $\mathbf{x}^{*T} = (1,1,1)$ y $\lambda^* = -2$, satisfacen las condiciones de primer orden. En este caso, la matriz $\mathbf{L}=\mathbf{F}+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{H}$ se convierte en

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

que no es ni positiva ni negativa definida. Sin embargo, en el subespacio

$$T = \{\mathbf{y} : y_1 + y_2 + y_3 = 0\}$$

se observa que

$$\mathbf{y}^T\mathbf{L}\mathbf{y} = (y_1, y_2, y_3)\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = (y_1, y_2, y_3)\begin{pmatrix} y_2 + y_3 \\ y_1 + y_3 \\ y_1 + y_2 \end{pmatrix}$$
$$= y_1(y_2 + y_3) + y_2(y_1 + y_3) + y_3(y_1 + y_2) = -\left(y_1^2 + y_2^2 + y_3^2\right)$$

y así, $\mathbf{L}$ es definida negativa en T. Por tanto, la solución es un máximo local.

Restricciones de Desigualdad

Se tratarán ahora problemas de la forma

$$\begin{aligned} &Minimizar \quad f(\mathbf{x}) \\ &Sujeta\ a: \quad \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ &\qquad\qquad \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \end{aligned} \tag{5.49}$$

Se supone que $f(\mathbf{x})$ y $\mathbf{h}(\mathbf{x})$ son como antes y que $\mathbf{g}(\mathbf{x})$ es una función l-dimensional. Inicialmente, se supone que $f(\mathbf{x})$, $\mathbf{h}(\mathbf{x})$ y $\mathbf{g}(\mathbf{x})$ tienen primeras derivadas parciales continuas.

Condiciones Necesarias de Primer Orden

Con la consiguiente generalización de la definición anterior, se puede establecer un paralelismo con el desarrollo de las condiciones necesarias para las restricciones de igualdad.

Definición 5.3:

Sea $\mathbf{x}^*$ un punto que satisfaga las restricciones

$$\begin{aligned} \mathbf{h}(\mathbf{x}^*) &= \mathbf{0} \\ \mathbf{g}(\mathbf{x}^*) &\leq \mathbf{0} \end{aligned} \tag{5.50}$$

y sea K el conjunto de índices k para el cual $g_k(\mathbf{x}^*) = 0$. Entonces se dice que $\mathbf{x}^*$ es un punto regular de las restricciones (5.50) si los vectores gradientes $\nabla h_j(\mathbf{x}^*)$, $\nabla g_k(\mathbf{x}^*), 1 \leq j \leq m,\ k \in K$ son linealmente independientes.

Se observa que, según la definición de restricciones activas, un punto $\mathbf{x}^*$ es un punto regular si los gradientes de las restricciones activas son linealmente independientes. O de otra forma, $\mathbf{x}^*$ es regular para las restricciones si es regular en el sentido de la definición anterior para restricciones de igualdad aplicadas a las restricciones activas.

Teorema 5.9 Condiciones de Karush-Kuhn-Tucker

Sea $\mathbf{x}^*$ un punto mínimo relativo para el problema

$$\begin{aligned} &\textit{Minimizar} && f(\mathbf{x}) \\ &\textit{Sujeta a}: && \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ & && \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \end{aligned} \tag{5.51}$$

y supóngase que $\mathbf{x}^*$ es un punto regular para las restricciones. Entonces, existe un vector $\boldsymbol{\lambda} \in E^m$ y un vector $\boldsymbol{\mu} \in E^l$, $\boldsymbol{\mu} \geq \mathbf{0}$ tal que

$$\nabla f(\mathbf{x}^*) + \boldsymbol{\lambda}^T \nabla \mathbf{h}^T(\mathbf{x}^*) + \boldsymbol{\mu}^T \nabla \mathbf{g}^T(\mathbf{x}^*) = \mathbf{0} \tag{5.52}$$

$$\boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}^*) = 0 \tag{5.53}$$

Ejemplo 5.7: Considérese el problema

$$\begin{aligned} &\textit{Minimizar} && f(\mathbf{x}) = 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2 \\ &\textit{Sujeta a}: && g_1(\mathbf{x}) = x_1^2 + x_2^2 - 5 \leq 0 \\ & && g_2(\mathbf{x}) = 3x_1 + x_2 - 6 \leq 0 \end{aligned}$$

Solución:

Las condiciones necesarias de primer orden, además de las restricciones, son:

$$4x_1 + 2x_2 - 10 + 2\mu_1 x_1 + 3\mu_2 = 0$$

$$2x_1 + 2x_2 - 10 + 2\mu_1 x_2 + \mu_2 = 0$$

$$\mu_1\left(x_1^2 + x_2^2 - 5\right) = 0$$

$$\mu_2\left(3x_1 + x_2 - 6\right) = 0$$

$$\mu_1 \geq 0$$

$$\mu_2 \geq 0$$

Para hallar una solución, se definen varias combinaciones de restricciones activas y se verifican los signos de los multiplicadores de Lagrange resultantes. En este problema, se puede intentar hacer activas ninguna, una o dos restricciones. Al suponer que la primera restricción es activa, y la segunda inactiva, resultan las ecuaciones

$$4x_1 + 2x_2 - 10 + 2\mu_1 x_1 = 0$$

$$2x_1 + 2x_2 - 10 + 2\mu_1 x_2 = 0$$

$$x_1^2 + x_2^2 - 5 = 0$$

que tienen la solución $\mathbf{x}^{*T} = (1,2)$, $\mu_1^* = 1$. Esto produce $3x_1 + x_2 = 5$ y, por tanto, se satisface la segunda restricción. Así como $\mu_1^* = 1 > 0$, se concluye que esta solución satisface las condiciones necesarias de primer orden, ($\mu_2^* = 0$).

Condiciones de Segundo Orden

Las condiciones de segundo orden, necesarias y suficientes para problemas con restricciones de desigualdad, se deducen esencialmente teniendo en cuenta sólo el problema con restricciones de igualdad implicado por las restricciones activas. El plano tangente apropiado para estos problemas es el plano tangente a las restricciones activas.

Teorema 5.10 Condiciones Necesarias de Segundo Orden

Supóngase que las funciones $f(\mathbf{x}^*)$, $\mathbf{h}(\mathbf{x})$ y $\mathbf{g}(\mathbf{x})$ tiene segundas derivadas parciales continuas y que $\mathbf{x}^*$ es un punto regular de las restricciones (5.50). Si $\mathbf{x}^*$ es un punto mínimo relativo para el problema (5.49), entonces existe $\boldsymbol{\lambda} \in E^m$, $\boldsymbol{\mu} \in E^l$, $\boldsymbol{\mu} \geq \mathbf{0}$ tales que se cumplen (5.52) y (5.53) y tales que

$$\mathbf{L}(\mathbf{x}^*) = \mathbf{F}(\mathbf{x}^*) + \boldsymbol{\lambda}^T \mathbf{H}(\mathbf{x}^*) \tag{5.54}$$

es positiva semidefinida en el subespacio tangente de las restricciones activas en $\mathbf{x}^*$.

Al igual que en la teoría de la minimización sin restricciones, se puede formular una inversa del teorema de la condición necesaria de segundo orden, para obtener un teorema de la condición suficiente de segundo orden. Por analogía, con la situación sin restricciones, se puede esperar que la hipótesis requerida sea que $\mathbf{L}(\mathbf{x}^*)$ sea positiva definida en el plano tangente T. De hecho, esto es suficienteen la mayoría de las situaciones. Sin embargo, si hay *restricciones de desigualdad degeneradas* (esto es, restricciones de desigualdad activas que tengan cero como multiplicador de Lagrange asociado), se debe exigir que $\mathbf{L}(\mathbf{x}^*)$ sea positiva definida en un subespacio mayor que T.

Teorema 5.11 Condiciones de Suficiencia de Segundo Orden

Sean $f(\mathbf{x})$, $\mathbf{h}(\mathbf{x})$ y $\mathbf{g}(\mathbf{x}) \in C^2$. Las condiciones suficientes para que un punto $\mathbf{x}^*$ que satisfaga (5.50) sea un punto mínimo relativo estricto del problema (5.49) es que exista $\boldsymbol{\lambda} \in E^m$,$\boldsymbol{\mu} \in E^l$ tal que

$$\boldsymbol{\mu} \geq \mathbf{0} \tag{5.55}$$

$$\boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}^*) = 0 \tag{5.56}$$

$$\nabla f(\mathbf{x}^*) + \boldsymbol{\lambda}^T \nabla \mathbf{h}^T(\mathbf{x}^*) + \boldsymbol{\mu}^T \nabla \mathbf{g}^T(\mathbf{x}^*) = \mathbf{0} \tag{5.57}$$

y la matriz Hessiana

$$\mathbf{L}(\mathbf{x}^*) = \mathbf{F}(\mathbf{x}^*) + \boldsymbol{\lambda}^T \mathbf{H}(\mathbf{x}^*) + \boldsymbol{\mu}^T \mathbf{G}(\mathbf{x}^*) \tag{5.58}$$

sea positiva definida en el subespacio

$$T' = \left\{ \mathbf{y} : \nabla \mathbf{h}^T(\mathbf{x}^*)\mathbf{y} = \mathbf{0}, \nabla^T g_k(\mathbf{x}^*)\mathbf{y} = 0, \forall\, k \in K \right\}$$

donde $K = \left\{ k : g_k(\mathbf{x}^*) = 0, \mu_k > 0 \right\}$.

Se observa, sobre todo, que si todas las restricciones de desigualdad tienen multiplicadores de Lagrange correspondientes estrictamente positivos (sin desigualdades degeneradas), entonces el conjunto K incluye todas las desigualdades activas. En este caso, la condición suficiente es que el Lagrangiano sea positivo definido en T, el plano tangente de las restricciones activas.

Condiciones Suficientes para Problemas Convexo

Para problemas de programación convexa, las condiciones de primer orden de Karush-Kuhn-Tucker (KKT) también resultan ser suficientes. De esta manera, si podemos mostrar la convexidad de un problema, cualquier solución de las condiciones necesarias, satisfacerán automáticamente las condiciones suficientes. En adición, la solución será un mínimo global.

Teorema 5.12 Condición Suficiente para un Problema Convexo

Si $f\left(\mathbf{x}\right)$ es una función objetivo convexa definida sobre una región factible convexa (conjunto de restricciones), entonces las condiciones de primer orden KKT son necesarias así como suficientes para un mínimo global.

Ejemplo 5.8: Considérese el siguiente problema

$$\text{Minimizar} \quad f\left(\mathbf{x}\right) = x_1^3 + 4x_2^2 - 4x_1$$

$$\text{Sujeta a:} \quad g\left(\mathbf{x}\right) = 2x_2 - x_1 - 12 \geq 0$$

¿Este problema es un problema de programación convexa?.

Solución:

Las matrices Hessianas asociadas a la función objetivo y la restricción son:

$$\mathbf{F}\left(\mathbf{x}\right) = \begin{pmatrix} 6x_1 & 0 \\ 0 & 8 \end{pmatrix} \quad \text{y} \quad \mathbf{G}\left(\mathbf{x}\right) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

obsérvese que la Hessiana asociada a la restricción de desigualdad es cóncava y convexa, pero la Hessiana asociada a la función objetivo sólo es convexa para la región $x_1 \geq 0$ e indefinida en la región $x_1 < 0$.

Problemas

Resuelva el siguiente conjunto de problemas usando los métodos de reducción de variables y/o el método de multiplicadores de Lagrange.

5.1 Un cartel debe contener 300 cm^2 de texto impreso con márgenes superior e inferior de 6 cm y 4 cm en los laterales. Encuentre las dimensiones del cartel que minimizan el área total.

5.2 Una caja con base cuadrada y sin tapa debe retener 1000 cm^3. Determine las dimensiones que requieren el menor material para construir la caja.

5.3 Halle el volumen del cilindro circular recto más grande que puede inscribirse dentro de una esfera de radio R.

5.4 Se desea construir un recipiente cilíndrico de metal con tapa que tenga una superficie total de 100 cm^2. Encuentre las dimensiones de modo que tenga el mayor volumen posible.

5.5 Inscribir en una esfera de radio 1 m, un cilindro circular que tenga:

a) volumen máximo

b) área lateral máxima.

En ambos casos encuentre las dimensiones, radio de la base y altura.

5.6 Un alambre de 100 cm. de longitud, se corta en dos partes formando con una de ellas un círculo y con la otra un cuadrado. Cómo debe cortarse el alambre para que:

a) la suma de las áreas de las dos figura sea máxima y

b) la suma de las áreas de las dos figura sea mínima.

5.7

$$\text{Optimizar: } f(\mathbf{x}) = x_1 + 3x_2$$

$$\text{Sujeta a: } 2x_1 + 3x_2 \leq 6$$

$$-x_1 + 4x_2 \leq 4$$

$$x_1, x_2 \geq 0$$

5.8

$$\text{Optimizar: } f(\mathbf{x}) = \left(x_1 - \frac{9}{4}\right)^2 + (x_2 - 2)^2$$

$$\text{Sujeta a: } -x_1^2 + x_2 \geq 0$$

$$x_1 + x_2 \leq 6$$

$$x_1, x_2 \geq 0$$

Apéndice A

Matemáticas preliminares

A.1 Introducción

Para entender los métodos de la optimización matemática y su análisis moderno, es importante familiarizarse con el álgebra lineal, específicamente con las operaciones vectoriales y matriciales, y el cálculo básico. Así, los conceptos fundamentales del cálculo de las funciones, de una y varias variables, también debe incluirse dentro del estudio de estos temas. En este capítulo se define la *terminología* y su *notación estándar* usadas en todo el texto. Es en extremo importante entender esto debido a que sin ellos podría ser difícil seguir el resto del texto. La notación definida aquí no es complicada; de hecho, es muy simple y casi directa. Cualquier persona con el conocimiento básico en álgebra de matrices y álgebra lineal no encontrará dificultades. Se recomienda una revisión de conceptos y definiciones.

A.2 Terminología básica y notación

Escalares, vectores y matrices

Definición A.1 El símbolo δ_{ij} representa la delta de Kronecker y se define como

$$\delta_{ij} = \begin{cases} 0 & si \quad i \neq j \\ 1 & si \quad i = j \end{cases}; \quad i, j = 1, 2, ..., n. \tag{A.1}$$

Este símbolo se usa con frecuencia en combinación con la sumatoria para contraer términos y reducir las expresiones a formas más simples.

Definición A.2 Un *escalar* es aquella propiedad física o matemática que posee únicamente magnitud.

Las operaciones con escalares obedecen las mismas reglas del álgebra.

Definición A.3 Un vector $\mathbf{x}$ es un conjunto ordenado de n números reales $x_1, x_2, ..., x_n$, tal que n es un número natural cualquiera indicando su tamaño, la dimensión o su orden.

Estas entidades físicas o matemáticas tienen tanto magnitud como dirección y suelen representarse en letras minúsculas y negrillas. Los vectores pueden ser de dos tipos; el llamado *vector renglón*, cuya representación es

$$\mathbf{x}^T = \left(x_1, x_2, ..., x_n\right) \tag{A.2}$$

y el denominado vector columna, representado por

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \tag{A.3}$$

Cuando se transforma un vector columna a un vector renglón, lo que se hace es escribir la columna como renglón y poner el superíndice T en el símbolo $\mathbf{x}$; esta operación se conoce como transposición. Con frecuencia es conveniente expresar un vector en función de sus componentes según los ejes de coordenadas, una vez establecido el convenio para la designación de éstos últimos. Con este propósito se definen los vectores elementales correspondientes a los ejes de coordenadas.

Definición A.4 Un *vector elemental* en el espacio euclidiano E^n es un vector de magnitud unitaria con la i-ésima componente igual a la unidad y las restantes componentes nulas, y su dirección apunta a lo largo del correspondiente eje de coordenadas.

Así pues,

$$\mathbf{e}_i^T = \underbrace{(0,0,...,\overset{\text{i-ésima componente}}{\overset{\downarrow}{1}},...,0)}_{n\ componentes} \tag{A.4}$$

es una representación de un vector *renglón elemental* de n componentes; el único valor distinto de cero ocupa la i-ésima posición, como se indica en la ecuación (A.4).

Definición A.5 El *producto interno o escalar* de dos vectores elementales cualesquiera de orden n se define como

$$\mathbf{e}_i^T \mathbf{e}_j = \delta_{ij}\ ; \quad i,j = 1,...,n. \tag{A.5}$$

El resultado es un escalar con valor 0 o 1 obtenido de la definición (A.1) y expresa las *condiciones de ortonormalidad* de los vectores elementales entre sí. Obsérvese que ambos vectores elementales deben ser del mismo orden.

Definición A.6 Una *matriz* es un arreglo rectangular de cantidades que pueden ser números reales o complejos, símbolos, funciones de varias variables, etc. escritos de la siguiente forma:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}. \qquad \text{(A.6)}$$

Por lo general, las matrices se representan en letras mayúsculas y en negrillas. Se dice que $\mathbf{A}$ es una matriz de orden o dimensión $m\ x\ n$ si ésta tiene m renglones y n columnas. La matriz $\mathbf{A}$ también se puede representar en forma compacta por medio de sus elementos, de la siguiente manera:

$$\mathbf{A} = \left(a_{ij}\right)_{mxn}; \quad \begin{matrix} i = 1,...,m \\ j = 1,...,n \end{matrix} \qquad \text{(A.7)}$$

donde a_{ij} es el elemento de $\mathbf{A}$ ubicado en el renglón i y la columna j, $m\ x\ n$ es el orden de la matriz. En este contexto, algunas veces es conveniente pensar en un *escalar*, como una matriz de orden $1\ x\ 1$, en un *vector renglón* como una matriz de orden $1\ x\ n$, y de un *vector columna* como una matriz de orden $nx1$.

Definición A.7 La transpuesta de una matriz $\mathbf{A}$ se forma intercambiando los renglones de $\mathbf{A}$ por sus columnas, es decir

$$\mathbf{A}^T = \left(a_{ji}\right)_{nxm}; \quad \begin{matrix} i = 1,...,m \\ j = 1,...,n \end{matrix} \tag{A.8}$$

Algunas propiedades de la transposición

1) $\left(\mathbf{A}^T\right)^T = \mathbf{A}$

2) $\left(\mathbf{A}+\mathbf{B}\right)^T = \mathbf{A}^T + \mathbf{B}^T$

3) $\left(\mathbf{AB}\right)^T = \mathbf{B}^T\mathbf{A}^T$

Definición A.8 El *producto externo* entre dos vectores elementales de orden diferente (m y n respectivamente) se expresa como

$$\mathbf{e}_i\mathbf{e}_j^T = \left(\varepsilon_{kl}\right)_{mxn}; \quad \begin{matrix} k = 1,...,m \\ l = 1,...,n \end{matrix} \tag{A.9}$$

donde i y j están fijos y son arbitrarios; además,

$$\varepsilon_{kl} = \begin{cases} 0 & si\ la\ ubicación\ del\ elemento\ kl\ no\ coincide\ con\ ij \\ 1 & si\ la\ ubicación\ del\ elemento\ kl\ coincide\ con\ ij \end{cases}$$

El resultado es una matriz elemental de orden $m\,x\,n$, ya que $\mathbf{e}_i$ es un vector elemental columna de orden $m\,x\,1$ y $\mathbf{e}^T_j$ es un vector elemental renglón de orden $1xn$. Si los vectores elementales tienen el mismo orden, resulta una matriz cuadrada.

En términos de vectores elementales, cualquier matriz $\mathbf{A}$ puede escribirse como

$$\mathbf{A} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}\mathbf{e}_i\mathbf{e}_j^T \qquad \text{(A.10)}$$

Operaciones básicas entre matrices

Igualdad de matrices

$$\mathbf{A}_{mxn} = \mathbf{B}_{mxn} \Leftrightarrow\ a_{ij} = b_{ij}\ ; \quad \begin{matrix} i = 1,...,m \\ j = 1,...,n \end{matrix} \qquad \text{(A.11)}$$

Es aparente que la igualdad de matrices no tiene significado a menos que ellas sean del mismo orden.

Desigualdad de matrices

$$\mathbf{A}_{mxn} \leq \mathbf{B}_{mxn} \Leftrightarrow a_{ij} \leq b_{ij}\ ; \quad \begin{matrix} i = 1,...,m \\ j = 1,...,n \end{matrix} \qquad \text{(A.12)}$$

Suma de matrices

$$\mathbf{A}_{mxn} \pm \mathbf{B}_{mxn} = \mathbf{C}_{mxn} \Leftrightarrow c_{ij} = a_{ij} \pm b_{ij}\ ; \quad \begin{matrix} i = 1,...,m \\ j = 1,...,n \end{matrix} \quad \text{(A.13)}$$

Multiplicación de matrices

$$\mathbf{A}_{mxl}\mathbf{B}_{lxn} = \mathbf{C}_{mxn} \Leftrightarrow c_{ij} = \sum_{k=1}^{l} a_{ik}b_{kj}\ ; \quad \begin{matrix} i = 1,...,m \\ j = 1,...,n \end{matrix} \quad \text{(A.14)}$$

En general, el producto $\mathbf{AB}$ de dos matrices $\mathbf{A}$ y $\mathbf{B}$ se define como la matriz $\mathbf{C}$ tal que el elemento c_{ij} en el i-ésimo renglón y la j-ésima columna de $\mathbf{C}$ se obtiene sumando los productos de los elementos del i-ésimo renglón de $\mathbf{A}$ y los correspondientes elementos de la j-ésima columna de $\mathbf{B}$ tomando cada uno de ellos en orden. Observe que el número de columnas de $\mathbf{A}$ debe ser el mismo número de renglones de $\mathbf{B}$.

Ejemplo A.1 Calcule $\mathbf{AB}$ para las siguientes matrices:

$$\mathbf{A} = \begin{pmatrix} 2 & 3 \\ 1 & -4 \end{pmatrix} \text{ y } \mathbf{B} = \begin{pmatrix} 3 & -2 & 2 \\ 1 & 0 & -1 \end{pmatrix}$$

Solución:

$$\mathbf{AB} = \begin{pmatrix} 2 & 3 \\ 1 & -4 \end{pmatrix}\begin{pmatrix} 3 & -2 & 2 \\ 1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 9 & -4 & 1 \\ -1 & -2 & 6 \end{pmatrix}.$$

Definición A.9 El determinante de una matriz cuadrada $\mathbf{A}$ se puede expresar por recurrencia sobre n de la siguiente manera:

Para $n=1$

$$\det(a_{11}) = a_{11}$$

Para $n=2$

$$\det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

Para $n=3$

$$\det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}$$

En el caso general $n\,x\,n$ se define

$$\det \mathbf{A} = \left|\mathbf{A}\right| = \sum_{k=1}^{n} a_{1k}c_{1k}\ ; \quad c_{1k} = (-1)^{1+k} D_{1k}\ ; \quad n = 2,3,...$$

(A.15)

siendo D_{1k} el determinante de la matriz $(n-1)$ por $(n-1)$ resultante de suprimir en $\mathbf{A}$ la fila 1 y la columna k.

Definición A.10 El *menor* $\mathbf{M}_{ij}$ de la matriz cuadrada $\mathbf{A}$ es el arreglo matricial resultante después de eliminar el renglón i y la columna j de la matriz $\mathbf{A}$.

Definición A.11 El *cofactor* c_{ij} de la matriz cuadrada $\mathbf{A}$ es el escalar resultante de la siguiente expresión

$$c_{ij} = (-1)^{i+j} \det \mathbf{M}_{ij} \tag{A.16}$$

Determinante de una matriz por expansión de menores

Partiendo de las definiciones A.10 y A.11 se tienen dos maneras para calcular el valor del determinante de una matriz cuadrada $\mathbf{A}$.

Desarrollo por renglón fijando i arbitrariamente

$$|\mathbf{A}| = \sum_{j=1}^{n} a_{ij} c_{ij} = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det \mathbf{M}_{ij} \tag{A.17}$$

Usando la notación de la ecuación (A.15), $D_{ij} = \det \mathbf{M}_{ij}$.

Observación: véase entonces que la ecuación (A.17) es un caso particular de la ecuación (A.15).

Desarrollo por columna fijando j arbitrariamente

$$|\mathbf{A}| = \sum_{i=1}^{n} a_{ij} c_{ij} = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det \mathbf{M}_{ij} \qquad \text{(A.18)}$$

Si la matriz es triangular o diagonal las expansiones anteriores se reducen a

$$|\mathbf{A}| = \prod_{i=1}^{n} a_{ii} \qquad \text{(A.19)}$$

Definición A.12 La *matriz adjunta* $\mathbf{A}^*$ de $\mathbf{A}$ es la transpuesta de la matriz de cofactores $\mathbf{C}$ de $\mathbf{A}$, es decir

$$\mathbf{A}^* = adj\ \mathbf{A} = \mathbf{C}^T \Leftrightarrow a_{ij}^* = c_{ji}\ ; \quad i, j = 1, ..., n \qquad \text{(A.20)}$$

Definición A.13 La matriz inversa de $\mathbf{A}$ es la matriz única $\mathbf{A}^{-1}$ que satisface

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \qquad \text{(A.21)}$$

La matriz inversa existe si y sólo si $\mathbf{A}$ es no-singular, es decir, si $|\mathbf{A}| \neq 0$. Si $|\mathbf{A}| = 0$ la matriz es singular y su inversa no existe. Una derivación de la matriz inversa de $\mathbf{A}$ esta dada por

$$\mathbf{A}^{-1} = \frac{adj\ \mathbf{A}}{\det \mathbf{A}} = \frac{\mathbf{A}^*}{|\mathbf{A}|} \tag{A.22}$$

Ejemplo A.2 Obtenga la matriz inversa $\mathbf{A}^{-1}$ de $\mathbf{A}$ usando la ecuación (A.22), si

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{pmatrix}$$

Solución:

Si la matriz de cofactores $\mathbf{C}$ y la matriz adjunta $\mathbf{A}^*$ están dadas por

$$\mathbf{C} = \begin{pmatrix} 2 & 2 & -3 \\ 4 & -11 & 6 \\ -3 & 6 & -3 \end{pmatrix} \text{ y } \mathbf{A}^* = \mathbf{C}^T = \begin{pmatrix} 2 & 4 & -3 \\ 2 & -11 & 6 \\ -3 & 6 & -3 \end{pmatrix}$$

y el $\det \mathbf{A} = -3$, entonces la inversa de $\mathbf{A}$ es

$$\mathbf{A}^{-1} = \frac{\mathbf{A}^*}{|\mathbf{A}|} = \frac{1}{-3}\begin{pmatrix} 2 & 4 & -3 \\ 2 & -11 & 6 \\ -3 & 6 & -3 \end{pmatrix}$$

Se deja al lector comprobar que $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} = \mathbf{A}^{-1}\mathbf{A}$.

Algunas propiedades de la matriz inversa

1) $\left|\mathbf{A}^{-1}\right| = \dfrac{1}{|\mathbf{A}|}$

2) $\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$

3) $\left(\mathbf{A}^{T}\right)^{-1} = \left(\mathbf{A}^{-1}\right)^{T}$

4) Si $\mathbf{A}$ es simétrica ($\mathbf{A}^{T} = \mathbf{A}$), entonces $\left(\mathbf{A}^{-1}\right)^{T} = \mathbf{A}^{-1}$

5) Si $\mathbf{A}$ es diagonal invertible, entonces $\mathbf{A}^{-1} = \left(\dfrac{1}{a_{ii}}\right)_{nxn}$

6) $\left(\mathbf{AB}\right)^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

Productos y normas de vectores y matrices

Sea $\mathbf{x}$ un vector en el espacio *Euclidiano de* n dimensiones ($\mathbf{x} \in R^{n}$) y $\mathbf{A}$ una matriz simétrica ($\mathbf{A}^{\mathrm{T}}$=$\mathbf{A}$) de orden $n\ x\ n$; otros productos entre vectores y matrices comúnmente encontrados incluyen los siguientes:

1) Producto entre un vector renglón y una *matriz*:

$$\mathbf{x}^{T}\mathbf{A} = \sum_{i,j=1}^{n} x_i a_{ij} \mathbf{e}_j^{T} \qquad \text{(A.23)}$$

y el resultado es un *vector renglón de 1 x n*.

2) Producto entre una *matriz* y un *vector columna:*

$$\mathbf{Ax} = \sum_{i,j=1}^{n} a_{ij} x_j \mathbf{e}_i \tag{A.24}$$

y el resultado es un *vector columna de n x 1*.

3) Producto entre un *vector renglón*, una *matriz y un vector columna*:

$$\mathbf{x}^T \mathbf{Ax} = \sum_{i,j=1}^{n} x_i a_{ij} x_j \tag{A.25}$$

resulta ser un *escalar*.

4) Producto entre un *vector renglón y un vector columna*:

$$\mathbf{xx}^T = \sum_{i,j=1}^{n} x_i x_j \mathbf{e}_i \mathbf{e}_j^T \tag{A.26}$$

resulta ser una *matriz* de orden $n\,x\,n$.

Definición A.14 La *magnitud*, la *longitud* o la *norma* (euclidiana) de un vector $\mathbf{x}$ se expresa como

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{A.27}$$

Ejemplos de otras definiciones de normas vectoriales posibles son

1) $\|\mathbf{x}\|_\infty = \max_{1\le i\le n} |x_i|$ (norma máxima).

2) $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$ (norma de la sumatoria).

Estos casos y la norma euclidiana son casos particulares de normas de Hölder para $p=\infty$, $p=1$ y $p=2$.

$$\|\mathbf{x}\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p} \tag{A.28}$$

Definición A.15 La *magnitud*, la *longitud* o la *norma* de Frobenius de un matriz $\mathbf{A}$ se expresa como

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2} \tag{A.29}$$

Ejemplos de otras normas matriciales posibles son

1) $\|\mathbf{A}\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$

2) $\|\mathbf{A}\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|$

Dependencia e independencia lineal, rango de una matriz

Definición A.16

Una *combinación lineal* de vectores $\mathbf{x}_i$, $i = 1,...n$ con un conjunto de escalares α_i , $i = 1,...n$ se define como

$$\sum_{i=1}^{n} \alpha_i \mathbf{x}_i = \mathbf{y} = \mathbf{X}\boldsymbol{\alpha} \qquad \text{(A.30)}$$

Donde $\mathbf{x}_i$, $i = 1,...n$ son las columnas de $\mathbf{X}$. Entonces:

1. Las columnas de $\mathbf{X}$ son *linealmente dependientes* si

a) $\mathbf{X}\boldsymbol{\alpha} = \mathbf{0}$ puede satisfacerse para una $\boldsymbol{\alpha} \neq \mathbf{0}$ ó

b) $\mathbf{X}$ es singular, es decir, $\det \mathbf{X} = 0$ y $\mathbf{X}^{-1}$ no existe.

2. Las columnas de $\mathbf{X}$ son *linealmente independientes* si

a) $\mathbf{X}\boldsymbol{\alpha} = \mathbf{0}$ solo para una $\boldsymbol{\alpha} = \mathbf{0}$ ó

b) $\mathbf{X}$ es no singular, es decir, $\det \mathbf{X} \neq 0$ y $\mathbf{X}^{-1}$ existe.

Definición A.17 El *rango* de una matriz es igual al número de renglones o columnas linealmente independientes y se denota por r_{A} .

Obsérvese que el rango no depende de si se toman renglones o columnas linealmente independientes.

Valores y vectores propios de una matriz

Definición A.18 Asociada con cada matriz cuadrada $\mathbf{A}$ de orden $n\,x\,n$ existe una función

$$f(\lambda) = \det\left(\mathbf{A} - \lambda\mathbf{I}\right) \tag{A.31}$$

llamada la *función característica* de $\mathbf{A}$.

Definición A.19 La ecuación

$$f(\lambda) = \det\left(\mathbf{A} - \lambda\mathbf{I}\right) = 0 \tag{A.32}$$

puede expresarse en forma polinomial como

$$\sum_{k=0}^{n} c_k \lambda^{n-k} = 0 \tag{A.33}$$

y se llama la ecuación característica de la matriz $\mathbf{A}$.

Definición A.20 Las n raíces de la ecuación característica de una matriz cuadrada $\mathbf{A}$ se conocen como *valores propios* de $\mathbf{A}$ y también se conocen como *valores característicos* o incluso *eigenvalores*.

Ejemplo A.3

Dada la matriz ,

$$\mathbf{A}=\begin{pmatrix} 3 & 0 & 2 \\ -6 & 5 & 1 \\ 9 & -5 & 1 \end{pmatrix}$$

determine su ecuación característica y valores propios.

Solución:

Para hallar la ecuación característica de $\mathbf{A}$ debe calcularse el determinante de

$$\left|\begin{pmatrix} 3 & 0 & 2 \\ -6 & 5 & 1 \\ 9 & -5 & 1 \end{pmatrix}-\lambda\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right|=\lambda^3-9\lambda^2+10\lambda=0$$

Esta es la ecuación característica y para determinar los valores propios de $\mathbf{A}$, debe resolverse la ecuación obteniéndose todas sus raíces. Entonces

$$\lambda^3-9\lambda^2+10\lambda^2=\lambda\left(\lambda^2-9\lambda+10\right)=0$$

cuyas raíces son

$$\lambda_1=0,\qquad \lambda_2=\frac{9}{2}+\frac{\sqrt{41}}{2},\qquad \lambda_3=\frac{9}{2}-\frac{\sqrt{41}}{2}$$

las cuales son los valores propios de la matriz $\mathbf{A}$.

Definición A.21 Cualquier vector columna no nulo denotado por $\mathbf{x}_i$, de manera que

$$\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i \tag{A.34}$$

se llama *vector propio* de $\mathbf{A}$ asociado al valor propio λ_i.

Un importante e interesante teorema de la teoría de matrices es el *Teorema de Hamilton-Cayley*.

Teorema A.1 Toda matriz cuadrada $\mathbf{A}$ satisface su propia ecuación característica.

Si λ es reemplazada por la matriz $\mathbf{A}$ de orden n y el número real c_n es reemplazado por el múltiplo matricial $c_n\mathbf{I}$, donde $\mathbf{I}$ es la matriz identidad de orden n, entonces la ecuación característica de la matriz $\mathbf{A}$ resulta en una ecuación matricial válida, es decir,

$$\sum_{k=0}^{n} c_k \mathbf{A}^{n-k} = \mathbf{0} \tag{A.35}$$

El teorema de Hamilton-Cayley puede aplicarse al problema de determinar la matriz inversa de una matriz no singular $\mathbf{A}$. Sea

$$c_0\lambda^n + c_1\lambda^{n-1} + \cdots + c_{n-1}\lambda + c_n = 0$$

la ecuación característica de $\mathbf{A}$. Observe que como $\mathbf{A}$ es una matriz no singular, $\lambda_i \neq 0$, es decir, todo valor propio es no nulo y $c_n \neq 0$. Por el teorema de Hamilton-Cayley,

$$c_0\mathbf{A}^n + c_1\mathbf{A}^{n-1} + \cdots + c_{n-1}\mathbf{A} + c_n\boldsymbol{I} = \mathbf{0}$$

y entonces

$$\mathbf{I} = -\frac{1}{c_n}\left(c_0\mathbf{A}^n + c_1\mathbf{A}^{n-1} + \cdots + c_{n-1}\mathbf{A}\right) \tag{A.36}$$

si ambos lados de (A.36) son multiplicados por $\mathbf{A}^{-1}$, el resultado es

$$\mathbf{A}^{-1} = -\frac{1}{c_n}\left(c_0\mathbf{A}^{n-1} + c_1\mathbf{A}^{n-2} + \cdots + c_{n-1}\mathbf{I}\right) \tag{A.37}$$

.

Ejemplo A.4 Use el teorema de Hamilton-Cayley para hallar la matriz inversa de .

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ -1 & 1 & -3 \\ 2 & 2 & 4 \end{pmatrix}$$

Solución:

La ecuación característica de $\mathbf{A}$ es

$$\lambda^3 - 6\lambda^2 + 13\lambda - 6 = 0$$

.

Por el teorema de Hamilton-Cayley resulta

$$\mathbf{A}^3 - 6\mathbf{A}^2 + 13\mathbf{A} - 6\mathbf{I} = \mathbf{0},$$
$$\mathbf{I} = \frac{1}{6}\left(\mathbf{A}^3 - 6\mathbf{A}^2 + 13\mathbf{A}\right),$$
$$\mathbf{A}^{-1} = \frac{1}{6}\left(\mathbf{A}^2 - 6\mathbf{A} + 13\mathbf{I}\right)$$

y por lo tanto,

$$\mathbf{A}^{-1} = \frac{1}{6}\left[\begin{pmatrix} 1 & 0 & 1 \\ -1 & 1 & -3 \\ 2 & 2 & 4 \end{pmatrix}^2 - 6\begin{pmatrix} 1 & 0 & 1 \\ -1 & 1 & -3 \\ 2 & 2 & 4 \end{pmatrix} + 13\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right]$$

$$\mathbf{A}^{-1} = \begin{pmatrix} \frac{5}{3} & \frac{1}{3} & -\frac{1}{6} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ -\frac{2}{3} & -\frac{1}{3} & \frac{1}{6} \end{pmatrix}$$

Observe que el cálculo de una matriz inversa por el uso de la ecuación (A.37) es muy adaptable a una computadora digital y no es difícil de calcular manualmente para pequeños valores de n. Por último, usando la ecuación (A.37) ahora es posible expresar cualquier potencia entera negativa de una matriz no singular $\mathbf{A}$ de orden n en términos de una función lineal de las primeras (n-1) potencias de $\mathbf{A}$.

Clasificación de las matrices simétrica

Definición A.22 Una *matriz simétrica* $\mathbf{A}$ de orden $n\ x\ n$ puede dividirse en submatrices escaladas, como se muestra a continuación:

$$\mathbf{A} = \left(\begin{array}{ccccc} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{array} \right) \tag{A.38}$$

Los determinantes de las *submatrices principales* de $\mathbf{A}$

son $D_1 = a_{11},\ D_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix},\ D_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, \ldots,\ D_n = |\mathbf{A}|$

Todos los determinantes de las submatrices escaladas de $\mathbf{A}$ se conocen como los *menores principales* de $\mathbf{A}$.

Menores principales de una matriz $\mathbf{A}$ simétrica

Para $i = 1, 2, ..., n$

1) $\mathbf{A}$ es *positiva definida* si y sólo si $D_i > 0$.

2) $\mathbf{A}$ es *positiva semidefinida* si y sólo si $D_i \geq 0$.

3) $\mathbf{A}$ es *negativa definida* si y sólo si $(-1)^i D_i > 0$.

4) $\mathbf{A}$ es *negativa semidefinida* si y sólo si $(-1)^i D_i \geq 0$.

5) $\mathbf{A}$ es *indefinida* si y sólo si la sucesión D_i es distinta a lo anterior.

Ejemplo A.5 Dada la matriz

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$$

determine usando el criterio de determinantes de los menores principales, si la matriz es: positiva definida, positiva semidefinida, negativa definida, negativa semidefinida o indefinida.

Solución:

Los determinantes de los menores principales de la matriz $\mathbf{A}$ son

$$D_1 = a_{11} = 1 > 0 \quad \text{y} \quad D_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = -1 < 0$$

según la clasificación anterior, esta matriz es indefinida.

Valores propios de una matriz $\mathbf{A}$ simétrica

Para $i = 1, 2, ..., n$

1) $\mathbf{A}$ es *positiva definida* si y sólo si $\lambda_i > 0$.

2) $\mathbf{A}$ es *positiva semidefinida* si y sólo si $\lambda_i \geq 0$.

3) $\mathbf{A}$ es *negativa definida* si y sólo si $\lambda_i < 0$.

4) $\mathbf{A}$ es *negativa semidefinida* si y sólo si $\lambda_i \leq 0$.

5) $\mathbf{A}$ es *indefinida* si y sólo si la sucesión para algunas i, $\lambda_i \geq 0$ y las restantes $\lambda_i < 0$.

Ejemplo A.6 Dada la matriz del ejemplo A.5, determine si ésta es; positiva definida, positiva semidefinida, negativa definida, negativa semidefinida e indefinida usando el criterio de los valores propios.

Solución:

Los valores propios de la matriz $\mathbf{A}$ son

$$\lambda_1 = 2 + \sqrt{5} = 4.2360 > 0 \text{ y } \lambda_2 = 2 - \sqrt{5} = -0.2360 < 0$$

según la clasificación anterior, esta matriz es indefinida, resultado que concuerda en el ejemplo anterior.

Más propiedades de las matrices simétricas

1) Si $\lambda_i = 0$ para alguna i, entonces $\det \mathbf{A} = 0$.

2) Los valores propios de $\mathbf{A}$ y $\mathbf{A}^T$ son los mismos.

3) Los valores propios de una matriz diagonal son iguales a los elementos de la diagonal principal.

4) Si λ_i son los valores propios de $\mathbf{A}$, entonces λ_i^{-1} son los valores propios de $\mathbf{A}^{-1}$.

5) Si los valores propios de una matriz son distintos, entonces los vectores propios asociados son linealmente independientes.

6) Si $\mathbf{A}$ es una matriz real, entonces los valores propios de $\mathbf{A}$ son reales.

7) Si $\mathbf{A}$ es una matriz real, entonces los vectores propios de $\mathbf{A}$ asociados con distintos valores propios son vectores mutuamente ortogonales.

Formas cuadráticas

Definición A.23 Una forma cuadrática real es un polinomio homogéneo de segundo grado en n variables $x_1, x_2, \ldots, x_n$ es decir, es una función polinomial de la forma

$$f(\mathbf{A}, \mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \qquad \text{(A.39)}$$

Toda forma cuadrática puede expresarse como un producto matricial de la siguiente manera

$$f(\mathbf{A}, \mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} \qquad \text{(A.40)}$$

donde $\mathbf{A}$, si es una matriz simétrica real, es única; si no es simétrica, entonces $\mathbf{A}$ no es única.

Ejemplo A.7 Exprese la siguiente forma cuadrática en forma matricial

$$f(\mathbf{A},\mathbf{x}) = 3x_1^2 + 10x_1x_2 + 3x_2^2$$

Solución:

La transformación queda expresada por

$$f(\mathbf{A},\mathbf{x}) = 3x_1^2 + 10x_1x_2 + 3x_2^2 = \begin{pmatrix} x_1, x_2 \end{pmatrix}^T \begin{pmatrix} 3 & 5 \\ 5 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{x}^T\mathbf{A}\mathbf{x}$$

o bien, podría ser

$$f(\mathbf{A},\mathbf{x}) = 3x_1^2 + 10x_1x_2 + 3x_2^2 = \begin{pmatrix} x_1, x_2 \end{pmatrix}^T \begin{pmatrix} 3 & 10 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{x}^T\mathbf{A}\mathbf{x}$$

o

$$f(\mathbf{A},\mathbf{x}) = 3x_1^2 + 10x_1x_2 + 3x_2^2 = \begin{pmatrix} x_1, x_2 \end{pmatrix}^T \begin{pmatrix} 3 & 3 \\ 7 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{x}^T\mathbf{A}\mathbf{x}$$

Obsérvese que la primera forma matricial contiene a una matriz simétrica y esta es única, mientras que en las otras formas matriciales se tienen dos matrices no simétricas.

Clasificación de las formas cuadrática

1) Si $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ para todo valor de $\mathbf{x}$, la forma cuadrática se llama *positiva definida.*

2) Si $\mathbf{x}^T\mathbf{A}\mathbf{x} < 0$ para todo valor de $\mathbf{x}$, entonces se le llama *negativa definida.*

3) Si $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ la forma cuadrática se conoce con el nombre de *positiva semidefinida.*

4) Si $\mathbf{x}^T\mathbf{A}\mathbf{x} \leq 0$ se conoce como *negativa semidefinida.*

5) Si $\mathbf{x}^T\mathbf{A}\mathbf{x}$ toma valores negativos o cero para ciertos valores de $\mathbf{x}$, y positivos para otros valores de $\mathbf{x}$, entonces la forma cuadrática se llama *indefinida.*

Ejemplo A.8 La forma cuadrática del ejemplo A.7 sería una forma cuadrática indefinida porque ésta toma valores positivos para ciertos valores de $\mathbf{x}$ y valores negativos para otros valores de $\mathbf{x}$, como puede verificar el lector.

Referencias

Referencias

Bartholomew-Biggs, M., Nonlinear Optimization with Engineering Applications, Springer, 2008.
http://dx.doi.org/10.1007/978-0-387-78723-7

Bazaraa, M.S., Sherali, H.D. y C.M. Shetty, Nonlinear Programming Theory and Algorithms, Third Edition, John Wiley & Sons, Inc., 2006.
http://dx.doi.org/10.1002/0471787779

Bertsekas Dimitri P., Constrained Optimization and Lagrange Multiplier Methods, Athena Scientific Belmont, Massachusetts, 1996.

Beveridge, G.S.G. y R.S. Schetcher, Optimization: Theory and Practice, McGraw-Hill, New York, 1970.

Bevington, P.R., Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill Book Company, 1969.

Box, M.J. Computer J., 8:42, 1965.
http://dx.doi.org/10.1093/comjnl/8.1.42

Brown, K.M. y J.B. Dennis, Derivative free analogues of the Levenberg-Marquardt and Gauss algorithms for non linear least squares approximation, Numer. Math. 18, pp 289-297, 1972.
http://dx.doi.org/10.1007/BF01404679

Broyden, C.G., A class of methods for solving nonlinear simultaneous equations, Math. Comp., 19, pp 577-593, 1965.
http://dx.doi.org/10.1090/S0025-5718-1965-0198670-6

Burden, R.I. y J.D. Faires, Análisis Numérico, Grupo Editorial Iberoamericana, 1996.

Chapra, S.C. y R.P. Canale, Métodos Numéricos para Ingenieros, McGraw-Hill Interamericana Editores S.A. de C.V., 1999.

Constantinides, A. y N. Mostoufi Numerical Methods for Chemical Engineers with MATLAB Applications, Prentice-Hall International Series, 1999.

Conte, S.D., y C. de Boor, Análisis Numérico Elemental, McGraw-Hill, México, 1974.

Dennis, J.B. y R.B. Schnabel, Numerical Methods for Unconstrained and Nonlinear Equations, Prentice-Hall, Englewood Cliffs, New Jersey 1983.

Edgar, T.F., Himmelblau, D.M. y L.S. Lasdon, Optimization of Chemical Processes, Second Edition, McGraw-Hill International Edition, 2001.

Eves, H., Elementary Matrix Theory, Dover Publications, Inc., New York, 1966.

Fadeeva, N.V., Computational Methods of Linear Algebra, Dover Publications, Inc., New York, 1966.

Fletcher, R. y C.M. Reeves, Computer J., 7:149, 1964.
http://dx.doi.org/10.1093/comjnl/7.2.149

Fletcher, R., A modifie Marquardt subroutine for nonlinear least squares, Report R6799, AERE, Harwell, 1971.

Fletcher, R., Practical Methods of Optimization, Second Edition, John Wiley & Sons, Inc., 1993.

Forst, Wilhelm and Dieter Hoffmann, Optimization: Theory and Practice, Springer, 2010.
http://dx.doi.org/10.1007/978-0-387-78977-4

Forsythe, G. y C.B. Moler, Computer Solution of Linear Algebraic Equations, Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

Fox, R.L. Optimization Methods for Engineering Design, Addison-Wesley, Reading. Massachusetts 1971.

Gill, P.E., Murray, W., y Wright, M.H., Practical Optimization, Academic Press, New York, 1981.

Golub, G.H. y C.F. Van Loan, Matrix Computations, Third Edition, The John Hopkins University Press, Baltimore and London, 1996.

Hestenes, M.R., Conjugate-Direction Methods in Optimization, Springer Verlag, New York, 1980.
http://dx.doi.org/10.1007/978-1-4612-6048-6

Himmelblau, D.M., Applied Nonlinear Programming, McGraw-Hill Book Company, 1972.

Kuester, J.L. y J.H. Mize, Optimization Techniques with FORTRAN, McGraw-Hill Book Company, 1973.

Levenberg, K., A method for the solution of certain nonlinear problems in least squares, Quart, Appl. Math. 2, pp 164-168, 1944.

Luenberguer, D.G. y Y. Ye, Linear and Nonlinear Programming, 3rd Ed., Springer, 2008.

Marquardt, D.W., An algorithm for least squares estimation of nonlinear parameters, SIAM J. Appl. Math., 11, pp 431-441, 1963.
http://dx.doi.org/10.1137/0111030

Mathews, J.H. y K.D. Fink, Métodos Numéricos con MATLAB, Prentice-Hall Iberia S.R.L., 2000.

Nelder, J.A. y R. Mead, Computer J., 7:308, 1964.
http://dx.doi.org/10.1093/comjnl/7.4.308

Paviani, D. Ph.D. Dissertation, The University of Texas, Austin, Tex. 1969.

Pettofrezzo, A.J., Matrices and Transformations, Dover Publications, Inc., New York, 1966.

Reklaitis, G.V., Ravindran, R.A. y K.M. Ragsdell, Engineering Optimization Methods and Applications, Wiley-Interscience, New York, 1983.

Searle, S.R., Matrix Algebra Useful for Statistics, John Wiley & Sons. 1982.

Seinfeld, J.H. y L. Lapidus, Process Modeling, Estimation and Identification Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

Spendley, W., Hext, G.R. y F.R. Himsworth, Technometrics, 4:441, 1962.
http://dx.doi.org/10.1080/00401706.1962.10490033

Spiegel, M.R., Análisis Vectorial y una Introducción al Análisis Tensorial, McGraw-Hill Interamericana Editores S.A. de C.V., 1998.

Strang, G., Linear Algebra, Second Edition, Academic Press, New York, 1980.

Wilde, D.J., Optimum Seeking Methods, Prentice-Hall, Englewood Cliffs, New Jersey 1964.

Wilde, D.J. y C.S. Beightler, Foundations of Optimization, Prentice-Hall, Englewood Cliffs, New Jersey 1967.

El contenido del libro conjunta el material fundamental de un curso introductorio de optimización no lineal utilizado por los autores, en un período de más de veinte años. La motivación principal para escribir esta obra no ha sido la enseñanza de las matemáticas en sí (sin presentar extensas deducciones matemáticas), sino para equipar a los estudiantes en los fundamentos de la optimización matemática y sus algoritmos de manera integral: conceptos, desarrollo de habilidades para la implementación de programas de cómputo y metodología en la solución de problemas que se pueden abordar con los métodos de la optimización no lineal en particular.

La obra tiene una presentación gradual, de lo simple a lo complejo, y se inicia con la introducción de los conceptos elementales de optimización considerando la clasificación de los problemas, las propiedades de las funciones de una y varias variables, las condiciones necesarias y suficientes de optimalidad sin restricciones y de optimización cuadrática, y su interpretación geométrica. Para ello se presentan ejemplos de cada temática y problemas propuestos.

science