PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Análisis e Implementación del Método de Levenberg-Marquardt para la Estimación de Parámetros Mecánicos de un Altavoz

Tesis para obtener el grado académico de Magíster en Física Aplicada

que presenta:

Víctor Raúl Medina Chávez

Asesor:

Jorge Nestor Moreno Ruiz

Lima, 2017

Resumen

Si bien el modelo elástico tradicional de la parte mecánica de un altavoz de radiación directa, consistente en un sistema masa-resorte-amortiguador, es ampliamente utilizado en aplicaciones electroacústicas debido a su simplicidad y aceptable concordancia (para fines prácticos) con datos experimentales, existen características en el comportamiento del altavoz que dicho modelo no reproduce. Una de esas características, relacionada al fenómeno de creep que presenta la suspensión mecánica viscoelástica de los altavoces, es el incremento de los valores de la resistencia mecánica en el rango de las frecuencias bajas. El presente trabajo presenta un modelo viscoelástico de la parte mecánica de un altavoz de radiación directa y muestra como este resulta ser un modelo más completo respecto al modelo elástico tradicional al reproducir el mencionado comportamiento del altavoz asociado a la presencia del creep. Este resultado comparativo se valida mediante el ajuste del modelo a las curvas de datos experimentales obtenidos de la medición de distintos altavoces. El método de Levenberg–Marquardt, empleado para el ajuste del modelo por mínimos cuadrados no lineales, es estudiado con detalle.

Introducción

El presente trabajo de tesis se desarrolla en dos planos. Uno general que aborda el tratamiento matemático del criterio de los mínimos cuadrados y del método de Levenberg–Marquardt, y otro específico que muestra la aplicación de dicho método al caso particular de la estimación experimental de los parámetros de un modelo viscoelástico de la suspensión mecánica de un altavoz de radiación directa.

Por lo general, la aplicación de los mínimos cuadrados presentada a los estudiantes de ingeniería se limita al problema trivial de ajustar una recta a un conjunto de datos sujetos a errores de medición, lo que en última instancia se reduce a resolver un sistema de ecuaciones lineales. No es poco común, además, que los mínimos cuadrados se introduzcan como axioma, o que sus fundamentos matemáticos se discutan de forma superficial. En este contexto, raramente se tratan problemas reales que involucren modelos no lineales, y cuando esto ocurre la solución recae en el uso de software comercial. Este trabajo intenta cubrir esta brecha exponiendo algunos aspectos relevantes del tema y ofreciendo referencias a literatura complementaria.

La justificación matemática de los mínimos cuadrados como criterio para el ajuste óptimo de un modelo a un conjunto de datos experimentales se analiza en el capítulo 2. Los primeros intentos por darle un marco lógico a los mínimos cuadrados se basaron en el análisis probabilístico del error aleatorio de medición. Ejemplo de ello son las deducciones de Gauss y Laplace tratadas en este capítulo, las cuales se basan en la ley normal del error y el teorema del límite central respectivamente.

En la última sección de este capítulo se desarrolla una expresión general para la discrepancia entre el modelo y los datos medidos y se muestra que los minimos cuadrados son un caso particular. Este desarrollo se basa en ciertos requerimientos de naturaleza elemental y no considera las propiedades del

error de medición.

Una vez establecidos los mínimos cuadrados como criterio de ajuste el siguiente paso consiste en diseñar un procedimiento que implemente tal ajuste. Dos métodos precursores que cumplen con tal fin, el del máximo descenso y el de Gauss–Newton, son tratados con regular detalle en el capítulo 3. Sin embargo, no es poco frecuente que dichos métodos y sus posteriores variantes presenten problemas de convergencia cuando los modelos son complejos. Un método que hereda las ventajas de los dos anteriores y supera el problema de convergencia es el de Levenberg–Marquardt analizado en el capítulo 4. Por su eficiencia y confiabilidad, este método se ha convertido en la herramienta estándar para el ajuste de modelos no lineales.

Yendo al plano específico, los capítulos 1 y 5 muestran una aplicación del método de Levenberg–Marquardt al campo de la electroacústica. Este material puede ser de utilidad tanto para quien busque un complemento práctico de la teoría, como para el acústico interesado en los detalles específicos que involucra la medición de los elementos presentes en un modelo viscoelástico de la parte mecánica de un altavoz de radiación directa.

Dada la naturaleza de los materiales usados en la construcción de los altavoces, la resistencia mecánica asociada a las pérdidas de energía varía con la frecuencia. Este hecho, comprobado experimentalmente, sugiere la necesidad de ampliar el modelo tradicional de resistencia mecánica constante. El modelo viscoelástico propuesto en el capítulo 1 cumple este propósito y reproduce con mayor fidelidad el comportamiento observado experimentalmente.

El ejemplo de aplicación se completa en el capítulo 5 donde se muestran los resultados obtenidos al emplear el método de Levenberg–Marquardt en la estimación de los parámetros mecánicos de un altavoz de muestra según el modelo viscoelástico presentado en el primer capítulo.

En términos generales, la importancia del método de Levenberg–Marquardt radica en el proceso de ajuste de modelos, el cual permite al investigador teórico contrastar la exactitud de sus modelos contra los datos experimentales para así saber qué partes de estos se comprueban correctas y que otras necesitan ser refinadas. Al investigador práctico, por otro lado, le ofrece la posibilidad de extraer de forma eficiente información específica respecto a los parámetros o elementos de un sistema a partir de un conjunto de datos

medidos.

En cuanto al rango de aplicación, el método es potencialmente aplicable a cualquier sistema susceptible de ser modelado matemáticamente. Son diversas las áreas en las que el método es comúnmente empleado. La industría química, por ejemplo, extrae información respecto a las cantidades óptimas de los componentes que intervienen en una reacción química basándose en modelos ajustados experimentalmente. Otros ejemplos se dan en campos como las finanzas, en las que se ajustan los modelos de las series de tiempo correspondientes a los precios de las acciones. El método también se aplica en los modelos de riesgos de las empresas aseguradoras, en la optimización de técnicas agropecuarias, en sistemas de diagnóstico clínico de ciertas enfermedades crónicas, y en general en casi todos los ámbitos de las ciencias e ingeniería.

El método de Levenberg–Marquardt es sin duda uno de los implementos más importantes dentro de la caja de herramientas matemáticas con que cuenta la física aplicada.

Índice general

Resumen II

Introducción III

1. Parámetros Mecánicos de un Altavoz según el Modelo TGM **1**

1.1. Modelo Tradicional del Altavoz 2

1.2. Modelo TGM de la suspensión 5

1.3. Arreglo Experimental 8

2. Mínimos Cuadrados como Criterio para el Ajuste Óptimo de Modelos **10**

2.1. Formulación del Problema 10

2.2. Deducción basada en la Teoría de la Probabilidad 13

2.2.1. Modelo Lineal y Reducción de Modelos No Lineales . . 14

2.2.2. Enunciado de Gauss de la Media Aritmética 17

2.2.3. Demostración de Gauss de la Ley Normal del Error . . 18

2.2.4. Deducción de Gauss de los Mínimos Cuadrados 20

2.2.5. Deducción Alterna de Laplace 22

2.3. Interpretación Geométrico Vectorial del Método 25

2.3.1. Obtención de las Ecuaciones Normales Mediante Métodos Vectoriales 26

2.3.2. Evaluación de la Confiabilidad de los Parámetros Estimados 29

2.4. Mínimos Cuadrados y Modelos Aproximados 33

2.4.1. Expresión General para la Función de Discrepancia . . 34

2.4.2. Ejemplos Particulares de la Función de Discrepancia . 37

3. Algoritmos Precursores: Métodos del Máximo Descenso y de Gauss-Newton **39**
3.1. Introducción 40
3.2. Método del Máximo Descenso 42
3.2.1. Elección de la Dirección de Minimización 42
3.2.2. Reescalamiento de la Función a Minimizar 45
3.2.3. Cálculo de la Longitud del Paso 50
3.2.4. Demostración de la Convergencia del Método 52
3.3. Método de Gauss-Newton 61
3.3.1. Elección de la Dirección de Minimización 61
3.3.2. Cálculo de la Longitud del Paso 63
3.3.3. Demostración de la Convergencia del Método 64

4. Método de Levenberg-Marquardt **68**
4.1. Enfoque de Levenberg 69
4.1.1. Construcción del Método 69
4.1.2. Demostración de la Utilidad del Método 74
4.1.3. Cálculo del Vector de Paso 79
4.2. Enfoque de Marquardt 80
4.2.1. Base Teórica del Método 80
4.2.2. Escalamiento del Espacio de Parámetros 84
4.2.3. Construcción del Algoritmo 85
4.3. Diferencia entre los Dos Enfoques 88

5. Estimación de Parámetros y Conclusiones **90**
5.1. Procedimiento de Estimación de Parámetros 90
5.2. Conclusiones 93

Bibliografía **95**

Capítulo 1

Parámetros Mecánicos de un Altavoz según el Modelo TGM

El altavoz, visto como sistema electro-mécano-acústico formado por bloques que agrupan elementos de un mismo comportamiento y función, puede ser caracterizado por un modelo discreto de parámetros concentrados cuya complejidad dependerá de la exactitud que se busque obtener. La mayoría de estos modelos son válidos únicamente bajo dos condiciones de operación:

1. Régimen lineal del altavoz, obtenido a pequeños niveles de señal de entrada, con lo cual la presencia de las características no lineales tanto del factor de fuerza como de la compliancia mecánica son mínimas y despreciables.

2. Rango de frecuencias de pistón rígido, ubicado por debajo de las frecuencias donde empiezan a presentarse los modos del cono, lo que permite aproximar al sistema móvil del altavoz como un cuerpo rígido.

En el presente capítulo se introduce y analiza el modelo TGM (Truncated Generalized Maxwell) de la suspensión mecánica de un altavoz, el cual amplía el modelo tradicional incorporando el carácter dependiente de la frecuencia que presenta la parte real de la impedancia mecánica. Asímismo, se describe el procedimiento para estimar el valor de los parámetros que mejor ajusten el modelo a los datos experimentales de un altavoz en particular. Finalmente, se exponen detalles técnicos del arreglo de medición, y se comparan los resultados experimentales con aquellos simulados por el modelo.

1.1. Modelo Tradicional del Altavoz

Antes de tratar el modelo TGM mencionado, conviene hacer una breve descripción del modelo tradicional. Esto permite introducir al modelo TGM como una ampliación del modelo tradicional.

El modelo tradicional lineal del altavoz ha demostrado ser, a pesar de su simplicidad, una buena aproximación en la mayoría de los casos. El pequeño número de parámetros que contiene, la evidente interpretación física de los mismos, así como la disponibilidad de numerosos métodos para medirlos, tanto en el dominio del tiempo como de la frecuencia [28], [29], [18] y [24], hacen que el uso de este modelo sea común en aplicaciones prácticas.

En la figura 1.1 se muestra el esquema de un altavoz dinámico con el fin de facilitar la identificación de los elementos que conforman el modelo. El funcionamiento de este tipo de altavoces se puede simplificar de la siguiente manera: la corriente aplicada a través de los terminales de la bobina circula perpendicularmente a las líneas de campo magnético presentes en el entrehierro que completa el circuito magnético conformado por el polo, el imán, y las placas frontal y posterior. Las fuerzas de Lorentz resultantes actúan sobre las cargas que fluyen por la bobina. Esto genera el movimiento axial del diafragma formado por el cono y la tapa, ambos adheridos a la bobina y centrados por el sistema de suspensión compuesto por la araña y el anillo. Este movimiento se transmite a las partículas de aire circundantes y de ahí se propaga al resto del medio en forma de onda acústica. Cabe mencionar que el aire ejerce sobre el diafragma una fuerza de resistencia al movimiento.

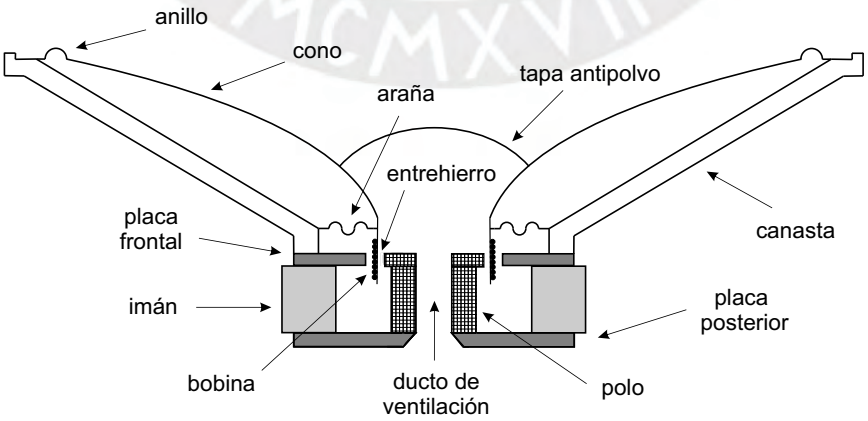


Figura 1.1: Esquema de la sección transversal de un altavoz dinámico.

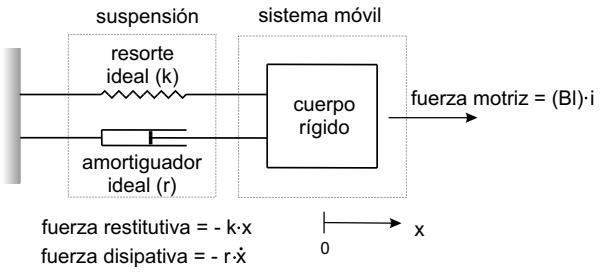


Figura 1.2: Modelo tradicional de la parte mecánica del altavoz.

Este efecto se toma en cuenta introduciendo en el modelo lo que se conoce como impedancia de radiación, que por simplificación muchas veces es descartada sin que esto afecte apreciablemente el desempeño del modelo, ya que su influencia es pequeña en comparación con los otros parámetros.

En el modelo tradicional [1] se asume que la fuerza motriz desarrollada en la bobina tiene un valor igual a la corriente que la atraviesa multiplicada por una constante denominada factor de fuerza (Bl), donde 'B' es la densidad de flujo magnético en el entrehierro y 'l' la longitud de la bobina. Asimismo, se asume que las fuerzas restitutivas y disipativas ejercidas por la suspensión sobre el diafragma son respectivamente proporcionales al desplazamiento y la velocidad del mismo. Para la parte eléctrica el modelo asume un comportamiento ideal de la bobina y de su resistencia intrínseca.

Bajo estos supuestos, la parte mecánica del modelo tradicional está conformada por un resorte más un amortiguador lineales y sin masa que caracterizan la suspensión, y por un cuerpo rígido que caracteriza al sistema móvil que incluye al diafragma, a la bobina y a la parte de la suspensión sujeta a movimiento. La conexión entre estos elementos se indica en la figura 1.2.

Considerando algunas leyes físicas básicas (ley de voltajes de Kirchhoff y segunda ley de Newton) es posible derivar la expresión matemática que

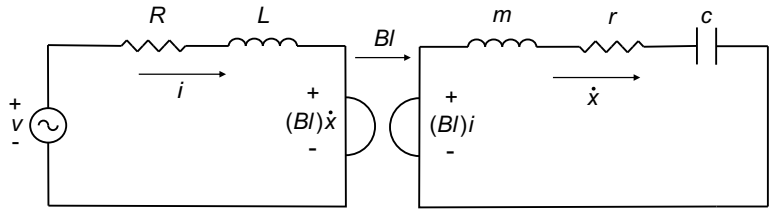


Figura 1.3: Circuito eléctrico equivalente.

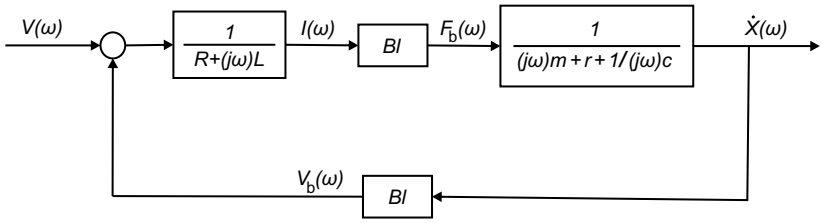


Figura 1.4: Diagrama de bloques del modelo tradicional.

describe, de acuerdo al modelo, los procesos eléctricos y mecánicos presentes en el altavoz. Esta expresión está compuesta por dos ecuaciones diferenciales, donde las funciones se pueden conocer experimentalmente y los coeficientes son las incógnitas a determinar:

$$v(t) = R\,i(t) + L\,\frac{\mathrm{d}i(t)}{\mathrm{d}t} + v_b(t) \qquad v_b(t) = Bl\,\frac{\mathrm{d}x(t)}{\mathrm{d}t}, \tag{1.1}$$

$$f_b(t) = m\,\frac{\mathrm{d}^2x(t)}{\mathrm{d}t^2} + r\,\frac{\mathrm{d}x(t)}{\mathrm{d}t} + \frac{1}{c}\,x(t) \qquad f_b(t) = Bl\,i(t). \tag{1.2}$$

La definición de los símbolos se puede ver en el cuadro 1.1.

Estas ecuaciones son idénticas a las que se obtienen del circuito eléctrico equivalente mostrado en la figura 1.3. En este circuito se utiliza un girador, el cual es el componente ideal que separa la parte eléctrica de la mecánica para así enfatizar la interdependencia entre ambas. Dicho en términos generales, un girador es un elemento discreto eléctrico ideal de dos puertos cuya ley de funcionamiento es la siguiente: el voltaje presente en los terminales del primer puerto es proporcional a la corriente que fluye por el segundo puerto y, similarmente, el voltaje presente en los terminales del segundo puerto es proporcional a la corriente que fluye por el primer puerto. Este comportamiento puede observarse en la figura 1.3.

Vistas en el dominio de la frecuencia, dichas ecuaciones se pueden representar mediante el diagrama de bloques de la figura 1.4, donde las mayúsculas denotan la transformada de Fourier; e.g., $X(\omega) = \mathscr{F}\{x(t)\}$.

La ventaja del diagrama de bloques es que muestra en forma sencilla la relación entre las señales. Esto permite que tanto las funciones de transferencia como las impedancias puedan ser derivadas directamente. Así por ejemplo

$v(t)$	Voltaje aplicado [V]
$i(t)$	Corriente en la bobina [A]
R	Resistencia de la bobina [Ω]
L	Inductancia de la bobina [H]
Bl	Factor de fuerza [N/A]
$v_b(t)$	Voltaje inducido en la bobina [V]
$f_b(t)$	Fuerza motriz desarrollada en la bobina [N]
$x(t)$	Desplazamiento de la bobina [m]
m	Masa del sistema móvil más la carga de aire [kg]
r	Coeficiente de amortiguamiento viscoso de la suspensión [N s/m]
c	Compliancia de la suspensión [m/N]
k	Coeficiente de elasticidad de la suspensión (= 1/c) [N/m]
t	Tiempo [s]
ω	Frecuencia angular [rad/s]

Cuadro 1.1: Nomenclatura (modelo tradicional).

se obtiene la función de transferencia del altavoz

$$\frac{I(\omega)}{V(\omega)} = \frac{1}{R + (j\omega)L + \dfrac{(Bl)^2}{(j\omega)m + r + 1/(j\omega)c}}, \tag{1.3}$$

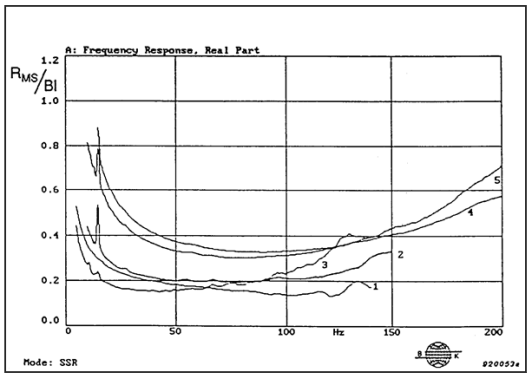
y la impedancia mecánica

$$\frac{F_b(\omega)}{\dot{X}(\omega)} = Z_M(\omega) = (j\omega)m + r + 1/(j\omega)c. \tag{1.4}$$

1.2. Modelo TGM de la suspensión

En la sección anterior se vio cómo el modelo tradicional posee un amortiguador viscoso como único elemento disipador de energía (elemento que ejerce una fuerza que se opone al movimiento). En consecuencia, la expresión de la resistencia mecánica, definida como la parte real de (1.4)) y por ende responsable de las pérdidas de energía en el sistema, es una constante r cuyo valor es el coeficiente de amortiguación viscoso.

Sin embargo, esta aproximación básica adoptada por el modelo tradicional respecto a la resistencia mecánica no reproduce satisfactoriamente los resultados obtenidos experimentalmente, como se menciona en [13], donde se



Muestra	Modelo del Altavoz	Diámetro
1	Peerless 831727	10 in
2	JBL 2241H	18 in
3	JBL 2245H	18 in
4	JBL 2206H	12 in
5	JBL 2226H	15 in

Figura 1.5: Resistencia Mecánica (dividida entre Bl) de los altavoces en el cuadro contiguo.

midió la resistencia mecánica dividida entre el factor de fuerza de distintos altavoces. El resultado de esas mediciones se muestra en la figura 1.5. En ella se observa que en realidad la resistencia mecánica es dependiente de la frecuencia y se aleja notoriamente del valor constante asumido.

Esta limitación motiva la elaboración de un modelo más complejo que incluya propiedades físicas adicionales a la de restitución y amortiguamiento que presenta la suspensión. Una de tales propiedades es la que se conoce como creep, comúnmente presente en materiales viscoelásticos como los que

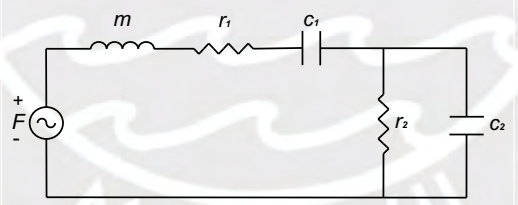


Figura 1.6: Circuito eléctrico equivalente.

F	Fuerza motriz [V]
m	Masa del sistema móvil más la carga de aire [kg]
r_1	Resistencia mecánica de la rama Voigt del modelo TGM [N·s/m]
c_1	Compliancia mecánica de la rama Voigt del modelo TGM [m/N]
r_2	Resistencia mecánica de la rama Maxwell del modelo TGM [N·s/m]
c_2	Compliancia mecánica de la rama Maxwell del modelo TGM [m/N]

Cuadro 1.2: Nomenclatura (modelo TGM).

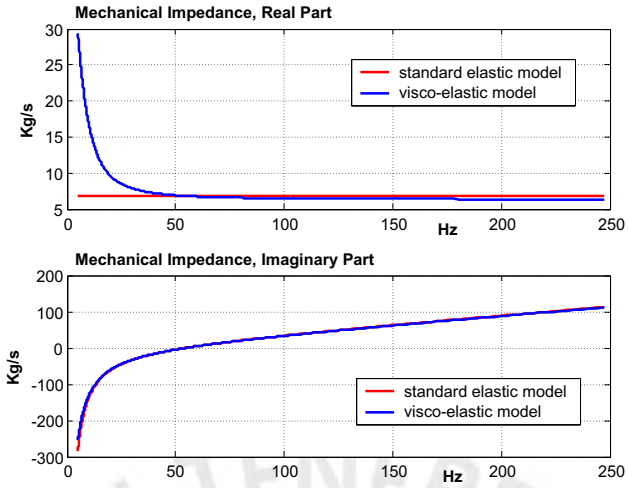


Figura 1.7: Comparación de los resultados obtenidos para la impedancia mecánica calculada por el método tradicional y el método TGM.

componen la suspensión. El fenómeno de creep se manifiesta como un continuo y lento desplazamiento del diafragma bajo una fuerza constante aplicada, hecho que impide alcanzar el estado estacionario esperado. La presencia de creep en los altavoces fue reportada por primera vez en [6] y modelada según ciertas estructuras en [14].

Un modelo particular, entre los diversos modelos existentes para caracterizar materiales que exhiben creep, que ha probado ser apropiado para modelar la suspensión (resistencia y compliancia mecánica) es el modelo generalizado de Maxwell truncado, o TGM por sus siglas en inglés. Reemplazando por tanto el modelo tradicional de la suspensión por el modelo TGM, se consigue un modelo mejorado de la impedancia mecánica del altavoz. Para este caso, el circuito eléctrico equivalente de la parte mecánica, ya sin incluir el girador, se muestra en la figura 1.6.

En la expresión correspondiente a la impedancia mecánica del modelo TGM

$$Z_M = r_1 + \frac{r_2}{1+(\omega c_2 r_2)^2} + j\left[(\omega m - \frac{1}{\omega c_1}) - \frac{\omega c_2 r_2^2}{1+(\omega c_2 r_2)^2}\right], \qquad (1.5)$$

puede observarse que la parte real ya no es constante sino que depende de la frecuencia. Nótese en la figura 1.7 que el efecto de los parámetros r_2 y c_2, introducidos por el modelo TGM (a diferencia del modelo tradicional), se manifiesta en el rango de las frecuencias bajas de la parte real de la impe-

dancia mecánica, reproduciendo el comportamiento asociado al creep que se desea modelar. El efecto de r_2 y c_2 es en cambio irrelevante en todo el rango de frecuencias en la parte imaginaria de la impedancia mecánica, la cual es dominada por los parámetros del modelo tradicional.

Para ilustrar las diferencias entre el modelo tradicional y el TGM, en la figura 1.7 se compara tanto la parte real como la imaginaria de las impedancias mecánicas simuladas según cada modelo. La impedancia medida del altavoz no se muestra en la figura.

De las gráficas puede observarse que a diferencia del modelo tradicional el modelo TGM presenta un incremento de la resistencia mecánica en el rango de las frecuencias bajas similar al que presentan los datos experimentales (figura 1.5). Esta característica, como se verá más adelante, permite una mejor coincidencia entre los datos simulados con el modelo TGM y los datos obtenidos experimentalmente.

1.3. Arreglo Experimental

En la figura 1.8 se muestra el arreglo experimental empleado en este trabajo para la toma de datos.

El altavoz se sujeta verticalmente mediante una abrazadera de bronce que evita fugas de flujo magnético. Como señal de entrada se utiliza un barrido sinusoidal generado por el analizador y transmitido al altavoz a través de un amplificador de potencia. El nivel de esta señal debe ser suficientemente pequeño como para no excitar las no linealidades del altavoz y suficientemente grande como para alcanzar una buena relación señal a ruido. Tres señales, conectadas a las entradas del analizador multicanal, son medidas simultáneamente: el voltaje en los terminales del altavoz, el voltaje a través de una resistencia de 1Ω en serie con el altavoz y por ende la corriente que circula por la bobina, y la velocidad del cono mediante un transductor láser de velocidad.

Las componentes de frecuencia de estas tres señales son medidas por el analizador mediante la técnica SSR (Steady State Response) que determina la amplitud y fase de cada señal medida tomando como referencia la señal armónica de entrada y descartando los primeros periodos para conseguir que los transitorios se extingan.

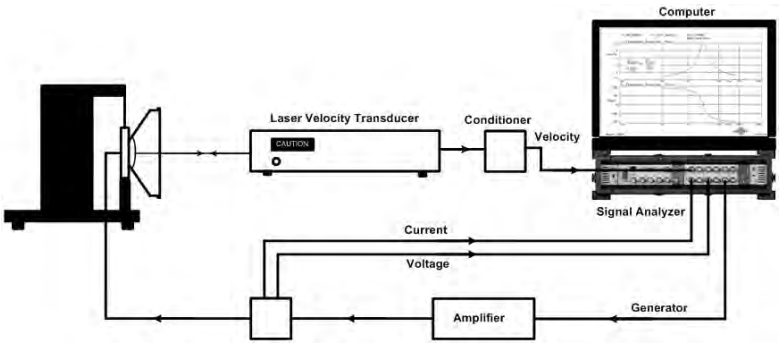


Figura 1.8: Arreglo experimental

Las impedancias eléctrica y mecánica del altavoz se calculan a partir de las componentes de frecuencia ya medidas de la velocidad, corriente, y voltaje. Asimismo, es posible calcular el factor de fuerza siguiendo el procedimiento descrito en [24].

La definición teórica de la impedancia mecánica del altavoz es la razón entre la fuerza ejercida en la bobina y la velocidad del cono. Sin embargo, en este caso la fuerza no se mide directamente sino que se calcula asumiendo idealmente que la misma es igual a la corriente multiplicada por el factor de fuerza. La validez de este supuesto ya se ha verificado en anteriores investigaciones. Para ello se mide la impedancia mecánica sin hacer uso del motor del altavoz, i.e., aplicando y midiendo directamente una fuerza externa al sistema móvil del altavoz. En [25] se dan detalles de un arreglo experimental usado con ese propósito.

Capítulo 2

Mínimos Cuadrados como Criterio para el Ajuste Óptimo de Modelos

En las diversas ramas de la ciencia se presenta con frecuencia el problema consistente en ajustar un modelo matemático parametrado a datos observados experimentalmente de cierto sistema. Para llevar a cabo el ajuste es necesario establecer un criterio que evalúe el grado de concordancia, o discrepancia, entre el modelo y los datos experimentales.

El presente capítulo analiza uno de esos criterios en particular, denominado método de los mínimos cuadrados, y justifica el uso del mismo para conseguir el ajuste óptimo de modelos.

En la primera sección se introduce una descripción del método; en la segunda, se dan detalles de la derivación del método hecha por Gauss [9] con base en la teoría matemática de la probabilidad; en la tercera, se presenta una interpretación del método propuesta por Kolmogorov [15] dentro del campo del álgebra lineal vectorial; y finalmente en la cuarta, se propone un enfoque distinto del problema, apartándolo de su interpretación probabilística convencional y asociándolo con algunos conceptos de espacios normados.

2.1. Formulación del Problema

En términos generales se puede definir un sistema como una caja que admite un conjunto de señales de entrada, las procesa, y produce como res-

puesta un grupo de señales de salida. Cuando este proceso está libre de elementos aleatorios, produce siempre las mismas salidas en respuesta a unas mismas entradas y se dice que el sistema es determinista. En este caso, las entradas y salidas del sistema están relacionadas de manera precisa mediante una expresión matemática que representa el modelo exacto del sistema.

Contar con una expresión matemática que describa el comportamiento de cierto sistema resulta muy útil en el proceso de diseño, análisis y optimización del mismo. Esta expresión o modelo matemático, al ser una forma equivalente del sistema, permite mediante operaciones simples simular la respuesta del sistema ante situaciones hipotéticas.

Sin embargo, conocer el modelo exacto de un sistema es solo un caso ideal, pues no es posible conocer la totalidad de elementos que intervienen en un sistema real. En la práctica se requiere un modelo que aproxime al sistema con un grado de exactitud suficiente. El proceso de construcción de un modelo implica identificar los elementos representativos que conforman un sistema.

Un modelo es ajustable cuando posee parámetros que pueden asumir valores arbitrarios que permitan adaptar el modelo a distintos sistemas particulares. Así, en términos más precisos, el proceso de ajuste consiste en estimar los valores particulares de los parámetros que produzcan un mínimo grado de discrepancia entre las respuestas observadas del sistema y las reproducidas por el modelo.

La figura 2.1 ilustra el proceso de generación de las respuestas reales (y^*), medidas (y) y predichas ($\hat{y}$). Dado que todo proceso de medición introduce inevitablemente una componente de error aleatorio (Δ), las respuestas medidas poseen la forma $y = y^* + \Delta$.

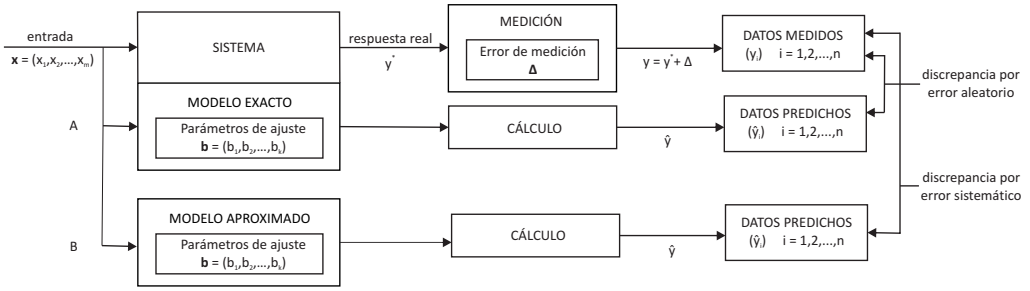


Figura 2.1: Proceso de generación de datos medidos (y) y predichos ($\hat{y}$).

Dos casos se muestran en esta figura:

Caso A: se tiene el modelo exacto del sistema, por lo que es posible conseguir que las respuestas predichas coincidan con las reales al asignarles valores apropiados a los parámetros de ajuste. En este caso, la discrepancia entre las respuestas medidas y predichas se debe únicamente al error aleatorio de medición y, por ende, el criterio para evaluarla debe fundarse en la naturaleza aleatoria del error.

Caso B: se tiene un modelo aproximado del sistema, por lo que aun cuando se estimen los valores óptimos de los parámetros de ajuste, habrá siempre una diferencia entre las respuestas reales y predichas. En este caso, la discrepancia se debe al error sistemático introducido por el uso de un modelo aproximado y, por ende, el criterio para evaluarla debe considerar la naturaleza sistemática del error.

Para expresar matemáticamente lo anteriormente dicho, se considera el modelo

$$\begin{aligned} \hat{y}_i &= f(x_{1i}, x_{2i}, \ldots, x_{mi}; b_1, b_2, \ldots, b_k) \\ &= f(\boldsymbol{x}_i, \boldsymbol{b}) \\ &= f_i(\boldsymbol{b}), \end{aligned} \tag{2.1}$$

donde $\boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{mi})$ es el i-ésimo vector de entradas o variables independientes, $\boldsymbol{b} = (b_1, b_2, \ldots, b_k)$ es el vector de parámetros de ajuste, e $\hat{y}_i$ es la i-ésima respuesta escalar predicha por el modelo. Asimismo, se define

el residuo (r_i) para cada uno de los n datos medidos

$$y_i = y_i^* + \Delta_i, \qquad i = 1, 2, \ldots, n,$$

como la diferencia entre las correspondientes respuestas medida y predicha

$$r_i = y_i - \hat{y}_i, \qquad i = 1, 2, \ldots, n; \tag{2.2}$$

y se asume que todo criterio de ajuste legítimo define la discrepancia entre los datos medidos y predichos en función de tales residuos.

En particular, el método de los mínimos cuadrados le asigna a la discrepancia Φ un valor igual a la sumatoria de los cuadrados de los residuos

$$\Phi = \sum_{i=1}^{n} r_i^2, \tag{2.3}$$

de modo que el vector de parámetros de ajuste $\boldsymbol{b}_{min}$ que minimice Φ debe ser solución del sistema de ecuaciones

$$\frac{\partial \Phi}{\partial b_j} = 2 \sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial b_j} = 0, \qquad j = 1, 2, \ldots, k. \tag{2.4}$$

En este punto es importante anotar dos hechos. Primero, si algún $\boldsymbol{b_0}$ es solución del sistema de ecuaciones

$$r_i(b_1, b_2, \ldots, b_k) = 0, \qquad i = 1, 2, \ldots, n, \tag{2.5}$$

dicho $\boldsymbol{b_0}$ es también solución de (2.4). Segundo, para el caso en que los r_i son una combinación lineal de las componentes de $\boldsymbol{b}$, se tiene que (2.5) posee infinitas soluciones si $n < k$. En dicho caso, $n \geq k$ es condición necesaria para que (2.4) posea una solución única $\boldsymbol{b}_{min}$.

2.2. Deducción de Gauss basada en la Teoría de la Probabilidad

La primera mención del método de los mínimos cuadrados aparece en una publicación de Legendre de 1806 [19], en la que el método es propuesto sin

fundamento matemático y a manera de principio. Tres años después, Gauss publica la primera deducción del método [9], fundada en principios de naturaleza más elemental, en la que conecta el método con la teoría matemática de la probabilidad.

En esta sección se analizan los detalles de la deducción de Gauss y se expone al final una deducción alterna sugerida por Laplace.

2.2.1. Modelo Lineal y Reducción de Modelos No Lineales a la Forma Lineal

Como punto de partida se presenta un ejemplo de aplicación del método. Se considera para ello el caso más sencillo en el que el modelo, expresado en (2.1), es lineal en los parámetros que se busca estimar. Se tiene entonces

$$f_i(\boldsymbol{b}) = c_{1i}b_1 + c_{2i}b_2 + \cdots + c_{ki}b_k, \tag{2.6}$$

de donde se desprende la forma de los residuos

$$r_i = y_i - \sum_{j=1}^{k} c_{ji}b_j. \tag{2.7}$$

El vector de parámetros $\boldsymbol{b}_{min}$ es en este caso la solución única del sistema de ecuaciones conocido como ecuaciones normales

$$\left.\begin{array}{ccccccccc} [\boldsymbol{c_1c_1}]b_1 & + & [\boldsymbol{c_1c_2}]b_2 & + & \dots & + & [\boldsymbol{c_1c_k}]b_k & = & [\boldsymbol{c_1y}] \\ [\boldsymbol{c_1c_2}]b_1 & + & [\boldsymbol{c_2c_2}]b_2 & + & \dots & + & [\boldsymbol{c_2c_k}]b_k & = & [\boldsymbol{c_2y}] \\ \vdots & & \vdots & & \ddots & & \vdots & & \vdots \\ [\boldsymbol{c_1c_k}]b_1 & + & [\boldsymbol{c_2c_k}]b_2 & + & \dots & + & [\boldsymbol{c_kc_k}]b_k & = & [\boldsymbol{c_ky}] \end{array}\right\} \tag{2.8}$$

obtenido al introducir (2.7) en (2.4). Cabe señalar que los corchetes [ ] en la notación de Gauss representan el producto interno de los vectores contenidos en ellos, i.e.,

$$[\boldsymbol{c_jc_h}] = \sum_{i=1}^{n} c_{ji}c_{hi}, \qquad j, h = 1, 2, \dots, k.$$

En la sección 2.3.1 se analizan las condiciones que deben cumplir los vectores $\boldsymbol{c_1}$, $\boldsymbol{c_2}, \dots,$ $\boldsymbol{c_k}$ para garantizar que (2.8) posea una única solución.

Aplicando técnicas de álgebra elemental es posible obtener la solución

de (2.8). Una de estas técnicas, basada en el uso de determinantes, permite expresar de forma explícita cada componente $(b_{min})_j$ del vector solución en función de los coeficientes presentes en el sistema de ecuaciones.

En particular para obtener $(b_{min})_1$ se define

$$A = \begin{bmatrix} [\boldsymbol{c_1 c_1}] & [\boldsymbol{c_1 c_2}] & \dots & [\boldsymbol{c_1 c_k}] \\ [\boldsymbol{c_1 c_2}] & [\boldsymbol{c_2 c_2}] & \dots & [\boldsymbol{c_2 c_k}] \\ \vdots & \vdots & \ddots & \vdots \\ [\boldsymbol{c_1 c_k}] & [\boldsymbol{c_2 c_k}] & \dots & [\boldsymbol{c_k c_k}] \end{bmatrix} \quad \text{y} \quad \boldsymbol{g} = \begin{bmatrix} [\boldsymbol{c_1 y}] \\ [\boldsymbol{c_2 y}] \\ \vdots \\ [\boldsymbol{c_k y}] \end{bmatrix}, \tag{2.9}$$

con lo cual (2.8) toma la forma vectorial

$$A\,\boldsymbol{b} = \boldsymbol{g}. \tag{2.8a}$$

Si $\boldsymbol{b} = \boldsymbol{b}_{min}$ es solución de esta ecuación, entonces también lo será de la ecuación

$$A\,\boldsymbol{b} = \frac{b_1}{(b_{min})_1}\boldsymbol{g}, \tag{2.10}$$

puesto que para $\boldsymbol{b}=\boldsymbol{b}_{min}$ ambas ecuaciones son idénticas al ser $b_1/(b_{min})_1=1$. Definiendo

$$A_1 = \begin{bmatrix} [\boldsymbol{c_1 c_1}] - \dfrac{[\boldsymbol{c_1 y}]}{(b_{min})_1} & [\boldsymbol{c_1 c_2}] & \dots & [\boldsymbol{c_1 c_k}] \\ [\boldsymbol{c_1 c_2}] - \dfrac{[\boldsymbol{c_2 y}]}{(b_{min})_1} & [\boldsymbol{c_2 c_2}] & \dots & [\boldsymbol{c_2 c_k}] \\ \vdots & \vdots & \ddots & \vdots \\ [\boldsymbol{c_1 c_k}] - \dfrac{[\boldsymbol{c_k y}]}{(b_{min})_1} & [\boldsymbol{c_2 c_k}] & \dots & [\boldsymbol{c_k c_k}] \end{bmatrix},$$

se expresa (2.10) en forma matricial

$$A_1\boldsymbol{b} = 0, \tag{2.10a}$$

la cual hace evidente que aparte de la solución trivial $\boldsymbol{b}=0$ existen infinitas soluciones de la forma

$$\boldsymbol{b} = \alpha\,\boldsymbol{b}_{min}, \qquad \alpha \in \mathbf{R},$$

por ende A_1 es singular y tanto su determinante como la determinante de su

transpuesta A_1^T son cero. Se tiene entonces la ecuación

$$|A_1^T| = \begin{vmatrix} [\boldsymbol{c_1 c_1}] - \frac{[\boldsymbol{c_1 y}]}{(b_{min})_1} & [\boldsymbol{c_1 c_2}] - \frac{[\boldsymbol{c_2 y}]}{(b_{min})_1} & \dots & [\boldsymbol{c_k c_k}] - \frac{[\boldsymbol{c_k y}]}{(b_{min})_1} \\ [\boldsymbol{c_1 c_2}] & [\boldsymbol{c_2 c_2}] & \dots & [\boldsymbol{c_2 c_k}] \\ \vdots & \vdots & \ddots & \vdots \\ [\boldsymbol{c_1 c_k}] & [\boldsymbol{c_2 c_k}] & \dots & [\boldsymbol{c_k c_k}] \end{vmatrix} = 0, \tag{2.11}$$

de donde se despeja el valor de $(b_{min})_1$

$$(b_{min})_1 = \frac{\begin{vmatrix} [\boldsymbol{c_1 y}] & [\boldsymbol{c_2 y}] & \dots & [\boldsymbol{c_k y}] \\ [\boldsymbol{c_1 c_2}] & [\boldsymbol{c_2 c_2}] & \dots & [\boldsymbol{c_2 c_k}] \\ \vdots & \vdots & \ddots & \vdots \\ [\boldsymbol{c_1 c_k}] & [\boldsymbol{c_2 c_k}] & \dots & [\boldsymbol{c_k c_k}] \end{vmatrix}}{\begin{vmatrix} [\boldsymbol{c_1 c_1}] & [\boldsymbol{c_1 c_2}] & \dots & [\boldsymbol{c_1 c_k}] \\ [\boldsymbol{c_1 c_2}] & [\boldsymbol{c_2 c_2}] & \dots & [\boldsymbol{c_2 c_k}] \\ \vdots & \vdots & \ddots & \vdots \\ [\boldsymbol{c_1 c_k}] & [\boldsymbol{c_2 c_k}] & \dots & [\boldsymbol{c_k c_k}] \end{vmatrix}}. \tag{2.12}$$

Las demás componentes del vector solución $\boldsymbol{b}_{min}$ se pueden obtener siguiendo el mismo procedimiento.

Para el caso general en el que el modelo es no lineal en los parámetros, el tratamiento estándar consiste en reducir el modelo $f_i(\boldsymbol{b})$ a la forma lineal $\langle f_i(\boldsymbol{b}) \rangle$ mediante su aproximación de Taylor de primer grado alrededor de un punto inicial $\boldsymbol{b}_0$. Introduciendo $\boldsymbol{\delta} = (\boldsymbol{b} - \boldsymbol{b}_0)$ se tiene

$$\langle y_i \rangle = \langle f_i(\boldsymbol{b}) \rangle = f_i(\boldsymbol{b}_0) + \sum_{j=1}^{k} \frac{\partial f_i(\boldsymbol{b}_0)}{\partial b_j} \delta_j, \tag{2.13}$$

donde $\langle y_i \rangle$ es el i-ésimo dato predicho por el modelo linealizado. El residuo correspondiente a esta versión aproximada del modelo se define como

$$\langle r_i \rangle = [y_i - f_i(\boldsymbol{b}_0)] - \sum_{j=1}^{k} \frac{\partial f_i(\boldsymbol{b}_0)}{\partial b_j} \delta_j, \tag{2.14}$$

el cual, al igual que (2.7), posee forma lineal.

A partir de este punto es posible seguir el procedimiento descrito al inicio

para obtener el vector solución $\boldsymbol{b}_{min}$; para ello basta hacer la identificación

$$c_{ji} = \frac{\partial f_i(\boldsymbol{b}_0)}{\partial b_j}.$$

2.2.2. Enunciado de Gauss de la Media Aritmética

Para deducir la ley de probabilidad del error aleatorio de medición, Gauss utilizó el siguiente enunciado: *dado cualquier número de medidas* y_1, y_2, $y_3, \ldots$, *de una cantidad desconocida* η, *el valor* η^* *con mayor probabilidad de ser dicho* η *desconocido, es la media aritmética de las medidas.* Cabe señalar que el valor real de η, denotado por y^*, no es necesariamente igual a η^*.

Este enunciado se puede derivar a partir de propiedades y axiomas de naturaleza más evidente, formulados de la siguiente manera:

Propiedad I - Las diferencias entre el valor más probable y las medidas individuales no dependen de la posición del punto cero desde el cual son calculadas.

Propiedad II - La razón entre el valor más probable y cualquier medida individual no depende de la unidad de medida utilizada.

Axioma I - El valor más probable es independiente del orden en el que se realizan las medidas, por lo tanto es una función simétrica de las medidas.

Axioma II - El valor más probable, considerado como una función de las medidas individuales, posee primeras derivadas continuas con respecto a ellas.

Si se expresa el valor más probable η^* en términos de las s medidas $y_1, y_2, \ldots, y_s$ mediante la función $g(y_1, y_2, \ldots, y_s)$, entonces por el teorema del valor medio, aplicable gracias al axioma II, se tiene

$$g(\boldsymbol{y}) = g(y_1, y_2, \ldots, y_s) = g(0, 0, \ldots, 0) + y_1\left[\frac{\partial g}{\partial y_1}\right] + y_2\left[\frac{\partial g}{\partial y_2}\right] + \ldots + y_s\left[\frac{\partial g}{\partial y_s}\right], \tag{2.15}$$

donde los corchetes denotan que las derivadas parciales están evaluadas en el punto $t\boldsymbol{y}$, con $t \in [0, 1]$.

Según la propiedad II se tiene que la ecuación

$$g(ky_1, ky_2, \ldots, ky_s) = k\, g(y_1, y_2, \ldots, y_s)$$

debe ser válida para todo $k \neq 0$. Evaluando en particular en el límite $k \to 0$, y considerando la continuidad de g, se obtiene

$$g(0, 0, \ldots, 0) = 0,$$

y así (2.15) se transforma en

$$g(y_1, y_2, \ldots, y_s) = \alpha_1(\boldsymbol{y})y_1 + \alpha_2(\boldsymbol{y})\, y_2 + \ldots + \alpha_s(\boldsymbol{y})\, y_s.$$

De acuerdo al axioma I, todos los $\alpha' s$ deben ser iguales, por lo tanto

$$g(y_1, y_2, \ldots, y_s) = \alpha(\boldsymbol{y})[y_1 + y_2 + \ldots + y_s].$$

Finalmente, la propiedad I se traduce en

$$g(y_1 + h, y_2 + h, \ldots, y_s + h) = g(y_1, y_2, \ldots, y_s) + h,$$

de donde se desprende

$$\alpha(\boldsymbol{y}) = \frac{1}{s}.$$

De estos resultados se concluye que la forma de g es

$$\eta^* = g(y_1, y_2, \ldots, y_s) = \frac{1}{s}\,(y_1 + y_2 + \ldots + y_s), \tag{2.16}$$

que corresponde a la media aritmética de las medidas.

2.2.3. Demostración de Gauss de la Ley Normal del Error

A partir del enunciado de la media aritmética es posible deducir que la densidad de probabilidad del error de medición Δ sigue una ley normal.

Si $p(\Delta)$ es la función de densidad de probabilidad de Δ, y ϵ es la menor cantidad a la cual el instrumento de medición es sensible, entonces la probabilidad de obtener un error de valor Δ_0 es $p(\Delta_0)\epsilon$.

Si para una cierta cantidad η, cuyo valor real es y^*, se tienen s medidas $y_1,\, y_2, \ldots,\, y_s$ en las que se introducen los errores $\Delta_1 = y_1 - y^*,\ \Delta_2 = y_2 - y^*, \ldots,$ $\Delta_s = y_s - y^*$, entonces la probabilidad de que ocurran dichas medidas es

$$\epsilon^s p(y_1 - y^*)p(y_2 - y^*) \ldots p(y_s - y^*).$$

Si se asume, antes del proceso de medición, que todos los valores de η tienen la misma probabilidad de ser el valor real a medir, entonces por el teorema de Bayes de la probabilidad inductiva [16] se concluye que, una vez realizadas las mediciones, la probabilidad de que el valor real de η esté en el intervalo $[y^*,\, y^* + \mathrm{d}y^*]$ es

$$\frac{\epsilon^s p(y_1 - y^*)p(y_2 - y^*) \ldots p(y_s - y^*)\,\mathrm{d}y^*}{\int_{-\infty}^{\infty} \epsilon^s p(y_1 - \eta)p(y_2 - \eta) \ldots p(y_s - \eta)\,\mathrm{d}\eta},$$

y por consiguiente la hipótesis más probable, con respecto al verdadero valor de η, es aquel valor de y^* que maximice la expresión

$$p(y_1 - y^*)p(y_2 - y^*) \ldots p(y_s - y^*).$$

Dicho y^*, simbolizado por η^*, satisface por ende la ecuación

$$\sum_{q=1}^{s} \frac{\mathrm{d}}{\mathrm{d}\eta} \ln\left[p(y_q - \eta^*)\right] = 0. \tag{2.17}$$

Por otro lado, según el postulado de la media aritmética, se tiene que

$$\sum_{q=1}^{s} (y_q - \eta^*) = 0. \tag{2.16a}$$

Condición suficiente para que exista compatibilidad entre (2.17) y (2.16a) es que la función de densidad p satisfaga la ecuación diferencial

$$\frac{\mathrm{d}}{\mathrm{d}\eta} \ln\left[p(y_q - \eta)\right] = c(y_q - \eta), \tag{2.18}$$

donde c es una constante arbitraria.

Efectuando el cambio de variable $\Delta = (y_q - \eta)$, se obtiene la solución de

(2.18)
$$p(\Delta) = Ae^{-\frac{1}{2}c\Delta^2}.$$

Para determinar el valor particular de la constante de integración A, se considera que $p(\Delta)$ posee, por ser función de densidad de probabilidad, la propiedad

$$\int_{-\infty}^{\infty} p(\Delta)\,\mathrm{d}\Delta = 1,$$

por lo tanto

$$A\int_{-\infty}^{\infty} e^{-\frac{1}{2}c\Delta^2}\,\mathrm{d}\Delta = 1;$$

y puesto que

$$\int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x = \sqrt{\pi},$$

entonces

$$A = \sqrt{\frac{c}{2\pi}}.$$

Finalmente, introduciendo $\sigma = \sqrt{2/c}$ se obtiene

$$p(\Delta) = \frac{1}{\sqrt{\pi}\sigma}\, e^{-\frac{\Delta^2}{\sigma^2}}, \tag{2.19}$$

la cual es la función de densidad normal.

2.2.4. Deducción de Gauss del Criterio de los Mínimos Cuadrados

Como paso previo a la deducción del criterio de los mínimos cuadrados, es conveniente introducir el concepto de 'peso' asociado a cada medida. En términos generales, el peso de una medida indica el grado de precisión involucrado en el proceso de medición del que resulta dicha medida. De esta manera, si una medida está menos propensa a error, mayor es su peso.

En conformidad con lo arriba expresado, algunas posibles definiciones del peso w asociado a una cierta medición, que introduce un error Δ con densidad de probabilidad $p(\Delta)$, son:

- la densidad de probabilidad de que la medida esté libre de errores

$$w = p(0),$$

- la probabilidad de que el error caiga dentro de un cierto rango de tolerancia $|\Delta| \leq \Delta_0$

$$w = \int_{-\Delta_0}^{\Delta_0} p(\Delta)\,\mathrm{d}\Delta,$$

- la inversa del tamaño del intervalo centrado en cero dentro del cual los errores caen con cierta probabilidad P_0

$$P_0 = \int_{-1/2w}^{1/2w} p(\Delta)\,\mathrm{d}\Delta.$$

Sin embargo, el peso se define formalmente como la inversa del cuadrado de la desviación estándar del error de medición

$$w = \left[\int_{-\infty}^{\infty} \Delta^2 p(\Delta)\,\mathrm{d}\Delta\right]^{-1}. \tag{2.20}$$

Esta definición es adecuada puesto que la desviación estándar, al ser la raiz cuadrada del error cuadrático medio, constituye una medida de la presencia del error, el cual está en relación inversa con la precisión y por tanto con el peso. Para el caso tratado, de densidad de error normal, se tiene a partir de (2.19) y (2.20)

$$w = \sigma^{-2}. \tag{2.21}$$

Dadas n respuestas con valores reales y_1^*, $y_2^*, \ldots,$ y_n^*, la probabilidad de obtener, a partir de estas, sendas medidas y_1, $y_2, \ldots,$ y_n es

$$P = \frac{1}{\sqrt{\pi}\sigma_1}\, e^{-\frac{(y_1 - y_1^*)^2}{\sigma_1^2}}\, \frac{1}{\sqrt{\pi}\sigma_2}\, e^{-\frac{(y_2 - y_2^*)^2}{\sigma_2^2}} \cdots \frac{1}{\sqrt{\pi}\sigma_n}\, e^{-\frac{(y_n - y_n^*)^2}{\sigma_n^2}},$$

donde se considera que los valores de la desviación estándar del error introducido en el proceso de medición de cada respuesta son, en general, distintos entre sí.

El criterio que define el método de los mínimos cuadrados establece que el ajuste óptimo de un modelo a un conjunto de medidas $y_1, y_2, \ldots, y_n$ se consigue cuando se maximiza la probabilidad de obtener dichas medidas, adoptando los correspondientes valores predichos por el modelo $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ como los valores reales de los que se obtienen las medidas, i.e., cuando el vector de parámetros de ajuste es tal que maximice la expresión

$$P = \frac{1}{\sqrt{\pi}\sigma_1} e^{-\frac{(y_1-\hat{y}_1)^2}{\sigma_1^2}} \frac{1}{\sqrt{\pi}\sigma_2} e^{-\frac{(y_2-\hat{y}_2)^2}{\sigma_2^2}} \cdots \frac{1}{\sqrt{\pi}\sigma_n} e^{-\frac{(y_n-\hat{y}_n)^2}{\sigma_n^2}},$$

o, en forma equivalente al tomar logaritmos y considerar (2.21), que minimice

$$\begin{aligned} \Phi &= w_1(y_1 - \hat{y}_1)^2 + w_2(y_2 - \hat{y}_2)^2 + \cdots + w_n(y_n - \hat{y}_n)^2 \\ &= \sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2. \end{aligned} \tag{2.22}$$

En este caso general, al método se le denomina mínimos cuadrados ponderados, sin embargo la mayor cantidad de aplicaciones del método se dan para el caso especial en que todos los procesos de medición poseen la misma precisión o peso $w_i = w$ $(i = 1, 2, \ldots, n)$, de manera que (2.22) se convierte en

$$\Phi = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

que no es otra cosa que (2.3).

2.2.5. Deducción Alterna de Laplace

Posteriormente a la primera deducción de Gauss, tratada en las subsecciones precedentes, el método de los mínimos cuadrados fue derivado de una manera completamente distinta por Laplace [17], cuyo enfoque influenció subsecuentes investigaciones en el tema; entre ellas, un trabajo de Gauss [10] en el que expone su segunda deducción del método.

El principio en el que se basa esta deducción se puede enunciar en los siguientes términos:

Para las n respuestas predichas por el modelo mediante las expresiones

lineales

$$\left.\begin{array}{rcccccccc} \hat{y}_1 & = & c_{11}b_1 & + & c_{21}b_2 & + & \ldots & + & c_{k1}b_k \\ \hat{y}_2 & = & c_{12}b_1 & + & c_{22}b_2 & + & \ldots & + & c_{k2}b_k \\ \vdots & & \vdots & & \vdots & & \ddots & & \vdots \\ \hat{y}_n & = & c_{1n}b_1 & + & c_{2n}b_2 & + & \ldots & + & c_{kn}b_k, \end{array}\right\}$$

se tienen respectivamente las estimaciones $y_1, y_2, \ldots, y_n$ obtenidas mediante medición, donde n se supone mayor que el número de parámetros k.

Para la cantidad

$$\begin{aligned} \mu_1(c_{11}b_1 + c_{21}b_2 + \cdots + c_{k1}b_k) + \mu_2(c_{12}b_1 + c_{22}b_2 + \cdots + c_{k2}b_k) + \ldots \\ + \mu_n(c_{1n}b_1 + c_{2n}b_2 + \cdots + c_{kn}b_k) \end{aligned} \qquad \text{(A)}$$

se tiene entonces la estimación

$$\mu_1 y_1 + \mu_2 y_2 + \cdots + \mu_n y_n. \qquad \text{(B)}$$

Si los μ_i se escogen de forma tal que en la expresión (A) el coeficiente resultante de b_1 es uno y los coeficientes resultantes de $b_2, b_3, \ldots, b_k$ son cero, i.e.,

$$\left.\begin{array}{rcccccccc} c_{11}\mu_1 & + & c_{12}\mu_2 & + & \ldots & + & c_{1n}\mu_n & = & 1 \\ c_{21}\mu_1 & + & c_{22}\mu_2 & + & \ldots & + & c_{2n}\mu_n & = & 0 \\ \vdots & & \vdots & & \ddots & & \vdots & & \vdots \\ c_{k1}\mu_1 & + & c_{k2}\mu_2 & + & \ldots & + & c_{kn}\mu_n & = & 0, \end{array}\right\} \qquad (2.23)$$

entonces la expresión (B) es una estimación de b_1.

Puesto que existen infinitos juegos de μ_i que son solución de (2.23), el problema consiste en establecer condiciones adicionales para determinar el juego de valores μ_i que asegure que (B) sea la mejor estimación posible de b_1. Esto se consigue considerando la noción de peso de una estimación y definiendo como mejor estimación a aquella que posea el mayor peso.

Si se tiene que los pesos asociados a las estimaciones $y_1, y_2, \ldots, y_n$ son respectivamente $w_1, w_2, \ldots, w_n$, entonces es posible demostrar [7] que el peso W asociado a la expresión lineal (B) satisface

$$\frac{1}{W} = \frac{\mu_1^2}{w_1} + \frac{\mu_2^2}{w_2} + \cdots + \frac{\mu_n^2}{w_n}. \qquad (2.24)$$

Para determinar el juego de valores μ_i que minimicen la inversa del peso $(1/W)$ y, simultáneamente, cumplan con el sistema de restricciones (2.23), se emplea el método de los multiplicadores de Lagrange (véase sección 4.2.1), de cuya aplicación se obtiene el sistema de ecuaciones

$$\left.\begin{array}{ccccccccccc} \mu_1 & + & w_1c_{11}\lambda_1 & + & w_1c_{21}\lambda_2 & + & \ldots & + & w_1c_{k1}\lambda_k & = & 0 \\ \mu_2 & + & w_2c_{12}\lambda_1 & + & w_2c_{22}\lambda_2 & + & \ldots & + & w_2c_{k2}\lambda_k & = & 0 \\ \vdots & & \vdots & & \vdots & & \ddots & & \vdots & & \vdots \\ \mu_n & + & w_nc_{1n}\lambda_1 & + & w_nc_{2n}\lambda_2 & + & \ldots & + & w_nc_{kn}\lambda_k & = & 0, \end{array}\right\} \tag{2.25}$$

que junto al sistema de ecuaciones (2.23) determinan unívocamente el juego de valores μ_i con los que se calcula la mejor estimación $(b_{min})_1$ del parámetro b_1 mediante la expresión

$$(b_{min})_1 = \mu_1 y_1 + \mu_2 y_2 + \cdots + \mu_n y_n. \tag{2.26}$$

Despejando los μ_i en (2.25) e introduciéndolos en (2.23) y (2.26) se obtienen las ecuaciones

$$\left.\begin{array}{ccccccccccc} (b_{min})_1 & + & [\boldsymbol{wc_1y}]\lambda_1 & + & [\boldsymbol{wc_2y}]\lambda_2 & + & \ldots & + & [\boldsymbol{wc_ky}]\lambda_k & = & 0 \\ 1 & + & [\boldsymbol{wc_1c_1}]\lambda_1 & + & [\boldsymbol{wc_1c_2}]\lambda_2 & + & \ldots & + & [\boldsymbol{wc_1c_k}]\lambda_k & = & 0 \\ & & [\boldsymbol{wc_1c_2}]\lambda_1 & + & [\boldsymbol{wc_2c_2}]\lambda_2 & + & \ldots & + & [\boldsymbol{wc_2c_k}]\lambda_k & = & 0 \\ & & \vdots & & \vdots & & \ddots & & \vdots & & \vdots \\ & & [\boldsymbol{wc_1c_k}]\lambda_1 & + & [\boldsymbol{wc_2c_k}]\lambda_2 & + & \ldots & + & [\boldsymbol{wc_kc_k}]\lambda_k & = & 0, \end{array}\right\}$$

a partir de las cuales, siguiendo la técnica descrita en la sección 2.2.1, se deduce

$$\begin{vmatrix} 1 & [\boldsymbol{wc_1y}] & [\boldsymbol{wc_2y}] & \ldots & [\boldsymbol{wc_ky}] \\ 1/(b_{min})_1 & [\boldsymbol{wc_1c_1}] & [\boldsymbol{wc_1c_2}] & \ldots & [\boldsymbol{wc_1c_k}] \\ 0 & [\boldsymbol{wc_1c_2}] & [\boldsymbol{wc_2c_2}] & \ldots & [\boldsymbol{wc_2c_k}] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & [\boldsymbol{wc_1c_k}] & [\boldsymbol{wc_2c_k}] & \ldots & [\boldsymbol{wc_kc_k}] \end{vmatrix} = 0,$$

de donde finalmente se despeja el valor buscado

$$(b_{min})_1 = \frac{\begin{vmatrix} [\boldsymbol{wc_1y}] & [\boldsymbol{wc_2y}] & \dots & [\boldsymbol{wc_ky}] \\ [\boldsymbol{wc_1c_2}] & [\boldsymbol{wc_2c_2}] & \dots & [\boldsymbol{wc_2c_k}] \\ \vdots & \vdots & \ddots & \vdots \\ [\boldsymbol{wc_1c_k}] & [\boldsymbol{wc_2c_k}] & \dots & [\boldsymbol{wc_kc_k}] \end{vmatrix}}{\begin{vmatrix} [\boldsymbol{wc_1c_1}] & [\boldsymbol{wc_1c_2}] & \dots & [\boldsymbol{wc_1c_k}] \\ [\boldsymbol{wc_1c_2}] & [\boldsymbol{wc_2c_2}] & \dots & [\boldsymbol{wc_2c_k}] \\ \vdots & \vdots & \ddots & \vdots \\ [\boldsymbol{wc_1c_k}] & [\boldsymbol{wc_2c_k}] & \dots & [\boldsymbol{wc_kc_k}] \end{vmatrix}}. \tag{2.27}$$

Para el caso especial en que $w_1 = w_2 = \cdots = w_n$, se puede apreciar que el valor resultante de $(b_{min})_1$ es idéntico al obtenido en (2.12), el cual es la solución de las ecuaciones normales ordinarias dictadas por el método de los mínimos cuadrados; por lo tanto, el enfoque alterno de Laplace conduce al establecimiento de dicho método.

2.3. Interpretación Geométrico Vectorial del Método según Kolmogorov

La originalidad e importancia del trabajo de Kolmogorov [15], concerniente al método de los mínimos cuadrados, radica no en la propuesta de algún nuevo procedimiento de deducción del método, sino en su interpretación enmarcada en el campo de la geometría vectorial n-dimensional.

En esta sección se presentan los aspectos más relevantes del mencionado trabajo en el que mediante un ejemplo del caso más simple, correspondiente a un modelo lineal, se muestra cómo la aplicación de métodos generales del álgebra lineal vectorial, tales como el concepto de ortogonalidad, permite obtener de manera mucho más transparente todos los resultados básicos del método, así como los instrumentos para evaluar la confiabilidad de las estimaciones involucradas.

2.3.1. Obtención de las Ecuaciones Normales Mediante Métodos Vectoriales

Considérese un sistema cuyo modelo exacto $f(\boldsymbol{x}, \boldsymbol{b})$ es lineal en los parámetros b_j, $j = 1, 2, \ldots, k$. Por definición, para dicho modelo existe un vector de parámetros exactos $\boldsymbol{\beta}$, a priori desconocidos, con los cuales se calculan las respuestas del sistema (y_i^*) mediante la expresión lineal

$$y_i^* = \sum_{j=1}^{k} c_{ji}\beta_j, \qquad i = 1, 2, \ldots, n, \tag{2.28}$$

donde los coeficientes c_{ji}, al estar en función de las entradas $\boldsymbol{x_i}$, son conocidos.

Asimismo, asúmase que dados los valores de y_i^* y $\boldsymbol{x_i}$, los β_j están determinados de manera única por (2.28). Esto implica que el rango de la matriz $[c_{ji}]^{n\times k}$ no es menor que k, por lo tanto se debe cumplir que $n \geq k$.

Sin embargo, la única forma de conocer las respuestas del sistema (y_i^*) es mediante un proceso de medición que inevitablemente introduce error. Así en lugar de obtener los verdaderos valores (y_i^*), se obtienen experimentalmente los valores

$$y_i = y_i^* + \Delta_i, \qquad i = 1, 2, \ldots, n, \tag{2.29}$$

lo que elimina la posibilidad de obtener los β_j a partir de (2.28), ya que los errores Δ_i se introducen en (2.28) como incógnitas adicionales.

Ante la imposibilidad de conocer los parámetros exactos β_j, se utilizan parámetros con valores alternos arbitrarios b_j, por lo que el modelo ya no entrega las respuestas reales (y_i^*), sino las respuestas predichas

$$\hat{y}_i = \sum_{j=1}^{k} c_{ji}b_j, \qquad i = 1, 2, \ldots, n, \tag{2.30}$$

las cuales, en general, son distintas a las primeras.

Como ya se indicó, el método de los mínimos cuadrados define los residuos

$$r_i = y_i - \hat{y}_i, \qquad i = 1, 2, \ldots, n, \tag{2.31}$$

y establece que los parámetros b_j que mejor substituyen a los parámetros exactos β_j en lograr que el modelo caracterice al sistema, son aquellos que

minimizan la norma del residuo

$$[\boldsymbol{rr}] = \sum_{i=1}^{n} r_i r_i, \tag{2.32}$$

donde los corchetes denotan el producto escalar de los vectores contenidos.

Es posible demostrar, aplicando cálculo elemental, que los b_j que minimizan (2.32) están determinados unívocamente por el sistema de ecuaciones normales

$$\sum_{j=1}^{k} [\boldsymbol{c_h c_j}] b_j = [\boldsymbol{c_h y}], \qquad h = 1, 2, \ldots, k. \tag{2.33}$$

Sin embargo, el uso del cálculo no es la única manera de establecer la equivalencia entre cumplir la condición de minimizar (2.32) y ser solución de las ecuaciones (2.33); también es posible obtener dicho resultado mediante el uso de métodos vectoriales.

Con ese fin, considérense y_i^*, c_{ji}, y_i, Δ_i, $\hat{y}_i$, r_i, $(i = 1, 2, \ldots, n)$, como las componentes de los vectores n-dimensionales $\boldsymbol{y^*}$, $\boldsymbol{c_j}$, $\boldsymbol{y}$, $\boldsymbol{\Delta}$, $\boldsymbol{\hat{y}}$, $\boldsymbol{r}$, con lo cual (2.28) - (2.31) se reescriben en forma vectorial

$$\boldsymbol{y^*} = \sum_{j=1}^{k} \beta_j \boldsymbol{c_j}, \tag{2.28a}$$

$$\boldsymbol{y} = \boldsymbol{y^*} + \boldsymbol{\Delta}, \tag{2.29a}$$

$$\boldsymbol{\hat{y}} = \sum_{j=1}^{k} b_j \boldsymbol{c_j}, \tag{2.30a}$$

$$\boldsymbol{r} = \boldsymbol{y} - \sum_{j=1}^{k} b_j \boldsymbol{c_j}. \tag{2.31a}$$

Si se denota con V al subespacio lineal generado dentro del espacio vectorial $\mathbf{R}^n$ por el conjunto de vectores $\{\boldsymbol{c_j}\}$, se desprende entonces de (2.30a) que el vector $\boldsymbol{\hat{y}}$ está contenido en V ($\boldsymbol{\hat{y}} \in V$).

En lenguaje vectorial, cumplir la condición de minimizar $[\boldsymbol{rr}]$ equivale a que $\boldsymbol{\hat{y}}$ sea la proyección ortogonal de $\boldsymbol{y}$ sobre V y $\boldsymbol{r}$ sea el vector complementario ortogonal a V, con lo cual

$$[\boldsymbol{r c_j}] = 0, \qquad j = 1, 2, \ldots, k. \tag{2.34}$$

La figura 2.2 ilustra el trazado, en conformidad con el método, de los

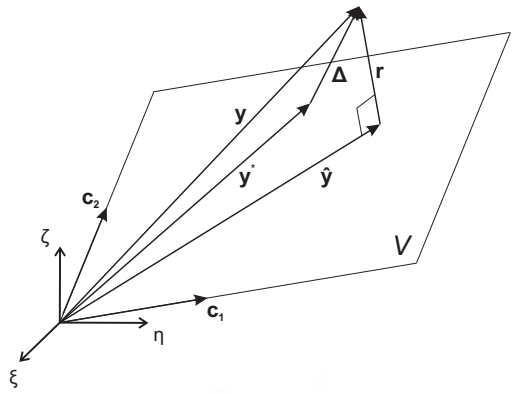


Figura 2.2: Representación vectorial del método para $n = 3$ y $k = 2$.

vectores mencionados para un caso arbitrario en que el número de muestras es $n = 3$ y el número de parámetros es $k = 2$. Nótese que en este caso el subespacio lineal V corresponde a un plano en $\mathbf{R}^3$ que pasa por el origen.

Reemplazando en (2.34) la expresión para $\boldsymbol{r}$ dada por (2.31a) se obtienen las ecuaciones

$$\sum_{j=1}^{k} [\boldsymbol{c_h c_j}] b_j = [\boldsymbol{c_h y}], \qquad h = 1, 2, \ldots, k, \tag{2.33}$$

correspondientes al sistema de ecuaciones normales formulado en (2.33).

Este sistema se puede reescribir en forma matricial (véase (2.9), en la página 15),

$$A\,\boldsymbol{b} = \boldsymbol{g}, \tag{2.35}$$

donde el determinante del sistema,

$$|A| = |[\boldsymbol{c_h c_j}]|,$$

es el determinante de Gram de los vectores $\boldsymbol{c_1}, \boldsymbol{c_2}, \ldots, \boldsymbol{c_k}$.

Bajo el supuesto, enunciado líneas arriba, que la matriz $[c_{ji}]^{n\times k}$ posee rango k, se deduce que los vectores $\boldsymbol{c_1}, \boldsymbol{c_2}, \ldots, \boldsymbol{c_k}$ son linealmente independientes y, en consecuencia, su determinante de Gram es distinto de cero ($|A| \neq 0$). Por lo tanto, al ser A invertible, (2.33) determina unívocamente los parámetros b_j.

2.3.2. Evaluación de la Confiabilidad de los Parámetros Estimados

Hasta este punto se ha visto cómo el método estima los parámetros b_j a partir de las respuestas medidas (sujetas a error) ante la inaccesibilidad de las respuestas reales necesarias para calcular los parámetros verdaderos β_j.

Es importante entonces establecer un criterio para evaluar el grado con que se desvían los parámetros estimados respecto a los parámetros reales; lo cual será un indicativo de la confiabilidad de los resultados. Esto se consigue analizando la esperanza y la varianza de los parámetros b_j. La esperanza $E\{b_j\}$ indica el valor al que converge la media de los parámetros, por lo que un valor distinto de β_j (de conocerse β_j) señalaría la presencia de error sistemático en los resultados. La varianza $V\{b_j\}$ por otro lado, al ser el valor cuadrático medio de la desviación de b_j respecto a su media, determina la dispersión o, visto a la inversa, la exactitud de los resultados.

Antes de proceder al cálculo de $E\{b_j\}$ y $V\{b_j\}$ es necesario asumir las siguientes condiciones respecto a la distribución de probabilidad de los errores de medición Δ_i:

C1) Δ_i y Δ_l, con $i \neq l$, son variables aleatorias independientes.

C2) $E\{\Delta_i\} = 0$.

C3) $V\{\Delta_i\} = E\{\Delta_i^2\} = s^2$, donde s^2 es finito e independiente de i.

C4) $E\{\Delta_i \Delta_l\} = 0$ para $i \neq l$.

Como punto de partida, defínase el sistema de vectores $\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_k}$, en el subespacio lineal V, de forma que sea δ_k-ortogonal al sistema $\boldsymbol{c_1}, \boldsymbol{c_2}, \ldots, \boldsymbol{c_k}$, i.e., que cumpla

$$[\boldsymbol{c_j} \boldsymbol{u_h}] = I_{jh}^{(k)}, \tag{2.36}$$

donde $I^{(k)}$ es la matriz identidad de $k \times k$.

Puesto que cada uno de los sistemas $\{\boldsymbol{c_1}, \boldsymbol{c_2}, \ldots, \boldsymbol{c_k}\}$ y $\{\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_k}\}$ es una base generadora de V, se tiene que

$$\boldsymbol{u_j} = \sum_{\lambda=1}^{k} Q_{j\lambda} \boldsymbol{c_\lambda}, \qquad \boldsymbol{c_j} = \sum_{\lambda=1}^{k} A_{j\lambda} \boldsymbol{u_\lambda}.$$

Tomando el producto escalar de la primera igualdad con $\boldsymbol{u_h}$ y el producto escalar de la segunda con $\boldsymbol{c_h}$, y considerando (2.36), se tiene

$$Q_{jh} = [\boldsymbol{u_j u_h}], \qquad A_{jh} = [\boldsymbol{c_j c_h}], \tag{2.37}$$

con lo cual,

$$\boldsymbol{u_j} = \sum_{h=1}^{k} [\boldsymbol{u_j u_h}] \boldsymbol{c_h}, \tag{2.38}$$

$$\boldsymbol{c_j} = \sum_{h=1}^{k} [\boldsymbol{c_j c_h}] \boldsymbol{u_h}. \tag{2.39}$$

De estas dos últimas ecuaciones y de (2.37), se desprende que la matriz Q es la inversa de la matriz A.

Tomando ahora el producto escalar de (2.28a) y $\boldsymbol{u_j}$, se obtiene

$$\beta_j = [\boldsymbol{y^* u_j}], \tag{2.40}$$

y en forma similar, (2.30a) implica que

$$b_j = [\boldsymbol{\hat{y}\, u_j}]. \tag{2.41}$$

Por otro lado, puesto que $\boldsymbol{r}$ es un vector ortogonal a V, se cumple que

$$[\boldsymbol{r\, u_j}] = 0,$$

con lo cual, de (2.41) y (2.31a), se determina que

$$b_j = [\boldsymbol{y\, u_j}]. \tag{2.42}$$

De esta última ecuación, junto con (2.29a) y (2.40), se deduce que

$$b_j = \beta_j + [\boldsymbol{\Delta\, u_j}], \tag{2.43}$$

y por ende, considerando C2), se llega al resultado

$$E\{b_j\} = \beta_j, \tag{2.44}$$

que establece que la esperanza de cada uno de los parámetros estimados (aproximados), es igual al correspondiente parámetro verdadero. Se deduce entonces que los valores de los parámetros estimados no contienen error sistemático sí y solo sí se cumple la condición C2).

Obtenido el valor de la esperanza de los parámetros b_j, resta determinar la varianza de los mismos, que indicará la precisión con la que cada parámetro b_j aproxima en promedio a su respectivo β_j. Para ello se calcula primero la expresión más general correspondiente a $E\{(b_j - \beta_j)(b_h - \beta_h)\}$, la cual, empleando (2.43), C3) y C4), da como resultado

$$E\{(b_j - \beta_j)(b_h - \beta_h)\} = E\{[\boldsymbol{\Delta}\, \boldsymbol{u_j}][\boldsymbol{\Delta}\, \boldsymbol{u_h}]\} =$$
$$\sum_{i=1}^{n}\sum_{l=1}^{n} E\{\Delta_i \Delta_l\} u_{ji} u_{hl} = s^2 \sum_{i=1}^{n} u_{ji} u_{hi} = [\boldsymbol{u_j u_h}] s^2,$$

que, en vista de (2.37), se puede expresar como

$$E\{(b_j - \beta_j)(b_h - \beta_h)\} = Q_{jh} s^2. \tag{2.45}$$

Finalmente, para el caso particular $h = j$, se obtiene el valor buscado de la varianza de b_j

$$V\{b_j\} = E\{(b_j - \beta_j)^2\} = Q_{jj} s^2. \tag{2.46}$$

Nótese que la varianza de los b_j es proporcional a la varianza de los errores de medición ($V\{\Delta_i\} = s^2$). Por esta razón es necesario, en caso de no conocerse *a priori* el valor de s, establecer un valor σ que sirva como aproximación del mismo.

Con ese fin, se define el vector

$$\boldsymbol{\Delta}^* = \hat{\boldsymbol{y}} - \boldsymbol{y}^*, \tag{2.47}$$

el cual, por (2.29a) y (2.31a), se puede expresar como

$$\boldsymbol{\Delta} = \boldsymbol{\Delta}^* + \boldsymbol{r}. \tag{2.48}$$

Puesto que $\boldsymbol{\Delta}^*$ está contenido en V junto con $\hat{\boldsymbol{y}}$ e $\boldsymbol{y}^*$, y $\boldsymbol{r}$ es ortogonal a V, se concluye que esta última ecuación representa la descomposición de $\boldsymbol{\Delta}$ en su proyección ortogonal sobre V y el complemento ortogonal a dicha proyección.

Si se define la base ortonormal del espacio vectorial n-dimensional

$$\boldsymbol{e_\tau} = (e_{\tau 1},\, e_{\tau 2}, \ldots,\, e_{\tau n}), \qquad \tau = 1, 2, \ldots, n,$$
$$[\boldsymbol{e_\tau e_{\tau'}}] = I^{(n)}_{\tau\tau'},$$

de forma tal que los k primeros vectores estén contenidos en V y los $n-k$ restantes sean ortogonales a V; entonces, considerando

$$\tilde{\Delta}_\tau = [\boldsymbol{\Delta e_\tau}], \tag{2.49}$$

se obtiene

$$\boldsymbol{\Delta} = \sum_{\tau=1}^{n} \tilde{\Delta}_\tau \boldsymbol{e_\tau}, \tag{2.50}$$

$$\boldsymbol{\Delta}^* = \sum_{\tau=1}^{k} \tilde{\Delta}_\tau \boldsymbol{e_\tau}, \tag{2.51}$$

$$\boldsymbol{r} = \sum_{\tau=k+1}^{n} \tilde{\Delta}_\tau \boldsymbol{e_\tau}. \tag{2.52}$$

A partir de la propiedad de ortonormalidad de los vectores $\boldsymbol{e_\tau}$ de la base, se deduce que

$$[\boldsymbol{rr}] = \sum_{\tau=k+1}^{n} \tilde{\Delta}_\tau \tilde{\Delta}_\tau, \tag{2.53}$$

y puesto que por (2.49) y C3)

$$E\{\tilde{\Delta}_\tau \tilde{\Delta}_\tau\} = E\{\sum_{i=1}^{n} \Delta_i e_{\tau i} \sum_{i'=1}^{n} \Delta'_i e_{\tau i'}\} =$$
$$\sum_{i=1}^{n}\sum_{i'=1}^{n} E\{\Delta_i \Delta'_i\} e_{\tau i} e_{\tau i'} = s^2 \sum_{i=1}^{n} e_{\tau i} e_{\tau i} = s^2 [\boldsymbol{e_\tau}] = s^2, \tag{2.54}$$

se concluye finalmente que

$$E\{[\boldsymbol{rr}]\} = \sum_{\tau=k+1}^{n} E\{\tilde{\Delta}_\tau \tilde{\Delta}_\tau\} = \sum_{\tau=k+1}^{n} s^2 = (n-k)\, s^2. \tag{2.55}$$

Definiendo

$$\sigma = \sqrt{[\boldsymbol{rr}]/(n-k)}, \tag{2.56}$$

se cumple que

$$E\{\sigma^2\} = s^2, \tag{2.57}$$

y además es posible demostrar, siguiendo un procedimiento similar al empleado, que

$$V\{\sigma^2\} = E\{(\sigma^2 - s^2)\} = 2s^2/(n-k), \tag{2.58}$$

lo que significa que σ^2 tiene un valor muy cercano a s^2, con probabilidad muy cercana a uno, para valores de $(n-k)$ grandes.

Se concluye así que conforme $(n-k)$ crece, la aproximación de s^2 mediante σ^2 mejora, y para estos casos se hace válido aproximar la varianza de los parámetros b_j dada en (2.46) mediante la expresión

$$V\{b_j\} \approx Q_{jj}\sigma^2. \tag{2.59}$$

2.4. Aplicabilidad del Criterio de los Mínimos Cuadrados cuando el Modelo es Aproximado y no Existe Error de Medición

Dado que hasta este punto se asumió conocido el modelo exacto de un sistema, el estudio del criterio de los mínimos cuadrados ha estado siempre confinado al caso en que el error aleatorio de medición es la única fuente de discrepancia entre los datos observados de un sistema y aquellos predichos por su modelo.

Sin embargo, es virtualmente imposible conocer todos los detalles internos de un sistema, lo que impide elaborar un modelo exacto del mismo. En la práctica solo se cuenta con un modelo aproximado incapaz de reproducir la verdadera respuesta del sistema que se obtiene a partir de un proceso de medición libre de error.

Para un modelo aproximado no existe un juego de parámetros de ajuste exactos o verdaderos, ya que la discrepancia entre el sistema y el modelo es ahora de naturaleza sistemática. En vista que los argumentos que justifican el uso del criterio de los mínimos cuadrados tienen su fundamento en el error aleatorio de medición, cabe cuestionarse qué tan correcto es aplicar dicho

criterio en el ajuste de modelos aproximados donde el error es sistemático y se asume la ausencia de error de medición.

La presente sección busca discutir tal cuestionamiento. Para ello se propone el diseño de una función Ψ a partir de ciertas propiedades elementales adecuadas para cuantificar la discrepancia entre el modelo aproximado y el sistema. La aplicación del criterio de los mínimos cuadrados en este nuevo contexto será entonces más o menos correcto según el grado de concordancia con dicha función Ψ.

2.4.1. Expresión General para la Función de Discrepancia

Como punto de partida hay que establecer que el error o residuo entre el valor real y_i^* (coincidente en este caso con el valor medido y_i) y el valor $\hat{y}_i$ predicho por el modelo, debe evaluarse según una escala que esté en función a su diferencia. Este residuo puede expresarse matemáticamente en forma general como

$$r_i = r(y_i, \hat{y}_i), \tag{2.60}$$

donde valores positivos o negativos significan error por exceso o defecto respectivamente.

Un caso particular de (2.60), conocido como error relativo, se presenta cuando la medida del error es proporcional a la desviación pero normalizada respecto al valor medido. La forma del residuo en este caso es

$$r_i = \frac{y_i - \hat{y}_i}{y_i}. \tag{2.61}$$

Otro caso particular se da cuando el error se define igual a la desviación. En este caso el residuo toma su forma más sencilla y se le conoce como error absoluto,

$$r_i = y_i - \hat{y}_i. \tag{2.62}$$

En este trabajo se adopta una definición del residuo independiente del signo de la desviación:

$$r_i = |y_i - \hat{y}_i|. \tag{2.63}$$

Una vez definido el error o residuo individual al representar cierto valor y_i

por medio de otro $\hat{y}_i$, resta definir el error o residuo global entre las n-tuplas $(y_1, y_2, \ldots, y_n)$ y $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$. La función de discrepancia es justamente la que cuantifica dicho residuo global mediante una expresión matemática (a determinar) en función de los residuos individuales.

Por consiguiente, cualquier intento por definir una estrategia para calcular el residuo global debe expresarse matemáticamente en función de los residuos individuales como una función de discrepancia

$$\Psi = \Psi(r_1, r_2, \ldots, r_n). \tag{2.64}$$

El diseño de Ψ se basa en las siguientes propiedades o requerimientos elementales:

P1) El residuo global no depende del orden de los residuos individuales en $\Psi(r_1, r_2, \ldots, r_n)$ y por convención se adoptará un orden ascendente de los mismos, i.e., $r_{i+1} \geq r_i$.

P2) La expresión general del residuo global contiene un número infinito de argumentos. El caso particular para n argumentos toma la forma $\Psi(r_1, r_2, \ldots, r_n) = \Psi(r_1, r_2, \ldots, r_n, 0, 0, \ldots)$. Nótese que los residuos individuales con valor cero no contribuyen al residuo global.

P3) El residuo global para el caso particular de $n = 1$ es igual al residuo individual, i.e., $\Psi(r_1) = r_1$.

P4) El residuo global es creciente respecto a cada uno de sus argumentos, i.e., $\partial\Psi/\partial r_i > 0$ en todos los casos.

P5) El residuo global posee la misma dimensión física que los residuos individuales.

La propiedad P5) significa que Ψ y los r_i se expresan en la misma unidad de medida. Por lo tanto, si se utiliza una unidad de medida contenida k veces en la unidad original, se cumple

$$\Psi(kr_1, kr_2, \ldots, kr_n) = k\,\Psi(r_1, r_2, \ldots, r_n), \qquad k > 0. \tag{2.65}$$

Esto revela que Ψ es una función homogénea de primer grado. Introduciendo $k = r_n^{-1}$ en (2.65) se obtiene la forma general que poseen todas las funciones

homogéneas de primer grado

$$\Psi = r_n\, f(\frac{r_1}{r_n},\, \frac{r_2}{r_n}, \ldots,\, \frac{r_{n-1}}{r_n}), \tag{2.66}$$

donde f es una función arbitraria cualquiera.

Derivando Ψ respecto a cada r_i y eliminando f en (2.66) resulta la ecuación diferencial denominada relación de homogeneidad de Euler

$$r_1\frac{\partial\Psi}{\partial r_1} + r_2\frac{\partial\Psi}{\partial r_2} + \cdots + r_n\frac{\partial\Psi}{\partial r_n} = \Psi, \tag{2.67}$$

cuyo conjunto solución es la totalidad de funciones homogéneas de primer grado [4].

La solución general de esta ecuación se puede construir mediante la combinación lineal de todas las soluciones particulares obtenidas por el método de la separación de variables [26].

Una función con variables separadas

$$\Psi = R_1(r_1)R_2(r_2)\cdots R_n(r_n), \tag{2.68}$$

es solución de (2.67) si y solo si

$$\frac{r_1}{R_1}\frac{\partial R_1}{\partial r_1} + \frac{r_2}{R_2}\frac{\partial R_2}{\partial r_2} + \cdots + \frac{r_n}{R_n}\frac{\partial R_n}{\partial r_n} = 1. \tag{2.69}$$

Como cada sumando en el miembro izquierdo de esta ecuación depende de una variable distinta, estos deberán ser constantes α_i de valor arbitrario que cumplan con la restricción

$$\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1.$$

Por lo tanto, (2.69) es equivalente a n ecuaciones diferenciales ordinarias de primer grado

$$r_i\frac{\partial R_i}{\partial r_i} - \alpha_i R_i = 0, \qquad i = 1, 2, \ldots, n,$$

cuyas soluciones, donde los c_i representan constantes de integración, son

$$R_i(r_i) = c_i r_i^{\alpha_i}, \qquad i = 1, 2, \ldots, n.$$

Considerando (2.66) y (2.68) se obtiene la solución particular

$$\Psi = c\, r_n \left(\frac{r_1}{r_n}\right)^{\alpha_1} \left(\frac{r_2}{r_n}\right)^{\alpha_2} \cdots \left(\frac{r_{n-1}}{r_n}\right)^{\alpha_{n-1}}, \tag{2.70}$$

donde c es una constante arbitraria.

Finalmente, la solución general de (2.67) se construye combinando linealmente las soluciones particulares (2.70) generadas al considerar todas las posibles $(n-1)$-tuplas $(\alpha_1, \alpha_2, \ldots, \alpha_{n-1})$. La solución general es entonces la integral

$$\Psi = r_n \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} c(\alpha_1, \alpha_2, \ldots, \alpha_{n-1}) \left(\frac{r_1}{r_n}\right)^{\alpha_1} \left(\frac{r_2}{r_n}\right)^{\alpha_2} \cdots \cdots \left(\frac{r_{n-1}}{r_n}\right)^{\alpha_{n-1}} \mathrm{d}\alpha_1\, \mathrm{d}\alpha_2 \ldots \mathrm{d}\alpha_{n-1} \tag{2.71}$$

donde $c(\alpha_1, \alpha_2, \ldots, \alpha_{n-1})$ es una función arbitraria compatible con las propiedades P1) – P5). Nótese además que (2.71) es una forma distinta pero equivalente de expresar (2.66).

2.4.2. Ejemplos Particulares de la Función de Discrepancia

Para identificar dentro de la expresión general (2.71) qué función de discrepancia es la apropiada para cierto problema de ajuste de modelos, es necesario derivar la forma particular de $c(\alpha_1, \alpha_2, \ldots, \alpha_{n-1})$ a partir a los requerimientos específicos de dicho problema.

Sin embargo, con el fin de mostrar algunos ejemplos de funciones de discrepancia particulares se seleccionarán arbitrariamente algunos c especiales.

Uno de los casos más sencillos se da cuando

$$\begin{aligned} c(\alpha_1, \alpha_2, \ldots, \alpha_{n-1}) = {} & \delta(\alpha_1)\delta(\alpha_2)\ldots\delta(\alpha_{n-1}) + \\ & \delta(\alpha_1 - 1)\delta(\alpha_2)\ldots\delta(\alpha_{n-1}) + \delta(\alpha_1)\delta(\alpha_2 - 1)\ldots\delta(\alpha_{n-1}) + \cdots \\ & \cdots + \delta(\alpha_1)\delta(\alpha_2)\ldots\delta(\alpha_{n-1} - 1), \end{aligned} \tag{2.72}$$

dando como resultado la función

$$\Psi = r_1 + r_2 + \ldots + r_n, \tag{2.73}$$

correspondiente al criterio denominado de los mínimos errores absolutos.

Un caso más general y complejo se presenta cuando

$$c(\alpha_1, \alpha_2, \ldots, \alpha_{n-1}) = \sum_{m=0}^{\infty} \binom{1/\beta}{m} \sum_{m_1=0}^{m} \sum_{m_2=0}^{m-m_1} \cdots \sum_{m_{n-2}=0}^{m-m_1-\ldots-m_{n-3}} \frac{m!}{m_1!\, m_2! \cdots m_{n-1}!} \delta(\alpha_1 - \beta m_1)\delta(\alpha_2 - \beta m_2) \ldots \delta(\alpha_{n-1} - \beta m_{n-1}), \tag{2.74}$$

donde $m_{n-1} = m - (m_1 + m_2 + \cdots + m_{n-2})$. Considerando la serie binomial y el teorema del multinomio es posible demostrar que cuando se cumple la condición de convergencia

$$r_1^\beta + r_2^\beta + \ldots + r_n^\beta < 2\, r_n^\beta, \tag{2.75}$$

se obtiene la siguiente forma para la función de discrepancia

$$\Psi = (r_1^\beta + r_2^\beta + \ldots + r_n^\beta)^{1/\beta}, \tag{2.76}$$

la cual coincide con la β-norma de un espacio vectorial de dimensión finita n.

Finalmente, nótese que el criterio de los mínimos cuadrados corresponde al caso especial $\beta = 2$.

Capítulo 3

Algoritmos Precursores: Métodos del Máximo Descenso y de Gauss-Newton

El objetivo del campo de la *optimización* consiste en el desarrollo de algoritmos para evaluar el punto $\boldsymbol{b}_{min}$ que minimice (o maximice) cierta función suave $\Phi : \mathbf{R}^k \to \mathbf{R}$ en un subconjunto dado $\mathcal{A}$ de su dominio, i.e.,

$$\Phi(\boldsymbol{b}_{min}) = \min_{\boldsymbol{b} \in \mathcal{A} \subset \mathbf{R}^k} \Phi(\boldsymbol{b}).$$

Un grupo especial de estos algoritmos lo constituyen aquellos diseñados para minimizar la función particular

$$\Phi(\boldsymbol{b}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{b}) = \|\boldsymbol{r}\|^2,$$

estudiada en el capítulo anterior.

En el presente capítulo se analizan dos de estos algoritmos, conocidos como: 'del máximo descenso' y 'Gauss-Newton' respectivamente. Si bien ambos son de naturaleza elemental y fueron concebidos en una etapa inicial, su importancia radica en que han sido precursores de algoritmos más sofisticados y de mejor desempeño, tales como el de 'Levenberg-Marquardt' tratado en el siguiente capítulo.

3.1. Introducción

Todos los algoritmos de minimización requieren como punto de partida un tanteo inicial del mínimo, denotado por $\boldsymbol{b}^{(0)}$, a partir del cual calculan, mediante un proceso iterativo, una secuencia ($\boldsymbol{b}^{(1)}$, $\boldsymbol{b}^{(2)}$, ..., $\boldsymbol{b}^{(s)}$, ...) de *puntos tentativos* que mejoran progresivamente el tanteo inicial y convergen en el mínimo $\boldsymbol{b}_{min}$.

Para calcular un nuevo punto tentativo $\boldsymbol{b}^{(s+1)}$, estos algoritmos utilizan información sobre Φ, y sus derivadas, en el punto tentativo actual $\boldsymbol{b}^{(s)}$ y, eventualmente, en los anteriores (..., $\boldsymbol{b}^{(s-2)}$, $\boldsymbol{b}^{(s-1)}$).

La condición fundamental a cumplir en este caso es que $\Phi(\boldsymbol{b}^{(s+1)}) < \Phi(\boldsymbol{b}^{(s)})$, aunque existen excepciones como en el caso de los llamados algoritmos no monótonos, en los que la condición es que la función decrezca luego de un cierto número p de iteraciones, $\Phi(\boldsymbol{b}^{(s+p)}) < \Phi(\boldsymbol{b}^{(s)})$.

El proceso de iteración finaliza cuando se obtiene un punto cuya distancia al punto anterior es menor a cierto umbral preestablecido, o cuando la diferencia entre los valores de Φ en los puntos anterior y actual es igualmente menor a cierto umbral. La rapidez con la que el algoritmo arriba a dicho punto final depende no solo de la eficiencia del mismo, reflejada en su razón de convergencia como medida de su velocidad en alcanzar el mínimo, sino también de qué tan cerca o lejos del mínimo esté el tanteo inicial.

Existen dos estrategias fundamentales para construir procedimientos iterativos que busquen el mínimo: de búsqueda en línea (line search) y de región de confianza (trust region). Cronológicamente, los algoritmos de búsqueda en línea, como los del máximo descenso y de Gauss-Newton, se desarrollaron antes que los algoritmos de región de confianza, los cuales tienen su génesis en el algoritmo de Levenberg-Marquardt.

En la estrategia de búsqueda en línea se elige en cada iteración un vector dirección unitario $\hat{\boldsymbol{\delta}}^{(s)}$, con el que se construye un vector de paso que conduce al nuevo punto tentativo $\boldsymbol{b}^{(s+1)}$ a partir del actual $\boldsymbol{b}^{(s)}$. La longitud α_s del paso se obtiene resolviendo el problema de minimización unidimensional

$$\Phi(\boldsymbol{b}^{(s)} + \alpha_s \hat{\boldsymbol{\delta}}^{(s)}) = \min_{\alpha > 0} \Phi(\boldsymbol{b}^{(s)} + \alpha \hat{\boldsymbol{\delta}}^{(s)}). \tag{3.1}$$

En la práctica, sin embargo, los algoritmos de búsqueda en línea no optan por

resolver este problema, por ser este un proceso de alto costo computacional, sino que generan un número limitado de longitudes de paso de prueba hasta conseguir una que se aproxime suficientemente a la longitud de paso óptimo que resulta de (3.1).

En la estrategia de región de confianza se extrae en cada iteración información de Φ en el punto tentativo actual $\boldsymbol{b}^{(s)}$ para crear, alrededor del mismo, un modelo m_s aproximado que reemplace a Φ en la búsqueda del mínimo. La vecindad V_s en la cual m_s constituye una buena aproximación de Φ se conoce como región de confianza. El paso $\boldsymbol{\delta}^{(s)}$ que conduce al nuevo punto tentativo $\boldsymbol{b}^{(s+1)}$ se calcula resolviendo aproximadamente el problema de minimización

$$m_s(\boldsymbol{b}^{(s)} + \boldsymbol{\delta}^{(s)}) = \min_{\boldsymbol{\delta} \in V_s} m_s(\boldsymbol{b}^{(s)} + \boldsymbol{\delta}). \tag{3.2}$$

Si la iteración no produce una disminución aceptable de Φ, es necesario encoger la región de confianza V_s y volver a resolver (3.2). Por lo general, V_s es una bola definida por $\|\boldsymbol{\delta}\| \leq \Delta^{(s)}$, donde al escalar $\Delta^{(s)}$ se le denomina radio de la región de confianza.

En la mayoría de los casos, el modelo m_s lo constituye la función cuadrática correspondiente a la serie de Taylor truncada de Φ alrededor de $\boldsymbol{b}^{(s)}$

$$m_s(\boldsymbol{b}^{(s)} + \boldsymbol{\delta}) = \Phi^{(s)} + (\boldsymbol{\nabla}^T \Phi^{(s)})\, \boldsymbol{\delta} + \frac{1}{2}\, \boldsymbol{\delta}^T\, B^{(s)}\, \boldsymbol{\delta},$$

donde para abreviar: $\Phi^{(s)} = \Phi(\boldsymbol{b}^{(s)})$, $\boldsymbol{\nabla}\Phi^{(s)} = \boldsymbol{\nabla}\Phi(\boldsymbol{b}^{(s)})$, y $B^{(s)}$ es el Hessiano en $\boldsymbol{b}^{(s)}$

$$B^{(s)} = \boldsymbol{\nabla}(\boldsymbol{\nabla}^T \Phi^{(s)}) = \begin{bmatrix} \dfrac{\partial^2 \Phi(\boldsymbol{b}^{(s)})}{\partial b_1^2} & \dfrac{\partial^2 \Phi(\boldsymbol{b}^{(s)})}{\partial b_1 \partial b_2} & \cdots & \dfrac{\partial^2 \Phi(\boldsymbol{b}^{(s)})}{\partial b_1 \partial b_k} \\ \dfrac{\partial^2 \Phi(\boldsymbol{b}^{(s)})}{\partial b_1 \partial b_2} & \dfrac{\partial^2 \Phi(\boldsymbol{b}^{(s)})}{\partial b_2^2} & \cdots & \dfrac{\partial^2 \Phi(\boldsymbol{b}^{(s)})}{\partial b_2 \partial b_k} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 \Phi(\boldsymbol{b}^{(s)})}{\partial b_1 \partial b_k} & \dfrac{\partial^2 \Phi(\boldsymbol{b}^{(s)})}{\partial b_2 \partial b_k} & \cdots & \dfrac{\partial^2 \Phi(\boldsymbol{b}^{(s)})}{\partial b_k^2} \end{bmatrix},$$

o alguna aproximación del mismo.

Antes de pasar a ver en detalle los dos algoritmos de minimización mencionados, es conveniente resaltar la importancia de la expansión en series de

Taylor como herramienta matemática fundamental empleada en la minimización de funciones continuamente diferenciables. Para el caso particular de interés, correspondiente a funciones de más de una variable independiente a ser expandidas hasta el término de segundo grado, el teorema de Taylor toma la forma enunciada a continuación.

Teorema de Taylor. *Sea* $\Phi(\boldsymbol{b}) : \mathbf{R}^k \to \mathbf{R}$ *una función continua y doblemente diferenciable en todos los puntos* $\boldsymbol{b}$ *de la bola* $\| \boldsymbol{b} - \boldsymbol{b_0} \| < \delta_0$. *Si se tiene además que* $\boldsymbol{\delta} \in \mathbf{R}^k$ *con* $\|\boldsymbol{\delta}\| = \delta < \delta_0$, *entonces existe un* $t \in [0, 1]$ *para el cual se cumple*

$$\Phi(\boldsymbol{b_0} + \boldsymbol{\delta}) = \Phi(\boldsymbol{b_0}) + (\boldsymbol{\nabla}^T \Phi(\boldsymbol{b_0}))\, \boldsymbol{\delta} + \frac{1}{2}\, \boldsymbol{\delta}^T\, \boldsymbol{\nabla}(\boldsymbol{\nabla}^T \Phi(\boldsymbol{b_0} + t\, \boldsymbol{\delta}))\, \boldsymbol{\delta}. \tag{3.3}$$

La demostración de esta versión multivariable del teorema se puede llevar a cabo generalizando el procedimiento seguido en la demostración del caso univariable que aparece en diversos textos de cálculo. El lector interesado puede consultar en particular el artículo de Pringsheim [27], que ofrece también un recuento histórico del desarrollo del tema.

3.2. Método del Máximo Descenso

Una manera sencilla de visualizar el método del máximo descenso, conocido también como 'steepest descent', es mediante un símil con la trayectoria que sigue una esfera que rueda sobre una superficie irregular por acción única de la gravedad hasta que alcanza un *valle*, o mínimo relativo, en el que queda encajonada y detenida.

En la presente sección se analizan en detalle las tres etapas que componen este método: la elección de la dirección de minimización, el reescalamiento de la función a minimizar, y el cálculo de la longitud del paso en cada iteración. La sección concluye con un análisis de la convergencia del método.

3.2.1. Elección de la Dirección de Minimización

En cada nueva iteración se busca elegir la dirección $\hat{\boldsymbol{\delta}}$ en la cual la función Φ a minimizar decrezca de forma más abrupta a partir del punto tentativo actual, i.e., la dirección de máximo descenso. Esta dirección, evidentemente,

está dada por el negativo de la gradiente en dicho punto, que apunta hacia donde la derivada o *razón de cambio* de Φ adquiere su menor valor.

En particular, para el caso de la $(s+1)$-ava iteración, la búsqueda del nuevo punto tentativo $\boldsymbol{b}^{(s+1)}$ se lleva a cabo a partir del actual $\boldsymbol{b}^{(s)}$, siguiendo la dirección

$$\hat{\boldsymbol{\delta}}^{(s)} = -\frac{\boldsymbol{\nabla}\Phi^{(s)}}{\|\boldsymbol{\nabla}\Phi^{(s)}\|}.$$

Si bien lo anteriormente expresado fundamenta la elección del negativo de la gradiente como dirección de minimización a ser usada por el método, es conveniente analizar, con cierto detalle, las limitaciones de tal elección.

Para tal fin, y por simplicidad, se considera el caso de Φ definida en $\mathbf{R}^2$. Definiendo una familia de circunferencias concéntricas a cierto punto $\boldsymbol{b}$ en el dominio de Φ, e introduciendo luego coordenadas polares con origen en $\boldsymbol{b}$, se tiene que el valor que toma Φ en los puntos $r\hat{\boldsymbol{e}}_\theta = r(\cos\theta,\, \operatorname{sen}\theta)$ que conforman la circunferencia de radio r, es función únicamente de θ y, de acuerdo a lo establecido por el teorema de Taylor en (3.3), está dado por

$$f(\theta) = \Phi(\boldsymbol{b}+r\hat{\boldsymbol{e}}_\theta) = \Phi(\boldsymbol{b}) + (\boldsymbol{\nabla}^T\Phi(\boldsymbol{b}))\,\hat{\boldsymbol{e}}_\theta\, r + \left[\frac{1}{2}\,\hat{\boldsymbol{e}}_\theta^{\,T}\,\boldsymbol{\nabla}(\boldsymbol{\nabla}^T\Phi(\boldsymbol{b}+t\,r\hat{\boldsymbol{e}}_\theta))\,\hat{\boldsymbol{e}}_\theta\right] r^2. \tag{3.4}$$

El valor θ_r correspondiente al punto de la circunferencia en el que Φ toma el menor valor, se obtiene como solución de $f'(\theta) = 0$, siempre y cuando se cumpla la condición $f''(\theta_r) > 0$. Por lo tanto, igualando a cero la derivada de (3.4) se llega a la expresión que define a θ_r

$$\operatorname{sin}(\theta_r - \theta_0) = \frac{g(r,\, \theta_r)}{\|\boldsymbol{\nabla}\Phi(\boldsymbol{b})\|}\, r, \tag{3.5}$$

donde se ha considerado que $\boldsymbol{\nabla}\Phi(\boldsymbol{b}) = \|\boldsymbol{\nabla}\Phi(\boldsymbol{b})\|(\cos\theta_0,\, \operatorname{sin}\theta_0)$, con la condición adicional que $\|\boldsymbol{\nabla}\Phi(\boldsymbol{b})\| \neq 0$, y a $g(r,\, \theta_r)$ como la derivada parcial respecto a θ de la expresión entre corchetes en (3.4).

Además, por la condición de doble diferenciabilidad impuesta a Φ, se puede afirmar que tanto $\|\boldsymbol{\nabla}\Phi(\boldsymbol{b})\|$ como $g(r,\, \theta_r)$ poseen valores finitos. Esto trae como consecuencia que al aplicar en (3.5) el límite cuando $r \to 0$, el lado derecho se haga cero, con lo cual se tiene

$$\lim_{r\to 0} \theta_r = \theta_0 + \pi. \tag{3.6}$$

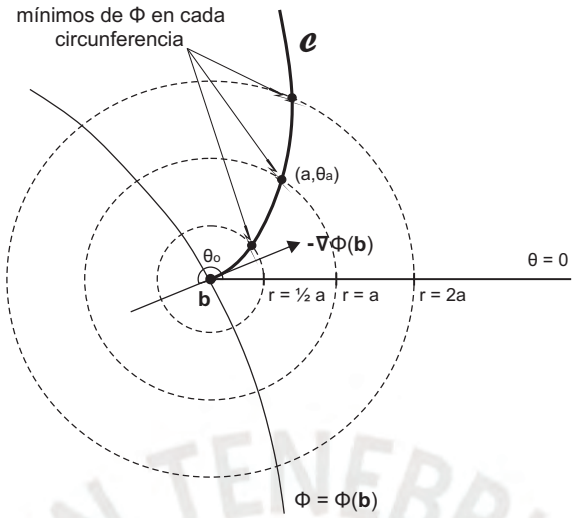


Figura 3.1: Curva $\mathcal{C}$ que une los puntos mínimos de Φ en cada circunferencia.

Este resultado indica que el mínimo sobre la circunferencia límite, cuyo radio tiende a cero, se encuentra en la dirección del negativo de la gradiente.

En la figura 3.1 se muestra la curva $\mathcal{C}$ formada por los puntos que minimizan Φ en cada circunferencia, la cual es tangente a $-\boldsymbol{\nabla}\Phi(\boldsymbol{b})$ en el punto $\boldsymbol{b}$. En ella se puede también observar que conforme las circunferencias se alejan de $\boldsymbol{b}$, los mínimos sobre estas se apartan en general de la dirección dada por $-\boldsymbol{\nabla}\Phi(\boldsymbol{b})$. Sin embargo, es posible advertir que en una pequeña vecindad de $\boldsymbol{b}$, delimitada por una circunferencia de radio menor a cierto umbral, es posible emplear la dirección del negativo de la gradiente como dirección de máximo descenso con un error de aproximación despreciable. En el caso límite $r \to 0$ el negativo de la gradiente es efectivamente, como ya se demostró, la dirección de máximo descenso.

Como comentario final es oportuno señalar que al recorrer la curva $\mathcal{C}$ partiendo de $\boldsymbol{b}$, el valor de Φ decrece hasta alcanzar un punto, de existir uno, en el que la tendencia decreciente se revierte y Φ empieza a crecer. Es posible demostrar que este punto de cambio es precisamente el punto mínimo $\boldsymbol{b}_{min}$ requerido. Otra observación, no evidente a simple vista, es que la curva $\mathcal{C}$ es en general distinta a la curva trazada desde $\boldsymbol{b}$ siguiendo progresivamente la dirección del negativo de la gradiente en cada uno de sus puntos, i.e., a la curva que parte de $\boldsymbol{b}$ y es ortogonal a las curvas o superficies de valor constante de Φ, denominadas *isolíneas* o *isosuperficies* dependiendo de si el dominio de Φ está en $\mathbf{R}^2$ o $\mathbf{R}^3$ respectivamente. Cabe mencionar que la

gradiente en cada punto de $\mathcal{C}$ está en la dirección del radio que va del centro $\boldsymbol{b}$ (de la familia de circunferencias) a dicho punto.

3.2.2. Reescalamiento de la Función a Minimizar

Uno de los factores que influye en el desempeño global del método del máximo descenso lo constituye la topología de la función Φ a minimizar, que puede en casos extremos degradar la eficiencia del método hasta el punto de hacerlo improductivo. Por este motivo es conveniente previamente 'mejorar' la topología de Φ mediante un proceso de reescalamiento.

Para ilustrar esta influencia, considérese una función Φ, definida por simplicidad en $\mathbf{R}^2$, con isolíneas que formen una familia de circunferencias concéntricas a su punto mínimo, e.g., $\Phi = x^2 + y^2$. Esta topología de Φ permite alcanzar el mínimo desde cualquier punto siguiendo la recta en la dirección del negativo de su gradiente. Desde el punto de vista del método, se dice que esta topología es *perfecta*, pues se obtiene la máxima eficiencia al alcanzar el mínimo en un solo paso. Generalizando, es hasta cierto punto válido afirmar que mientras más difiera una topología del caso perfecto, más pasos serán necesarios para alcanzar el mínimo de su respectiva función.

En particular, esta afirmación es correcta en la vecindad de un punto mínimo, donde la topología de cualquier función es aproximadamente la de su serie de Taylor truncada al segundo orden, i.e., una familia de elipses concéntricas al mínimo, las cuales se alejan más del caso perfecto, compuesto por circunferencias, conforme la razón entre sus ejes mayor y menor aumente.

La figura 3.2 muestra las isolíneas elípticas de cierta función Φ en una región cercana al mínimo $\boldsymbol{b}_{min}$. Para esta topología se cumple que desde cualquier punto, excepto aquellos sobre los ejes de las elipses, la dirección del negativo de la gradiente no conduce directamente a $\boldsymbol{b}_{min}$. Para analizar la relación entre la forma de las elipses y el número de pasos requeridos para alcanzar $\boldsymbol{b}_{min}$, es necesario tener en cuenta la forma en que el método construye cada paso.

En la versión convencional del método, cada paso parte de un punto y recorre la dirección del negativo de su gradiente hasta alcanzar un mínimo. En este mínimo, que es el punto de partida del siguiente paso, la isolínea y la recta recorrida son tangentes. Por consiguiente, los sucesivos pasos que lleva a cabo el método son ortogonales entre sí.

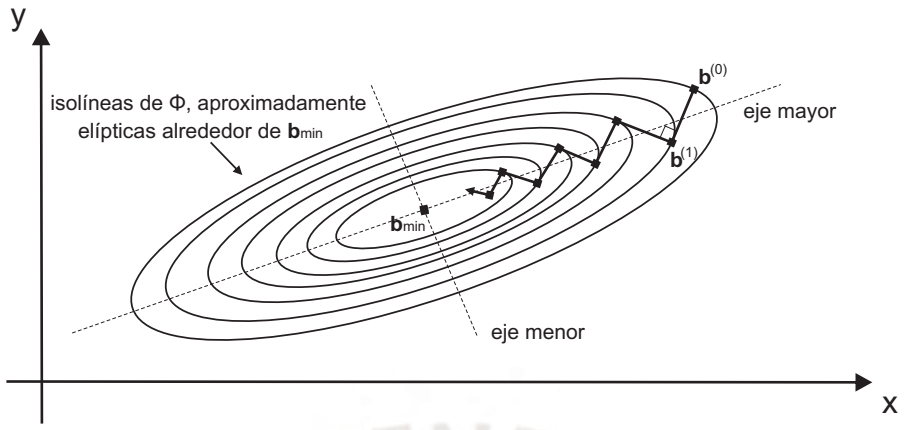


Figura 3.2: Pasos que lleva a cabo el método en una vecindad cercana del mínimo $\boldsymbol{b}_{min}$.

Volviendo al caso mostrado en la figura, donde las elipses son alargadas y el primer paso, que va de $\boldsymbol{b}^{(0)}$ a $\boldsymbol{b}^{(1)}$, forma un ángulo cercano a 45° con el eje mayor, se observa que el método alcanza el mínimo $\boldsymbol{b}_{min}$ luego de varios pequeños pasos decrecientes en zigzag. Nótese que si estas isolíneas fuesen aun más alargadas y angostas, i.e., más alejadas de la topología perfecta, los pasos en general serían más cortos y se necesitaría un mayor número de ellos para alcanzar el mínimo.

Tres de las técnicas de reescalamiento más útiles para superar las dificultades descritas se reseñan a continuación.

Una primera técnica, bastante efectiva cuando el punto tentativo actual $\boldsymbol{b}^{(s)}$ ha alcanzado suficiente cercanía al mínimo $\boldsymbol{b}_{min}$, consiste en reemplazar Φ por su serie de Taylor truncada de segundo grado alrededor de $\boldsymbol{b}^{(s)}$, denotada por Φ^*, para luego elegir, convenientemente, como dirección de minimización a aquella que conduce al centro de las isolíneas elípticas de Φ^*.

La figura 3.3 muestra precisamente el caso en que $\boldsymbol{b}_{min}$ se encuentra en la región, alrededor de $\boldsymbol{b}^{(s)}$, donde Φ^* constituye una buena aproximación de Φ. Por consiguiente, la separación entre $\boldsymbol{b}_{min}$ y el mínimo de Φ^*, que coincide con el centro $\boldsymbol{b}_0 = (x_0, y_0)$ de sus isolíneas elípticas, es razonablemente pequeña y tiende, además, a disminuir conforme $\boldsymbol{b}^{(s)}$ se acerca más a $\boldsymbol{b}_{min}$.

En consecuencia, cuando $\boldsymbol{b}^{(s)}$ está suficientemente cerca a $\boldsymbol{b}_{min}$, dar un paso en la dirección $\hat{\boldsymbol{\delta}}^{(s)}$, que conecta $\boldsymbol{b}^{(s)}$ con $\boldsymbol{b}_0$, representa una mejor opción para el objetivo de alcanzar $\boldsymbol{b}_{min}$, que darlo en la dirección del máximo descenso $-\boldsymbol{\nabla}\Phi^{(s)}$, tal como se aprecia en la figura.

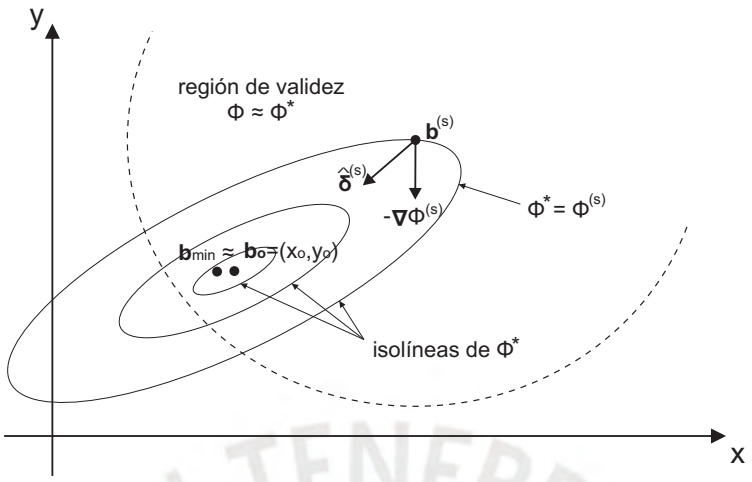


Figura 3.3: Dirección de paso modificada $\hat{\boldsymbol{\delta}}^{(s)}$, que apunta a $\boldsymbol{b}_0 \approx \boldsymbol{b}_{min}$ y reemplaza la dirección original $-\boldsymbol{\nabla}\Phi^{(s)}$.

Como se mencionó, el punto $\boldsymbol{b}_0$ es el centro de las isolíneas elípticas de la función

$$\Phi^*(\boldsymbol{b}) = \Phi^{(s)} + \boldsymbol{\nabla}\Phi^{(s)}(\boldsymbol{b} - \boldsymbol{b}^{(s)}) + \frac{1}{2}(\boldsymbol{b} - \boldsymbol{b}^{(s)})\boldsymbol{\nabla}(\boldsymbol{\nabla}^T\Phi^{(s)})(\boldsymbol{b} - \boldsymbol{b}^{(s)}), \quad (3.7)$$

que es la versión aproximada de Φ alrededor del punto tentativo actual $\boldsymbol{b}^{(s)}$. De esta familia de elipses, aquella que pasa por $\boldsymbol{b}^{(s)}$ está dada por la ecuación

$$\boldsymbol{\nabla}\Phi^{(s)}(\boldsymbol{b} - \boldsymbol{b}^{(s)}) + \frac{1}{2}(\boldsymbol{b} - \boldsymbol{b}^{(s)})\boldsymbol{\nabla}(\boldsymbol{\nabla}^T\Phi^{(s)})(\boldsymbol{b} - \boldsymbol{b}^{(s)}) = 0, \quad (3.8a)$$

que en coordenadas cartesianas, $\boldsymbol{b} = (x, y)$, toma la forma

$$\frac{1}{2}Ax^2 + \frac{1}{2}By^2 + Cxy + Dx + Ey + F = 0, \quad (3.8b)$$

donde

$$A = \frac{\partial^2\Phi^{(s)}}{\partial x^2}, \qquad B = \frac{\partial^2\Phi^{(s)}}{\partial y^2}, \qquad C = \frac{\partial^2\Phi^{(s)}}{\partial x \partial y},$$
$$D = \frac{\partial\Phi^{(s)}}{\partial x} - (A, C)\cdot\boldsymbol{b}^{(s)} \qquad \text{y} \qquad E = \frac{\partial\Phi^{(s)}}{\partial y} - (C, B)\cdot\boldsymbol{b}^{(s)}.$$

El centro de esta elipse, que lo es igualmente de toda la familia de elipses, es el único punto respecto al cual (3.8) es simétrica. Resolviendo la ecuación

correspondiente a esta condición se obtienen las coordenadas del centro

$$x_0 = \frac{CE - BD}{AB\,(1 - \frac{C^2}{AB})}, \qquad y_0 = \frac{CD - AE}{AB\,(1 - \frac{C^2}{AB})}, \tag{3.9}$$

cuyos valores están bien definidos, ya que $(1 - \frac{C^2}{AB}) = 0$ se da únicamente cuando (3.8) representa la ecuación de una recta o parábola y no, como en este caso, la de una elipse, para la cual se cumple siempre que $(1 - \frac{C^2}{AB}) > 0$.

Conocidas las coordenadas de $\boldsymbol{b}_0$, es posible definir un vector $\boldsymbol{\delta}^{(s)}$ arbitrario en la dirección que va de $\boldsymbol{b}^{(s)}$ a $\boldsymbol{b}_0$, el cual por simplicidad se elige

$$\boldsymbol{\delta}^{(s)} = (1 - \frac{C^2}{AB})(\boldsymbol{b}_0 - \boldsymbol{b}^{(s)}), \tag{3.10}$$

con respectivas componentes

$$\begin{aligned} \delta_x^{(s)} &= (\frac{1}{A}, -\frac{C}{AB}) \cdot (-\boldsymbol{\nabla}\Phi^{(s)}) = -\frac{1}{A}\frac{\partial \Phi^{(s)}}{\partial x} + \frac{C}{AB}\frac{\partial \Phi^{(s)}}{\partial y} \\ \delta_y^{(s)} &= (-\frac{C}{AB}, \frac{1}{B}) \cdot (-\boldsymbol{\nabla}\Phi^{(s)}) = -\frac{1}{B}\frac{\partial \Phi^{(s)}}{\partial y} + \frac{C}{AB}\frac{\partial \Phi^{(s)}}{\partial x}. \end{aligned} \tag{3.11}$$

Según este resultado, $\boldsymbol{\delta}^{(s)}$ se puede interpretar como una modificación de la dirección original $-\boldsymbol{\nabla}\Phi^{(s)}$, en un grado directamente relacionado a C.

Este factor C está asociado a la medida del alargamiento que posee la elipse, así como al ángulo de rotación que presenta la misma con respecto a la orientación de referencia en la que el eje mayor es horizontal. Explícitamente, C tiene la forma

$$C = K\,(1 - r)\sin(2\alpha),$$

donde r es la razón entre los ejes menor y mayor de la elipse, α el ángulo de rotación de la elipse, y K una constante de escalamiento positiva. De esta expresión se desprende que para valores de r cercanos a uno, correspondientes a elipses con formas cuasi circulares, la modificación de $-\boldsymbol{\nabla}\Phi^{(s)}$ es mínima y $\boldsymbol{\delta}^{(s)}$ apunta prácticamente en la dirección original; lo cual significa que la aplicación de este procedimiento tuvo un efecto despreciable.

Sin embargo, esta técnica de reescalamiento sí es relevante cuando las elipses son muy alargadas y, por consiguiente, r es pequeño. Asumiendo este caso y con el propósito de ahorrar operaciones, es posible omitir el cálculo de

C (consistente en aproximar numéricamente $\frac{\partial^2 \Phi^{(s)}}{\partial x \partial y}$) y reemplazarlo por un valor aproximado que solo dependa de los coeficientes A y B. Dicha aproximación dependerá de si el ángulo de rotación de la elipse es tal que origine que sus ejes se desvíen un ángulo pequeño respecto a los ejes coordenados, o de si este hace que los ejes de la elipse formen un ángulo cercano a los 45° con los ejes coordenados.

De un análisis más completo de la ecuación general de una elipse rotada se puede deducir que cuando $\frac{A}{B} + \frac{B}{A} > 4$, el ángulo de rotación corresponde al primer tipo descrito y $C \approx 0$, con lo cual

$$\delta_x^{(s)} = -\frac{1}{A}\frac{\partial \Phi^{(s)}}{\partial x}, \qquad \delta_y^{(s)} = -\frac{1}{B}\frac{\partial \Phi^{(s)}}{\partial y}, \tag{3.12a}$$

mientras que cuando $\frac{A}{B} + \frac{B}{A} \leq 4$, el efecto de la rotación es apreciable y $C \approx \mp\frac{A+B}{3}$, de modo que

$$\delta_x^{(s)} = -\frac{1}{A}\frac{\partial \Phi^{(s)}}{\partial x} \mp \frac{1}{3}\left(\frac{1}{A} + \frac{1}{B}\right)\frac{\partial \Phi^{(s)}}{\partial y}, \qquad \delta_y^{(s)} = -\frac{1}{B}\frac{\partial \Phi^{(s)}}{\partial y} \mp \frac{1}{3}\left(\frac{1}{A} + \frac{1}{B}\right)\frac{\partial \Phi^{(s)}}{\partial x}, \tag{3.12b}$$

donde los signos $-$ y $+$ corresponden a ángulos de rotación agudos y obtusos, respectivamente.

Cabe indicar que estos resultados, obtenidos para el caso bidimensional, pueden ser generalizados a n dimensiones.

Un segundo procedimiento [22] consiste en mapear el espacio k-dimensional de los vectores $\boldsymbol{b} = (b_1, b_2, \ldots, b_k)$ que conforman el dominio de Φ hacia un nuevo espacio $\boldsymbol{b}^*$ mediante la transformación

$$\left.\begin{array}{rcl} b_1^* & = & \arctan(b_1) \\ b_2^* & = & \arctan(b_2) \\ & \vdots & \\ b_k^* & = & \arctan(b_k) \end{array}\right\}, \tag{3.13}$$

que convierte las isosuperficies de Φ en el espacio original, en superficies aproximadamente esféricas en el nuevo espacio, mediante un proceso de acercamiento al origen, o compresión, de las partes distantes de las isosuperficies, responsables de alargarlas y apartarlas del caso esférico.

Al tomar $\Phi(\boldsymbol{b})$ la forma $\Phi^*(\boldsymbol{b}^*)$ en el nuevo espacio, se tiene que las deri-

vadas parciales en ambos espacios cumplen la relación

$$\frac{\partial \Phi}{\partial b_j} = \frac{\partial \Phi^*}{\partial b_j^*} \frac{\partial b_j^*}{\partial b_j}, \tag{3.14}$$

y puesto que de (3.13) resulta

$$\frac{\partial b_j^*}{\partial b_j} = \frac{1}{1 + b_j^2}\,, \tag{3.15}$$

se obtiene finalmente la expresión de la j-ésima derivada parcial de Φ en el espacio transformado a partir de su similar en el espacio original,

$$\frac{\partial \Phi^*}{\partial b_j^*} = (1 + b_j^2)\, \frac{\partial \Phi}{\partial b_j}. \tag{3.16}$$

Dado que la gradiente en este nuevo espacio está, en general, mejor dirigida hacia el mínimo que la gradiente en el espacio original, la presente técnica propone una dirección de minimización modificada que explota esta ventaja. Como resultado, las componentes del vector de paso $\hat{\boldsymbol{\delta}}^{(s)}$ para la s-ésima iteración toman la forma

$$\hat{\delta}_j^{(s)} = \frac{\left(1 + (b_j^{(s)})^2\right) \dfrac{\partial \Phi^{(s)}}{\partial b_j}}{\left(\displaystyle\sum_{h=1}^{k} \left[\left(1 + (b_h^{(s)})^2\right) \frac{\partial \Phi^{(s)}}{\partial b_h}\right]^2\right)^{1/2}}. \tag{3.17}$$

Una tercera técnica, que toma en cuenta el comportamiento estadístico de los datos medidos que el modelo busca reproducir, es tratada en el siguiente capítulo.

3.2.3. Cálculo de la Longitud del Paso

La última etapa del método consiste en calcular, para cada iteración, la longitud α_s del paso en la dirección de minimización $\hat{\boldsymbol{\delta}}^{(s)}$ que conduce del punto tentativo actual $\boldsymbol{b}^{(s)}$ al siguiente $\boldsymbol{b}^{(s+1)}$ según la fórmula

$$\boldsymbol{b}^{(s+1)} = \boldsymbol{b}^{(s)} + \alpha_s\, \hat{\boldsymbol{\delta}}^{(s)}. \tag{3.18}$$

Como ya se mencionó al inicio del capítulo, el beneficio de resolver de manera exacta (3.1) para obtener el valor ideal de α_s no justifica su alto costo computacional sino que degrada la eficiencia del método, por lo que es preferible la opción de calcular α_s mediante un procedimiento de tanteo y verificación [2] que ha probado tener un buen desempeño.

Este procedimiento toma en cuenta el hecho de que cuando la longitud del paso es mucho menor que la ideal, el vector de paso de la siguiente iteración es aproximadamente colineal al de la iteración actual, mientras que cuando la longitud del paso es mayor que la ideal, los ángulos entre ambos vectores de paso son mayores a 90°. La corrección de la longitud del paso se lleva a cabo entonces de acuerdo a este esquema, aumentándola o disminuyéndola según sea el caso. Evidentemente, en el caso que la longitud del paso sea la ideal, el ángulo entre dichos vectores de paso es exactamente 90°.

Para la $s+1$-ésima iteración, el procedimiento de cálculo de la longitud del paso α_s es el siguiente:

1. Calcular $\Phi^{(s)}$ y $\hat{\boldsymbol{\delta}}^{(s)}$.

2. Verificar que $\Phi^{(s)} < \Phi^{(s-1)}$. Si esto no se cumple, significa que el valor de α_{s-1} usado en la iteración anterior fue muy grande. En ese caso dividir α_{s-1} entre (por ejemplo) 4 y regresar al punto 1.

3. Calcular el coseno del ángulo θ que forma el vector de paso actual con el de la iteración anterior mediante: $\cos\theta = \hat{\boldsymbol{\delta}}^{(s)} \cdot \hat{\boldsymbol{\delta}}^{(s-1)}$.

4. Verificar que $\cos\theta \geq 0$ para evitar oscilaciones indeseables en el camino hacia el mínimo. Si $\cos\theta < 0$, entonces el valor de α_{s-1} fue muy grande. En ese caso dividir α_{s-1} entre (por ejemplo) 4 y regresar al punto 1.

5. Calcular por tanteo el valor de α_s en función al valor de $\cos\theta$. La estrategia consiste en asignarle a α_s un valor mayor al de α_{s-1} si $\cos\theta \approx 1$ y, correspondientemente, un valor menor si $\cos\theta \approx 0$. La siguiente fórmula a demostrado ser útil para este fin:

$$\alpha_s = (d_1 + d_2 \cos^4\theta)\, \alpha_{s-1},$$

donde d_1 y d_2 son parámetros arbitrarios que deben cumplir con $0 < d_1 < 1$

y $(1 - d_1) < d_2 \leq 1$. En particular, los valores $d_1 = 0{,}5$ y $d_2 = 1$ producen resultados satisfactorios.

6. Calcular a partir de (3.18) el nuevo punto tentativo $\boldsymbol{b}^{(s+1)}$ y ejecutar la siguiente iteración partiendo del punto 1.

Una de las principales ventajas de este procedimiento radica en que la longitud del paso en cada iteración se ajusta automáticamente a la topología local de Φ sin que sea necesario resolver un sistema de ecuaciones u otro tipo de operaciones complejas que no sean el simple cálculo de $\Phi^{(s)}$ y $\hat{\boldsymbol{\delta}}^{(s)}$.

Otra ventaja importante de este procedimiento es que garantiza, tal como se muestra a continuación, la convergencia al mínimo del proceso iterativo que compone el método del máximo descenso.

3.2.4. Demostración de la Convergencia del Método

La demostración de la convergencia del método del máximo descenso que aquí se presenta está basada en uno de los primeros trabajos que abordaron este tema [5]. Dado que la demostración expuesta en dicho trabajo es bastante breve y se fundamenta en una serie de proposiciones presentadas sin mayor detalle ni prueba, ha sido necesario presentar aquí la deducción de cada una de dichas proposiciones para subsanar esa carencia y proporcionar un análisis más completo.

Si bien esta demostración de convergencia es válida cuando en cada iteración el vector de paso se toma en la dirección del negativo de la gradiente o máximo descenso, es posible (con ligeras modificaciones) extender su validez a los casos en que el vector de paso es una versión modificada de la dirección del máximo descenso producto de alguna técnica de reescalamiento.

Considérese una región S en cuyo interior y contorno se asume Φ con primeras derivadas parciales continuas. Asimismo, sea C el camino segmentado que parte de algún punto tentativo inicial $\boldsymbol{b}^{(0)}$ al interior de S y sigue por el segmento de recta que lo une con el siguiente punto tentativo $\boldsymbol{b}^{(1)}$ hasta alcanzar primero ya sea el borde de S o el punto $\boldsymbol{b}^{(1)}$, y así sucesivamente con los puntos tentativos siguientes. Se asume además que a lo largo del camino C, producido por la secuencia de los $\boldsymbol{b}^{(s)}$, Φ es monótona decreciente.

Bajo este escenario existen tres posibilidades:

(1) el camino C termina en algún punto del borde de S,

(2) el camino C termina en un punto con gradiente cero, i.e., en un punto estacionario de Φ,

(3) el proceso continúa indefinidamente.

La primera posibilidad queda excluída si se elije $\boldsymbol{b}^{(0)}$ de manera que el valor de Φ en dicho punto, denotado por $\Phi^{(0)}$, sea menor que el valor de Φ en cualquier punto del borde de S. Para la segunda posibilidad, la prueba de convergencia es trivial. El caso relevante es por lo tanto el tercero, en que el proceso continúa indefinidamente.

De la condición previa sobre la diferenciabilidad de Φ y del supuesto que esta en el punto $\boldsymbol{b}^{(0)}$ interior a S toma un valor menor a los del contorno de S, se desprende la existencia de un punto estacionario, $\boldsymbol{b}_{min} \in S$, en el cual Φ adquiere su valor mínimo Φ_{min} en S. Establecido esto, la demostración de convergencia consiste en probar que el método converge a un punto estacionario de Φ, que en general puede corresponder a un mínimo relativo o a un punto de ensilladura (saddle point).

El paso inicial de esta demostración consiste en determinar la convergencia de la secuencia de puntos tentativos $\boldsymbol{b}^{(0)}, \boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \ldots$, que al mismo tiempo genera la serie $\Phi^{(0)}, \Phi^{(1)}, \Phi^{(2)}, \ldots$, la cual por ser monótona decreciente y acotada inferiormente por Φ_{min} posee un valor límite $\Phi_0 = \lim_{s\to\infty} \Phi^{(s)}$. Esto significa que los puntos $\boldsymbol{b}^{(s)}$ 'convergen' hacia la isosuperficie $\Phi(\boldsymbol{b}) = \Phi_0$, lo que implica la existencia de un umbral N_ϵ a partir del cual los puntos $\boldsymbol{b}^{(N_\epsilon)}, \boldsymbol{b}^{(N_\epsilon+1)}, \boldsymbol{b}^{(N_\epsilon+2)}, \ldots$, están contenidos en la región Ω_ϵ limitada por las isosuperficies $S_\epsilon : \Phi(\boldsymbol{b}) = \Phi_0 + \epsilon$ y $S_0 : \Phi(\boldsymbol{b}) = \Phi_0$, donde $\epsilon > 0$ es un parámetro que puede hacerse arbitrariamente pequeño para así conseguir que las isosuperficies S_ϵ y S_0 tiendan a ser paralelas y separadas una distancia d_ϵ infinitamente pequeña.

Afirmar que la secuencia de los $\boldsymbol{b}^{(s)}$ no converge en un punto equivale a afirmar que la serie $\{D_s = \|\boldsymbol{b}^{(s+1)} - \boldsymbol{b}^{(s)}\|\}$ de las distancias entre un punto tentativo y el siguiente no convergen en cero. En este caso siempre será posible encontrar, para un ϵ suficientemente pequeño, un $N'_\epsilon \geq N_\epsilon$ tal que $D_{N'_\epsilon} > d_\epsilon$,

lo que significa que $\boldsymbol{b}^{(N'_\epsilon)}$ y $\boldsymbol{b}^{(N'_\epsilon+1)}$ están al interior de Ω_ϵ y separados una distancia mayor a su grosor d_ϵ.

Sin embargo, el paso que conduce de $\boldsymbol{b}^{(N'_\epsilon)}$ a $\boldsymbol{b}^{(N'_\epsilon+1)}$, según el método del máximo descenso, es ortogonal a las isosuperficies aproximadamente paralelas S_ϵ y S_0 que delimitan Ω. Esto implica necesariamente que $\boldsymbol{b}^{(N'_\epsilon+1)}$ quede fuera de Ω, ya que la longitud del paso $D_{N'_\epsilon}$ es mayor que la distancia d_ϵ entre S_ϵ y S_0. Se observa entonces que asumir que la secuencia de los $\boldsymbol{b}^{(s)}$ no es convergente conlleva a conclusiones que se contradicen entre sí.

Queda así finalmente establecido que la secuencia $\boldsymbol{b}^{(0)}, \boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \ldots$, converge en el punto $\boldsymbol{b}^{(\infty)}$, para el cual se cumple

$$\Phi^{(\infty)} < \Phi^{(s)}, \qquad s = 0, 1, 2, \ldots \tag{3.19}$$

Como paso final de la demostración, resta probar que $\boldsymbol{b}^{(\infty)}$ es un punto estacionario de Φ, i.e., que se cumple la condición $\boldsymbol{\nabla}\Phi^{(\infty)} = \boldsymbol{0}$.

Adóptese la siguiente notación para el módulo y la dirección del negativo de la gradiente respectivamente,

$$h(\boldsymbol{b}) = \Big\|\boldsymbol{\nabla}\Phi(\boldsymbol{b})\Big\|, \qquad \hat{\boldsymbol{\delta}}(\boldsymbol{b}) = -\frac{\boldsymbol{\nabla}\Phi(\boldsymbol{b})}{\|\boldsymbol{\nabla}\Phi(\boldsymbol{b})\|}, \tag{3.20}$$

de modo que

$$\boldsymbol{\nabla}\Phi(\boldsymbol{b}) = -h(\boldsymbol{b})\hat{\boldsymbol{\delta}}(\boldsymbol{b}) = -\boldsymbol{\delta}(\boldsymbol{b}). \tag{3.21}$$

Del supuesto $h(\boldsymbol{b}^{(\infty)}) = h^{(\infty)} \neq 0$, contrario a lo que se pretende probar, y de la continuidad de $\boldsymbol{\nabla}\Phi(\boldsymbol{b})$ en $\boldsymbol{b}^{(\infty)}$, se desprende la existencia de una región U consistente en una bola de radio $R(\epsilon)$ centrada en $\boldsymbol{b}^{(\infty)}$ en la cual se cumple

$$\Big\|\boldsymbol{\nabla}\Phi(\boldsymbol{b}) - \boldsymbol{\nabla}\Phi^{(\infty)}\Big\| < \epsilon h^{(\infty)} \tag{3.22}$$

para todos los $\boldsymbol{b}$ en U.

Del teorema de la desigualdad triangular aplicado al triángulo formado por los vectores $\boldsymbol{\nabla}\Phi(\boldsymbol{b})$ y $\boldsymbol{\nabla}\Phi^{(\infty)}$ se obtiene

$$\Big|\Big\|\boldsymbol{\nabla}\Phi(\boldsymbol{b})\Big\| - \Big\|\boldsymbol{\nabla}\Phi^{(\infty)}\Big\|\Big| < \Big\|\boldsymbol{\nabla}\Phi(\boldsymbol{b}) - \boldsymbol{\nabla}\Phi^{(\infty)}\Big\| < \Big\|\boldsymbol{\nabla}\Phi(\boldsymbol{b})\Big\| + \Big\|\boldsymbol{\nabla}\Phi^{(\infty)}\Big\|, \tag{3.23}$$

de cuya parte izquierda, y utilizando (3.22), se concluye

$$\left|h(\boldsymbol{b}) - h^{(\infty)}\right| < \epsilon h^{(\infty)}. \tag{3.24}$$

Para derivar una desigualdad análoga a (3.24) que acote la desviación de los $\hat{\boldsymbol{\delta}}(\boldsymbol{b})$ en U respecto a $\hat{\boldsymbol{\delta}}^{(\infty)}$, se reescribe (3.24) en la forma

$$\left|\frac{h(\boldsymbol{b})}{h^{(\infty)}} - 1\right| < \epsilon, \tag{3.25}$$

y se divide (3.22) entre $h^{(\infty)}$ para obtener

$$\left\|\frac{h(\boldsymbol{b})}{h^{(\infty)}}\,\hat{\boldsymbol{\delta}}(\boldsymbol{b}) - \hat{\boldsymbol{\delta}}^{(\infty)}\right\| < \epsilon. \tag{3.26}$$

Esta desigualdad se expresa convenientemente en la forma equivalente

$$\left\|\Big(\frac{h(\boldsymbol{b})}{h^{(\infty)}} - 1\Big)\hat{\boldsymbol{\delta}}(\boldsymbol{b}) + \Big(\hat{\boldsymbol{\delta}}(\boldsymbol{b}) - \hat{\boldsymbol{\delta}}^{(\infty)}\Big)\right\| < \epsilon, \tag{3.27}$$

lo que implica

$$\left|\left\|\hat{\boldsymbol{\delta}}(\boldsymbol{b}) - \hat{\boldsymbol{\delta}}^{(\infty)}\right\| - \left|\frac{h(\boldsymbol{b})}{h^{(\infty)}} - 1\right|\right| < \epsilon, \tag{3.28}$$

y a la vez

$$\left|\frac{h(\boldsymbol{b})}{h^{(\infty)}} - 1\right| - \epsilon < \left\|\hat{\boldsymbol{\delta}}(\boldsymbol{b}) - \hat{\boldsymbol{\delta}}^{(\infty)}\right\| < \left|\frac{h(\boldsymbol{b})}{h^{(\infty)}} - 1\right| + \epsilon, \tag{3.29}$$

de cuya parte derecha y de (3.25) finalmente se concluye

$$\left\|\hat{\boldsymbol{\delta}}(\boldsymbol{b}) - \hat{\boldsymbol{\delta}}^{(\infty)}\right\| < 2\epsilon. \tag{3.30}$$

En general, para dos $\boldsymbol{b}_1$ y $\boldsymbol{b}_2$ cualesquiera en U se tienen entonces las expresiones

$$\left\|\hat{\boldsymbol{\delta}}(\boldsymbol{b}_1) - \hat{\boldsymbol{\delta}}^{(\infty)}\right\| < 2\epsilon \qquad \text{y} \qquad \left\|\hat{\boldsymbol{\delta}}(\boldsymbol{b}_2) - \hat{\boldsymbol{\delta}}^{(\infty)}\right\| < 2\epsilon, \tag{3.31}$$

que sumadas miembro a miembro resultan

$$\left\|\hat{\boldsymbol{\delta}}(\boldsymbol{b}_1) - \hat{\boldsymbol{\delta}}^{(\infty)}\right\| + \left\|\hat{\boldsymbol{\delta}}(\boldsymbol{b}_2) - \hat{\boldsymbol{\delta}}^{(\infty)}\right\| < 4\epsilon, \tag{3.32}$$

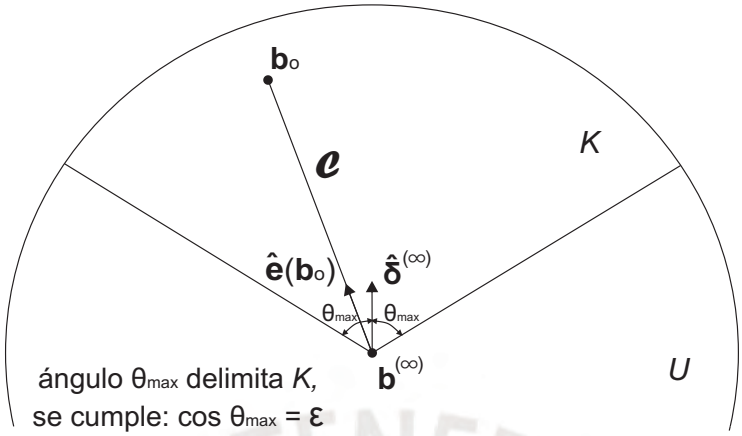


Figura 3.4: Región cónica $K \subset U$ de todos los puntos $\boldsymbol{b}$ que cumplen $\hat{\boldsymbol{e}}(\boldsymbol{b}) \cdot \hat{\boldsymbol{\delta}}^{(\infty)} \geq \epsilon$. Se demuestra que $\Phi(\boldsymbol{b}) < \Phi^{(\infty)}$, $\forall\, \boldsymbol{b} \in K$.

lo que implica, por el teorema de desigualdad triangular, que

$$\left\| \hat{\boldsymbol{\delta}}(\boldsymbol{b}_1) - \hat{\boldsymbol{\delta}}(\boldsymbol{b}_2) \right\| < 4\epsilon. \tag{3.33}$$

Para dos puntos tentativos consecutivos cualesquiera, $\boldsymbol{b}^{(s)}$ y $\boldsymbol{b}^{(s+1)}$, se cumple, debido a la ortogonalidad entre $\hat{\boldsymbol{\delta}}^{(s)}$ y $\hat{\boldsymbol{\delta}}^{(s+1)}$, que

$$\|\hat{\boldsymbol{\delta}}^{(s+1)} - \hat{\boldsymbol{\delta}}^{(s)}\| = \sqrt{2}. \tag{3.34}$$

Suponiendo que dichos $\boldsymbol{b}^{(s)}$ y $\boldsymbol{b}^{(s+1)}$ estén contenidos en U, debe cumplirse por (3.33) que $\|\hat{\boldsymbol{\delta}}^{(s+1)} - \hat{\boldsymbol{\delta}}^{(s)}\| < 4\epsilon$, lo cual contradice (3.34) para valores de ϵ menores a $\sqrt{2}/4$. De esto se concluye que $\boldsymbol{b}^{(s)}$ y $\boldsymbol{b}^{(s+1)}$ no pueden estar simultáneamente contenidos en U.

Definiendo asimismo para cada punto $\boldsymbol{b}$ en U el vector

$$\hat{\boldsymbol{e}}(\boldsymbol{b}) = \frac{\boldsymbol{b} - \boldsymbol{b}^{(\infty)}}{\|\boldsymbol{b} - \boldsymbol{b}^{(\infty)}\|}, \tag{3.35}$$

correspondiente a la dirección de su posición respecto a $\boldsymbol{b}^{(\infty)}$, es posible delimitar una región cónica $K \subset U$, como se muestra en la figura 3.4, compuesta por todos los puntos $\boldsymbol{b} \in U$ en los que el ángulo $\theta(\boldsymbol{b})$ entre su dirección $\hat{\boldsymbol{e}}(\boldsymbol{b})$ y la dirección fija $\hat{\boldsymbol{\delta}}^{(\infty)}$ es menor o igual a un ángulo máximo θ_{max} cuyo coseno

es ϵ, i.e.,

$$\cos\theta(\boldsymbol{b}) = \hat{\boldsymbol{e}}(\boldsymbol{b}) \cdot \hat{\boldsymbol{\delta}}^{(\infty)} \geq \epsilon = \cos\theta_{max}, \qquad \text{siempre que} \qquad 0 < \epsilon \leq 1. \tag{3.36}$$

El valor de Φ para un punto $\boldsymbol{b}_0$ cualquiera en K se puede calcular mediante la integral

$$\Phi(\boldsymbol{b}_0) = \Phi^{(\infty)} - \int_{\mathcal{C}} (-\boldsymbol{\nabla}\Phi(\boldsymbol{b})) \cdot \hat{\boldsymbol{e}}(\boldsymbol{b})\, \mathrm{d}c, \tag{3.37}$$

donde el camino de integración $\mathcal{C}$ es una línea recta que une $\boldsymbol{b}^{(\infty)}$ con $\boldsymbol{b}_0$.

De (3.21) y (3.22) se deriva

$$\boldsymbol{\nabla}\Phi(\boldsymbol{b}) = -h^{(\infty)}\hat{\boldsymbol{\delta}}^{(\infty)} + \boldsymbol{\gamma}(\boldsymbol{b}), \qquad \text{con} \qquad \left\|\boldsymbol{\gamma}(\boldsymbol{b})\right\| < \epsilon h^{(\infty)}, \tag{3.38}$$

con lo cual el integrando $I(\boldsymbol{b})$ de la integral en (3.37) adquiere convenientemente la forma

$$I(\boldsymbol{b}) = \left[h^{(\infty)}\hat{\boldsymbol{\delta}}^{(\infty)} - \boldsymbol{\gamma}(\boldsymbol{b})\right] \cdot \hat{\boldsymbol{e}}(\boldsymbol{b}), \tag{3.39}$$

y puesto que

$$\boldsymbol{\gamma}(\boldsymbol{b}) \cdot \hat{\boldsymbol{e}}(\boldsymbol{b}) \leq \left\|\boldsymbol{\gamma}(\boldsymbol{b})\right\| < \epsilon h^{(\infty)}, \tag{3.40}$$

se obtiene para $I(\boldsymbol{b})$ la cota inferior

$$I(\boldsymbol{b}) > h^{(\infty)}\left[\hat{\boldsymbol{\delta}}^{(\infty)} \cdot \hat{\boldsymbol{e}}(\boldsymbol{b}) - \epsilon\right]. \tag{3.41}$$

De esta cota inferior y de la condición en (3.36), que satisfacen los $\boldsymbol{b}$ en K, resulta

$$I(\boldsymbol{b}) > 0. \tag{3.42}$$

A partir de (3.37), y considerando que el integrando es positivo para todos los puntos en K (incluídos los que forman el camino de integración $\mathcal{C}$), se concluye para todos los puntos $\boldsymbol{b}_0$ contenidos en K que

$$\Phi(\boldsymbol{b}_0) < \Phi^{(\infty)}. \tag{3.43}$$

La última parte de la demostración de convergencia requiere la existencia de una región V construída a partir de una bola, de cierto radio $R(\epsilon')$, concéntrica e interior a U a la que se le extrae la región cónica correspondiente a su intersección con K. Dicha región V debe cumplir la condición de que el

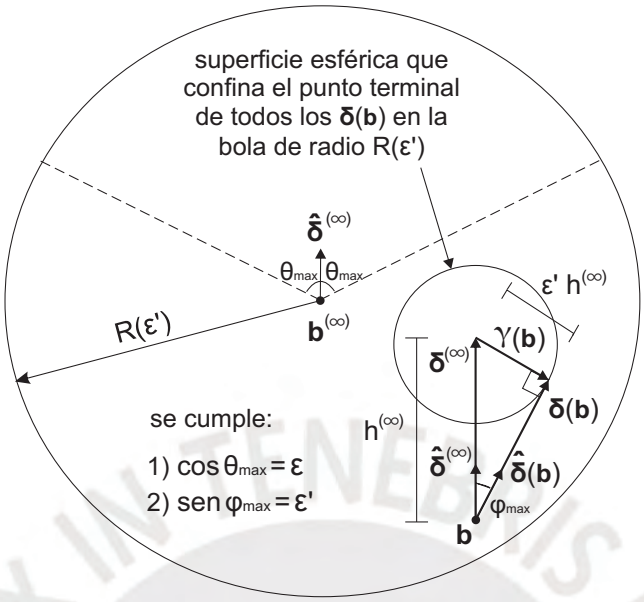


Figura 3.5: Construcción geométrica de $\boldsymbol{\delta}(\boldsymbol{b})$, cuyo punto terminal pertenece a la bola de radio $\epsilon' h^{(\infty)}$. El ángulo máximo φ_{max} se obtiene cuando $\boldsymbol{\delta}(\boldsymbol{b})$ es tangente a la superficie exterior de la bola.

rayo en la dirección $\hat{\boldsymbol{\delta}}(\boldsymbol{b})$ desde cualquier punto $\boldsymbol{b}$ en V atraviese K.

Antes de probar la existencia de dicha región V, es conveniente determinar el valor máximo φ_{max} que puede asumir el ángulo $\varphi(\boldsymbol{b})$ entre $\hat{\boldsymbol{\delta}}(\boldsymbol{b})$ y $\hat{\boldsymbol{\delta}}^{(\infty)}$ en cualquier punto $\boldsymbol{b}$ de la bola de radio $R(\epsilon')$.

De la definición de continuidad utilizada para deducir (3.22), se deduce igualmente para los puntos de dicha bola que

$$\boldsymbol{\delta}(\boldsymbol{b}) = h^{(\infty)}\hat{\boldsymbol{\delta}}^{(\infty)} + \boldsymbol{\gamma}(\boldsymbol{b}), \qquad \text{con} \qquad \left\|\boldsymbol{\gamma}(\boldsymbol{b})\right\| < \epsilon' h^{(\infty)}, \tag{3.44}$$

La interpretación geométrica de esta expresión, ilustrada en la figura 3.5, indica que si se coloca en $\boldsymbol{b}$ el punto inicial de $\boldsymbol{\delta}^{(\infty)}$ y $\boldsymbol{\delta}(\boldsymbol{b})$, entonces el punto terminal de este último estará en la bola con centro en el punto terminal de $\boldsymbol{\delta}^{(\infty)}$ y radio $\epsilon' h^{(\infty)}$. Esto significa que el caso de ángulo máximo, $\varphi(\boldsymbol{b}) = \varphi_{max}$, se presenta cuando algún $\boldsymbol{\delta}(\boldsymbol{b})$ particular es tangente a la bola. Por lo tanto

$$\operatorname{sen} \varphi_{max} = \epsilon'. \tag{3.45}$$

Cabe indicar que si el valor de ϵ' es tal que ocasiona que $\varphi_{max} > \theta_{max}$ en la bola de radio $R(\epsilon')$ correspondiente, puede entonces existir en dicha

bola algún punto $\boldsymbol{b}$ desde el cual el rayo en la dirección $\hat{\boldsymbol{\delta}}(\boldsymbol{b})$ posea un ángulo $\varphi(\boldsymbol{b}) > \theta_{max}$ respecto a $\hat{\boldsymbol{\delta}}^{(\infty)}$ que ocasione que este no atraviese a K. Tal bola, en consecuencia, no puede dar origen a la región V, pues no presenta la propiedad que por definición debe poseer V. Para evitar esta situación se introduce la condición

$$\epsilon' < \sqrt{1-\epsilon^2}, \qquad \text{con lo cual :} \qquad \varphi_{max} < \theta_{max}. \tag{3.46}$$

Considérese a continuación el cono Q construído tomando como generatriz la familia de rayos que forman un ángulo φ_{max} con $\hat{\boldsymbol{\delta}}^{(\infty)}$ y como vértice algún punto sobre el rayo que parte de $\boldsymbol{b}^{(\infty)}$ siguiendo la dirección $-\hat{\boldsymbol{\delta}}^{(\infty)}$, de manera tal que su base coincida con la base del cono que delimita K.

Para el valor umbral $\epsilon' = \sqrt{1-\epsilon^2}$, definido en (3.46), se tiene que el vértice de Q está en $\boldsymbol{b}^{(\infty)}$ y por ende Q coincide con el cono de K. Si a partir de dicho umbral se disminuye progresivamente el valor de ϵ', ocurre simultáneamente que la generatriz de Q se torna cada vez más vertical, causando que el vértice del mismo se aleje progresivamente de $\boldsymbol{b}^{(\infty)}$, mientras por otro lado la región V se contrae al disminuir el radio $R(\epsilon')$ que la limita. En el transcurso de este proceso se llega a un estado, mostrado en la figura 3.6a, en el que V resulta tangente e interior a Q. El valor de ϵ' que produce este estado se obtiene resolviendo la ecuación

$$\sqrt{1-\epsilon^2}\,\sqrt{1-\epsilon'^2} - \epsilon\,\epsilon' = R(\epsilon')/R(\epsilon). \tag{3.47}$$

Es posible probar gráficamente que la región V construída a partir de la bola de radio ϵ' obtenido de (3.47), efectivamente cumple el requisito de que desde cualquier punto interior $\boldsymbol{b}$ partan rayos con respectivas direcciones $\hat{\boldsymbol{\delta}}(\boldsymbol{b})$ que atraviesen K.

Considérese para ello el caso crítico en el que los rayos parten de los puntos ubicados en la circunferencia de contacto C_1 entre V y Q. Probar que dichos rayos atraviesan K es suficiente para garantizar que ocurra lo mismo para el resto de puntos en V.

En la figura 3.6b se muestra el segmento de una línea generatriz de Q que va del punto $\boldsymbol{b_1}$ en C_1 al punto $\boldsymbol{b_2}$ contenido en el plano de la circunferencia de intersección C_2 entre Q y K. Dado que el ángulo que forma dicho segmento con $\boldsymbol{b}^{(\infty)}$ es φ_{max}, todos los rayos que partan de $\boldsymbol{b_1}$ y posean ese mismo

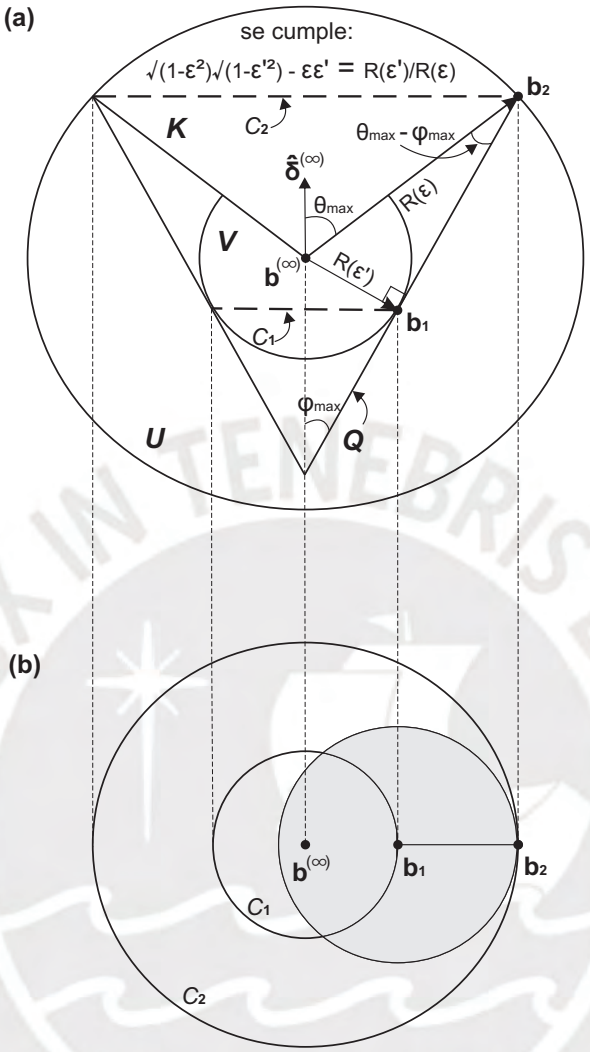


Figura 3.6: **(*a*)** Construcción de la región V a partir de un ϵ' que la hace tangente e interior a Q, garantizando así que todos los rayos que partan de algún $\boldsymbol{b}$ en V con dirección $\hat{\boldsymbol{\delta}}(\boldsymbol{b})$ atraviesen K.

(*b*) Vista de V, Q y K en la dirección de $\hat{\boldsymbol{\delta}}^{(\infty)}$, donde se observa el caso crítico en el que los rayos que parten de cualquier punto en la circunferencia de contacto C_1 entre V y Q caen necesariamente dentro de la región circular sombreada ubicada en el plano que contiene la circunferencia de intersección C_2 entre Q y K.

ángulo máximo intersectarán el plano de C_2 formando una circunferencia circundante a la región circular sombreada que aparece en la figura. Todos los demás rayos con ángulos menores a φ_{max} atravesarán necesariamente dicha región sombreada. Esto significa que los rayos que partan de $\boldsymbol{b_1}$ en todas las posibles direcciones permitidas atravesarán K puesto que la región sombreada es interior a K.

Aunque esta prueba gráfica asume un espacio geométrico tridimensional, sus resultados pueden extenderse al caso general en k dimensiones. De esta manera queda probada la existencia de la región V.

Finalmente, dado que V y K rodean $\boldsymbol{b}^{(\infty)}$, y este a su vez es el punto de convergencia de la secuencia $\boldsymbol{b}^{(0)}, \boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \ldots$, existe entonces algún punto $\boldsymbol{b}^{(s)}$ de la secuencia contenido en V, desde el cual el rayo con dirección $\hat{\boldsymbol{\delta}}^{(s)}$ atraviesa necesariamente K. Asimismo, sea $\boldsymbol{\zeta}$ un punto sobre la parte de dicho rayo interior a K. Este punto $\boldsymbol{\zeta}$ se encuentra antes del punto $\boldsymbol{b}^{(s+1)}$ en el recorrido a lo largo del camino de minimización C seguido por el método, pues $\boldsymbol{b}^{(s+1)}$ está impedido por (3.33) y (3.34) de pertenecer a U. Del caracter monótono decreciente de Φ en C y de (3.37) se tiene por lo tanto

$$\Phi^{(s+1)} < \Phi(\boldsymbol{\zeta}) < \Phi^{(\infty)},$$

resultado que contradice el supuesto inicial dado en (3.19).

Esta contradicción, al ser consecuencia de asumir $h^{(\infty)} \neq 0$, demuestra que efectivamente $\boldsymbol{b}^{(\infty)}$ es un punto estacionario de Φ en el cual se cumple $\boldsymbol{\nabla}\Phi^{(\infty)} = 0$.

3.3. Método de Gauss-Newton

A diferencia del método de la sección anterior en el que las características del modelo f a ajustar intervienen solo indirectamente, en el método de Gauss-Newton se utiliza una versión lineal de f correspondiente a su expansión en series de Taylor de primer grado alrededor del punto tentativo $\boldsymbol{b}^{(s)}$ de la iteración anterior para determinar la dirección de minimización sobre la cual se buscará el punto tentativo $\boldsymbol{b}^{(s+1)}$ de la iteración actual.

3.3.1. Elección de la Dirección de Minimización

Como se mencionó, la esencia del método se basa en sustituir el modelo no lineal f de k parámetros de ajuste $\boldsymbol{b} = (b_1, b_2, \ldots, b_k)$ y m variables independientes $\boldsymbol{x}_i = (x_{1,i}, x_{2i}, \ldots, x_{mi})$ que representan cada uno de los n juegos de datos de entrada $(i = 1, 2, \ldots, n)$,

$$\hat{y}_i = f(\boldsymbol{x}_i, \boldsymbol{b}) = f_i(\boldsymbol{b}),$$

donde $\hat{y}_i$ es la respuesta predicha por f correspondiente a la i-ésima respuesta medida y_i, por su expansión en serie de Taylor truncada en los términos lineales

$$\langle y_i \rangle = \langle f_i(\boldsymbol{b}^{(s)} + \boldsymbol{\delta}) \rangle = f_i(\boldsymbol{b}^{(s)}) + \sum_{j=1}^{k} \frac{\partial f_i(\boldsymbol{b}^{(s)})}{\partial b_j} \delta_j, \tag{3.48}$$

que en notación vectorial toma la forma

$$\langle \boldsymbol{y} \rangle = \boldsymbol{f}_0^{(s)} + P^{(s)} \boldsymbol{\delta}, \tag{3.48a}$$

donde evidentemente

$$\boldsymbol{f}_{0\ [n\times 1]}^{(s)} = \left(f_i(\boldsymbol{b}^{(s)}) \right), \qquad i = 1, 2, \ldots, n,$$

y

$$P_{[n\times k]}^{(s)} = \left(\frac{\partial f_i(\boldsymbol{b}^{(s)})}{\partial b_j} \right), \qquad i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, k.$$

Los corchetes $\langle\ \rangle$ se emplean para distinguir las respuestas predichas por la aproximación lineal del modelo de aquellas predichas por el modelo no lineal original.

La función de discrepancia original Φ dada en (2.3) es aproximada utilizando esta versión lineal del modelo mediante la expresión

$$\langle \Phi \rangle = \sum_{i=1}^{n} (y_i - \langle y_i \rangle)^2, \tag{3.49}$$

con lo cual el valor buscado es ahora aquel $\boldsymbol{\delta}$ que minimice $\langle \Phi \rangle$.

Dado que $\boldsymbol{\delta}$ aparece linealmente en (3.48), el cálculo del mismo se puede llevar a cabo directamente resolviendo el sistema de ecuaciones

$$\frac{\partial \langle \Phi \rangle}{\partial \delta_j} = 0, \qquad j = 1, 2, \ldots, k, \tag{3.50}$$

cuya forma explícita es equivalente a la ecuación matricial

$$A^{(s)} \boldsymbol{\delta} = \boldsymbol{g}^{(s)}, \tag{3.51}$$

donde

$$A_{[k\times k]}^{(s)} = [P^{(s)}]^T P^{(s)} \qquad \text{y} \qquad \boldsymbol{g}_{[k\times 1]}^{(s)} = [P^{(s)}]^T (\boldsymbol{y} - \boldsymbol{f}_0^{(s)}).$$

El valor así obtenido,

$$\boldsymbol{\delta} = [A^{(s)}]^{-1}\boldsymbol{g}^{(s)}, \tag{3.52}$$

define la dirección de minimización

$$\hat{\boldsymbol{\delta}}^{(s)} = \hat{\boldsymbol{\delta}}$$

a emplearse en la iteración actual.

En la práctica ha probado ser conveniente considerar únicamente una fracción de $\boldsymbol{\delta}$ como vector de paso $\boldsymbol{\delta}^{(s)}$; caso contrario la extrapolación podría estar fuera de la región donde (3.48) constituye una buena aproximación de f, lo que eventualmente originaría la divergencia del proceso iterativo.

Como consecuencia, el método establece, para la $(s+1)$-ava iteración, el vector de paso

$$\boldsymbol{\delta}^{(s)} = \kappa_0\,\boldsymbol{\delta}, \qquad 0 < \kappa_0 \leq 1. \tag{3.53}$$

3.3.2. Cálculo de la Longitud del Paso

De (3.53) se desprende la longitud del vector de paso

$$\alpha_s = \kappa_0\,\|\boldsymbol{\delta}\|, \tag{3.54}$$

cuyo cálculo se reduce a determinar el valor κ_0 que minimice, dentro del intervalo $]0,\,1]$, a la función

$$g(\kappa) = \Phi(\boldsymbol{b}^{(s)} + \kappa\boldsymbol{\delta}) \qquad \kappa \in\,]0,\,1], \tag{3.55}$$

obtenida al evaluar Φ en el punto tentativo resultante para cada valor de κ.

Dado que llevar a cabo la tarea de determinar de manera exacta dicho valor κ_0 es impráctico por el alto costo computacional involucrado, se prefiere resolver el problema de manera aproximada. Un procedimiento simple y confiable para tal fin consiste reemplazar $g(\kappa)$ por la parábola $g^*(\kappa)$ que pasa por los puntos $(0,\,g(0))$, $(1/2,\,g(1/2))$ y $(1,\,g(1))$, i.e.,

$$g^*(\kappa) = 2\,[g(0) - 2g(1/2) + g(1)]\,\kappa^2 - [3g(0) - 4g(1/2) + g(1)]\,\kappa + g(0), \tag{3.56}$$

y aproximar el valor κ_0 buscado mediante el valor

$$\kappa_0 = \frac{1}{2} + \frac{1}{4}\,\frac{g(0) - g(1)}{g(0) - 2g(1/2) + g(1)} \tag{3.57}$$

correspondiente al mínimo de la parábola. Dado que este mínimo debe estar en el intervalo mencionado, se le asigna el valor uno en caso este sobrepase dicho límite.

Finalmente, se debe comprobar que $\Phi^{(s+1)} < \Phi^{(s)}$, caso contrario habría que reducir la longitud del paso a la mitad.

3.3.3. Demostración de la Convergencia del Método

La demostración de la convergencia del método de Gauss-Newton que aquí se presenta es básicamente una adaptación del procedimiento empleado en la demostración descrita en [12].

Como paso previo se asume que el modelo $f(\boldsymbol{x}_i, \boldsymbol{b}) = f_i(\boldsymbol{b})$ satisface las siguientes condiciones:

a) Para todos los vectores de entrada $\boldsymbol{x}_i$ $(i = 1, 2, \ldots, n)$, tanto las primeras derivadas de $f_i(\boldsymbol{b})$ respecto a los b_j

$$\frac{\partial f_i(\boldsymbol{b})}{\partial b_j}, \qquad j = 1, 2, \ldots, k,$$

como las segundas derivadas

$$\frac{\partial^2 f_i(\boldsymbol{b})}{\partial b_j \partial b_h}, \qquad j = 1, 2, \ldots, k, \quad h = 1, 2, \ldots, k,$$

son funciones continuas de $\boldsymbol{b}$.

Este supuesto permite no solo la existencia de las primeras derivadas de la función de discrepancia original Φ

$$\frac{\partial \Phi}{\partial b_j}(\boldsymbol{b}) = -2 \sum_{i=1}^{n} (y_i - f_i(\boldsymbol{b})) \frac{\partial f_i(\boldsymbol{b})}{\partial b_j},$$

sino también la aproximación lineal del modelo según (3.48) y, con ello, la definición del sistema de ecuaciones correspondiente a los mínimos cuadrados lineales dado en (3.51).

b) Para todo vector $\boldsymbol{u} = (u_1, u_2, \ldots, u_k)$, con $\|\boldsymbol{u}\| \neq 0$, y todo $\boldsymbol{b}$ dentro de cierto subconjunto convexo S del espacio de parámetros, se cumple

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{k} u_j \frac{\partial f_i(\boldsymbol{b})}{\partial b_j} \right)^2 > 0, \tag{3.58}$$

lo que significa que el conjunto de n vectores

$$\boldsymbol{v}_i = \left(\frac{\partial f_i(\boldsymbol{b})}{\partial b_1}, \frac{\partial f_i(\boldsymbol{b})}{\partial b_2}, \ldots, \frac{\partial f_i(\boldsymbol{b})}{\partial b_k} \right), \qquad i = 1, 2, \ldots, n,$$

contiene una base del espacio de parámetros $\mathbf{R}^k$.

Este supuesto garantiza la existencia de un único vector solución $\boldsymbol{\delta}$ del sistema de ecuaciones correspondiente a los mínimos cuadrados lineales en (3.51), ya que de él se desprende que el rango de la matriz $A^{(s)}_{[k\times k]}$ de dicho sistema de ecuaciones es justamente k, lo que equivale a que $A^{(s)}$ sea invertible.

Una manera sencilla de derivar (3.58) consiste en considerar (3.48) y (3.49) para reescribir las ecuaciones (3.51) en la forma

$$\left.\begin{array}{ccccccc}
(y_1 - \langle y_1 \rangle)\dfrac{\partial f_1(\boldsymbol{b})}{\partial b_1} & + & (y_2 - \langle y_2 \rangle)\dfrac{\partial f_2(\boldsymbol{b})}{\partial b_1} & + \ldots + & (y_n - \langle y_n \rangle)\dfrac{\partial f_n(\boldsymbol{b})}{\partial b_1} & = & 0 \\
(y_1 - \langle y_1 \rangle)\dfrac{\partial f_1(\boldsymbol{b})}{\partial b_2} & + & (y_2 - \langle y_2 \rangle)\dfrac{\partial f_2(\boldsymbol{b})}{\partial b_2} & + \ldots + & (y_n - \langle y_n \rangle)\dfrac{\partial f_n(\boldsymbol{b})}{\partial b_2} & = & 0 \\
\vdots & & \vdots & \ddots & \vdots & & \vdots \\
(y_1 - \langle y_1 \rangle)\dfrac{\partial f_1(\boldsymbol{b})}{\partial b_k} & + & (y_2 - \langle y_2 \rangle)\dfrac{\partial f_2(\boldsymbol{b})}{\partial b_k} & + \ldots + & (y_n - \langle y_n \rangle)\dfrac{\partial f_n(\boldsymbol{b})}{\partial b_k} & = & 0
\end{array}\right\}, \tag{3.59}$$

donde se evidencia que estas serán linealmente independientes si y solo si para todo $\boldsymbol{u}$, con $\|\boldsymbol{u}\| \neq 0$, se cumple

$$\begin{array}{ccc}
(y_1 - \langle y_1 \rangle)\left(u_1 \dfrac{\partial f_1(\boldsymbol{b})}{\partial b_1} + u_2 \dfrac{\partial f_1(\boldsymbol{b})}{\partial b_2} + \ldots + u_k \dfrac{\partial f_1(\boldsymbol{b})}{\partial b_k} \right) & + & \\
(y_2 - \langle y_2 \rangle)\left(u_1 \dfrac{\partial f_2(\boldsymbol{b})}{\partial b_1} + u_2 \dfrac{\partial f_2(\boldsymbol{b})}{\partial b_2} + \ldots + u_k \dfrac{\partial f_2(\boldsymbol{b})}{\partial b_k} \right) & + & \\
\vdots & & \\
(y_n - \langle y_n \rangle)\left(u_1 \dfrac{\partial f_n(\boldsymbol{b})}{\partial b_1} + u_2 \dfrac{\partial f_n(\boldsymbol{b})}{\partial b_2} + \ldots + u_k \dfrac{\partial f_n(\boldsymbol{b})}{\partial b_k} \right) & \neq & 0,
\end{array} \tag{3.60}$$

expresión de la cual es posible deducir el requerimiento (3.58).

c) Dado el valor

$$\Phi_0 = \inf_{\bar{S}} \Phi(\boldsymbol{b}),$$

donde $\bar{S}$ es el complemento de S, es posible encontrar un vector $\boldsymbol{b}^{(0)}$ al interior de S tal que

$$\Phi(\boldsymbol{b}^{(0)}) < \Phi_0.$$

Para los fines de esta demostración es conveniente expresar (3.51) como un sistema de ecuaciones en forma explícita

$$2\sum_{j=1}^{k}\left[\sum_{i=1}^{n}\frac{\partial f_i(\boldsymbol{b}^{(s)})}{\partial b_h}\frac{\partial f_i(\boldsymbol{b}^{(s)})}{\partial b_j}\right]\delta_j = -\frac{\partial \Phi(\boldsymbol{b}^{(s)})}{\partial b_h}, \qquad h = 1, 2, \ldots, k, \quad (3.61)$$

el cual por la condición b) siempre posee un único vector solución $\boldsymbol{\delta}$ para cada $\boldsymbol{b}^{(s)}$ en S.

Como se indicó en la subsección anterior, para todo $\boldsymbol{b}^{(s)}$ se tiene un vector de paso $\boldsymbol{\delta}^{(s)}$ expresado en función de la correspondiente solución $\boldsymbol{\delta}$ como

$$\boldsymbol{\delta}^{(s)} = \kappa_0\,\boldsymbol{\delta}, \qquad 0 < \kappa_0 \leq 1,$$

donde el factor κ_0, introducido en (3.53), se calcula de manera que en el siguiente punto tentativo $\boldsymbol{b}^{(s+1)}$ se tenga $\Phi^{(s+1)} < \Phi^{(s)}$. Por lo tanto, la secuencia $\Phi^{(0)}, \Phi^{(1)}, \Phi^{(2)}, \ldots$, es decreciente. Asimismo, si $\boldsymbol{b}^{(0)}$ se escoge al interior de S se concluye entonces por la condición c) que todos los vectores de la secuencia $\boldsymbol{b}^{(0)}, \boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \ldots$, están igualmente contenidos en S.

La secuencia de los $\Phi^{(s)}$ $(s = 0, 1, 2, \ldots)$ converge a un valor finito por ser decreciente y acotada inferiormente. Bajo esta premisa es posible demostrar, siguiendo el mismo procedimiento utilizado para el caso análogo en el método del máximo descenso (ver pág. 53), que la secuencia de los $\boldsymbol{b}^{(s)}$ $(s = 0, 1, 2, \ldots)$ converge a un vector denotado como $\boldsymbol{b}^{(\infty)}$, i.e., se cumple

$$\lim_{s\to\infty} \boldsymbol{b}^{(s)} = \boldsymbol{b}^{(\infty)} \quad (3.62)$$

e igualmente, por la continuidad de Φ respecto a $\boldsymbol{b}$,

$$\lim_{s\to\infty} \Phi^{(s)} = \Phi(\boldsymbol{b}^{(\infty)}) = \Phi^{(\infty)}. \quad (3.63)$$

Se requiere demostrar ahora que en el punto $\boldsymbol{b}^{(\infty)}$ todas las derivadas parciales de Φ son cero.

Si se introduce $\boldsymbol{\delta}^*$ como la solución del sistema de ecuaciones

$$2\sum_{j=1}^{k}\ \sum_{i=1}^{n}\frac{\partial f_i(\boldsymbol{b}^{(\infty)})}{\partial b_h}\frac{\partial f_i(\boldsymbol{b}^{(\infty)})}{\partial b_j}\Bigg)\,{\delta_j}^* = -\frac{\partial \Phi(\boldsymbol{b}^{(\infty)})}{\partial b_h}, \qquad h = 1, 2, \ldots, k, \tag{3.64}$$

y se asume, contrario a lo que se quiere demostrar, que no todas las derivadas parciales de Φ en $\boldsymbol{b}^{(\infty)}$ son cero ($\boldsymbol{\nabla}\Phi(\boldsymbol{b}^{(\infty)}) \neq 0$), entonces se concluye

$$\sum_{j=1}^{k}\left({\delta_j}^*\right)^2 > 0,$$

puesto que la matriz de dicho sistema de ecuaciones es invertible.

Considerando tanto este resultado como la condición c), y tomando el producto escalar de cada miembro de (3.64) con $\boldsymbol{\delta}^*$, se obtiene

$$\sum_{j=1}^{k}\frac{\partial \Phi(\boldsymbol{b}^{(\infty)})}{\partial b_j}{\delta_j}^* = -2\sum_{i=1}^{n}\left(\sum_{j=1}^{k}{\delta_j}^*\frac{\partial f_i(\boldsymbol{b}^{(\infty)})}{\partial b_j}\right)^2 < 0, \tag{3.65}$$

lo que significa que en el punto $\boldsymbol{b}^{(\infty)}$ la derivada de Φ en la dirección $\hat{\boldsymbol{\delta}}^* = \hat{\boldsymbol{\delta}}^{(\infty)}$ es negativa. Asimismo, la continuidad de $\boldsymbol{\nabla}\Phi(\boldsymbol{b})$ y $\boldsymbol{\delta}(\boldsymbol{b})$ en S implica que para cada $\boldsymbol{b}^{(s)}$ en una pequeña vecindad alrededor de $\boldsymbol{b}^{(\infty)}$ la derivada de Φ en la respectiva dirección de minimización $\hat{\boldsymbol{\delta}}^{(s)}$ es también negativa.

A partir de este resultado se puede demostrar que para todos los $\boldsymbol{b}^{(s)}$ en la vecindad mencionada se cumple que el valor de $\Phi^{(s+1)}$ está por debajo del valor de $\Phi^{(s)}$ en al menos una cierta cantidad fija ϵ. Por lo tanto, la secuencia de los $\Phi^{(s)}$ tiende a un valor infinito negativo, lo cual contradice (3.63) que establece que tal secuencia converge en el valor finito $\Phi^{(\infty)}$.

Dado que el supuesto inicial ($\boldsymbol{\nabla}\Phi(\boldsymbol{b}^{(\infty)}) \neq 0$) conduce a una contradicción, queda entonces demostrado que todas las derivadas parciales de Φ en el punto límite $\boldsymbol{b}^{(\infty)}$ son cero.

Capítulo 4

Método de Levenberg-Marquardt

Cuando el problema de mínimos cuadrados aumenta en complejidad, los métodos de Gauss-Newton (de la serie de Taylor) y del máximo descenso (de la gradiente) se vuelven inaplicables o muy ineficientes debido a sus limitaciones.

En el caso del método de Gauss-Newton la principal deficiencia es la potencial divergencia de los sucesivos parámetros tentativos. Cuando la iteración produce un nuevo punto tentativo lejano al anterior alrededor del cual se linealizó el modelo, el supuesto de linealidad deja de ser válido y el valor de Φ en el nuevo punto tentativo podría aumentar respecto al anterior. En ese caso, la secuencia de puntos tentativos varían de forma errática, produciendo una convergencia muy lenta o incluso llegando a divergir.

Por otro lado, la principal deficiencia del método del máximo descenso es la convergencia considerablemente lenta que presenta luego de las primeras iteraciones. Como es bien sabido, cuando el modelo es lineal las isosuperficies de Φ son elipsoides concéntricos, mientras que cuando el modelo es no lineal, las isosuperficies se distorsionan según el grado de la no linealidad. En este último caso, la dirección de minimización (i.e., el negativo de la gradiente) en puntos lejanos al mínimo puede no ser adecuada. Independientemente de que el modelo sea o no lineal, las isosuperficies en puntos cercanos al mínimo son prácticamente elipsoides, que en la mayoría de los casos poseen formas alargadas. Por consiguiente, si algún punto tentativo se encontráse en los extremos de dichas elipsoides alargadas, los pasos subsiguientes serían

progresivamente más pequeños, ralentizándose así la convergencia (véase la fig. 3.2 en la pág. 46).

Varias modificaciones han sido sugeridas para limitar la longitud del paso en el método de Gauss-Newton y para acondicionar mediante reescalamiento la topología de Φ en el método del máximo descenso. Sin embargo, los problemas de divergencia por un lado, y de convergencia lenta por el otro, no han podido ser satisfactoriamente resueltos. Esto genera la necesidad de contar con un método que supere las limitaciones de ambos métodos y al mismo tiempo conserve sus ventajas. Uno de los métodos que cumple tales requisitos es el llamado método de Levenberg-Marquardt.

En el presente capítulo se exponen los razonamientos seguidos tanto por Levenberg [20] como por Marquardt [21] al desarrollar, cada uno por su cuenta, el método que lleva sus nombres.

4.1. Enfoque de Levenberg

El principal objetivo de Levenberg consistía en limitar o *atenuar* la longitud del paso dado por el método de Gauss-Newton estándar en cada iteración. Un paso que conduzca a un punto tentativo dentro de la región de validez de la aproximación lineal, evita que el valor de Φ se incremente con cada iteración, y por lo tanto que ocurra divergencia.

4.1.1. Construcción del Método

Como se vió en el capítulo anterior, el método de Gauss-Newton estándar resuelve iterativamente el problema de mínimos cuadrados no lineales consistente en estimar el vector de parámetros $\boldsymbol{b}_{min}$ que minimice

$$\Phi = \sum_{i=1}^{n} r_i^2,$$

donde

$$r_i = y_i - f_i(\boldsymbol{b}),$$

es el residuo entre el i-ésimo dato medido y_i y su correspondiente valor predicho por el modelo no lineal f_i evaluado en algún vector arbitrario de parámetros $\boldsymbol{b}$.

El proceso iterativo mencionado consiste en una secuencia de problemas de mínimos cuadrados lineales cuyas soluciones se van acercando progresivamente a la solución buscada $\boldsymbol{b}_{min}$. En estos problemas se reemplaza el modelo no lineal f_i por una versión lineal $\langle f_i \rangle$ correspondiente a su serie de Taylor de primer grado alrededor de el punto tentativo actual $\boldsymbol{b}^{(s)}$. En este punto, que representa la aproximación de $\boldsymbol{b}_{min}$ obtenida tras la s-ésima iteración, se asume que Φ no posee un valor estacionario. Tras introducir el vector de incrementos $\boldsymbol{\delta} = \boldsymbol{b} - \boldsymbol{b}^{(s)}$ se obtiene

$$\langle f_i(\boldsymbol{b}^{(s)} + \boldsymbol{\delta}) \rangle = f_i(\boldsymbol{b}^{(s)}) + \frac{\partial f_i(\boldsymbol{b}^{(s)})}{\partial b_1}\delta_1 + \frac{\partial f_i(\boldsymbol{b}^{(s)})}{\partial b_2}\delta_2 + \cdots + \frac{\partial f_i(\boldsymbol{b}^{(s)})}{\partial b_k}\delta_k, \quad (4.1)$$

de donde se desprende la versión lineal de los residuos

$$\langle r_i \rangle = y_i - \langle f_i(\boldsymbol{b}^{(s)} + \boldsymbol{\delta}) \rangle. \quad (4.2)$$

Dicho esto, el problema lineal consiste entonces en estimar el vector $\boldsymbol{\delta}^*$ que minimice

$$\langle \Phi(\boldsymbol{\delta}) \rangle = \sum_{i=1}^{n} \langle r_i \rangle^2, \quad (4.3)$$

el cual se obtiene resolviendo el sistema de ecuaciones lineales resultante de igualar a cero las derivadas parciales de $\langle \Phi \rangle$ respecto a cada una de las componentes de $\boldsymbol{\delta}$, sistema al que se conoce como ecuaciones normales lineales

$$\left.\begin{array}{rcccccccccc}
\frac{1}{2}\frac{\partial\langle\Phi\rangle}{\partial\delta_1} & = & [\boldsymbol{c_1 c_1}]\delta_1 & + & [\boldsymbol{c_1 c_2}]\delta_2 & + & \ldots & + & [\boldsymbol{c_1 c_k}]\delta_k & + & [\boldsymbol{c_1 c_0}] & = & 0 \\
\frac{1}{2}\frac{\partial\langle\Phi\rangle}{\partial\delta_2} & = & [\boldsymbol{c_2 c_1}]\delta_1 & + & [\boldsymbol{c_2 c_2}]\delta_2 & + & \ldots & + & [\boldsymbol{c_2 c_k}]\delta_k & + & [\boldsymbol{c_2 c_0}] & = & 0 \\
\vdots & & \vdots & & \vdots & & \ddots & & \vdots & & \vdots & & \vdots \\
\frac{1}{2}\frac{\partial\langle\Phi\rangle}{\partial\delta_k} & = & [\boldsymbol{c_k c_1}]\delta_1 & + & [\boldsymbol{c_k c_2}]\delta_2 & + & \ldots & + & [\boldsymbol{c_k c_k}]\delta_k & + & [\boldsymbol{c_k c_0}] & = & 0
\end{array}\right\} \quad (4.4)$$

donde los corchetes [ ], introducidos en la pág. 14, representan el producto interno de los vectores contenidos, de forma tal que

$$[\boldsymbol{c_1 c_0}] = \sum_{i=1}^{n} \left(\frac{\partial f_i(\boldsymbol{b}^{(s)})}{\partial b_1} \left[f_i(\boldsymbol{b}^{(s)}) - y_i \right] \right),$$
$$[\boldsymbol{c_1 c_1}] = \sum_{i=1}^{n} \left(\frac{\partial f_i(\boldsymbol{b}^{(s)})}{\partial b_1} \frac{\partial f_i(\boldsymbol{b}^{(s)})}{\partial b_1} \right), \; [\boldsymbol{c_1 c_2}] = \sum_{i=1}^{n} \left(\frac{\partial f_i(\boldsymbol{b}^{(s)})}{\partial b_1} \frac{\partial f_i(\boldsymbol{b}^{(s)})}{\partial b_2} \right), \ldots \quad (4.5)$$

En las versiones modificadas del método de Gauss-Newton el vector de paso se define como $\boldsymbol{\delta}^{(s)} = \kappa_0 \boldsymbol{\delta}^*$, donde $0 < \kappa_0 < 1$, mientras en la versión estándar del mismo se toma $\kappa_0 = 1$. De esta manera, si el módulo de $\boldsymbol{\delta}^*$ es lo suficientemente grande, la aproximación lineal (4.1) no sería válida alrededor del siguiente punto tentativo, con lo cual una disminución en $\langle\Phi\rangle$ podría no corresponder a una disminución en Φ.

En tales casos, Levenberg propone limitar o *atenuar* el módulo del vector de incrementos $\boldsymbol{\delta}$ con el fin de mejorar la aproximación de Taylor de primer grado en (4.1) y simultáneamente minimizar el valor de $\langle\Phi\rangle$ en (4.3) bajo estas condiciones de atenuación. Asimismo, propone que para conseguir que tanto el módulo de $\boldsymbol{\delta}$ como el valor de $\langle\Phi\rangle$ sean pequeños, se emplee la idea fundamental de los mínimos cuadrados. Como resultado, se origina el llamado método de Levenberg-Marquardt en el que la expresión a minimizar es

$$\bar{\Phi}_w = w\langle\Phi(\boldsymbol{\delta})\rangle + Q(\boldsymbol{\delta}), \qquad \text{con :} \quad Q(\boldsymbol{\delta}) = \|\boldsymbol{\delta}\|^2, \tag{4.6}$$

donde w es un factor de ponderación positivo que expresa la importancia relativa entre $\langle\Phi\rangle$ y Q en el proceso de minimización.

Por comodidad se opta por trasladar el origen de coordenadas del espacio de parámetros al punto $\boldsymbol{b}^{(s)}$. Por consiguiente, el punto en el cual $\bar{\Phi}_w$ alcanza su valor mínimo es el vector de incrementos $\boldsymbol{\delta}_w$ que minimiza (4.6). Este vector de incrementos es el vector de paso $\boldsymbol{\delta}^{(s)}$ calculado por método.

Para obtener $\boldsymbol{\delta}_w$, las derivadas parciales de $\bar{\Phi}_w$ respecto a las distintas componentes del vector de incrementos $\boldsymbol{\delta}$ se igualan a cero, con lo cual

$$\frac{\partial\bar{\Phi}_w}{\partial\delta_j} = w\frac{\partial\langle\Phi\rangle}{\partial\delta_j} + 2\delta_j = 0, \qquad j = 1, 2, \ldots, k. \tag{4.7}$$

Al dividir cada una de estas ecuaciones entre $2w$, y substituir las expresiones de las derivadas parciales de $\langle\Phi\rangle$ por (4.4), resultan las *ecuaciones normales*

atenuadas

$$\left.\begin{array}{rcrcrcrcl} ([\boldsymbol{c_1c_1}]+w^{-1})\delta_1 & + & [\boldsymbol{c_1c_2}]\delta_2 & + & \ldots & + & [\boldsymbol{c_1c_k}]\delta_k & = & -[\boldsymbol{c_1c_0}] \\ [\boldsymbol{c_2c_1}]\delta_1 & + & ([\boldsymbol{c_2c_2}]+w^{-1})\delta_2 & + & \ldots & + & [\boldsymbol{c_2c_k}]\delta_k & = & -[\boldsymbol{c_2c_0}] \\ \vdots & & \vdots & & \ddots & & \vdots & & \vdots \\ [\boldsymbol{c_kc_1}]\delta_1 & + & [\boldsymbol{c_kc_2}]\delta_2 & + & \ldots & + & ([\boldsymbol{c_kc_k}]+w^{-1})\delta_k & = & -[\boldsymbol{c_kc_0}] \end{array}\right\} \tag{4.8}$$

Se puede apreciar que estas ecuaciones son iguales a las ecuaciones normales ordinarias (4.4), excepto por los coeficientes de la diagonal principal, los cuales están incrementados en w^{-1}. Al igual que para (4.4), la solución de (4.8) se puede obtener mediante métodos simplificados que explotan la simetría de la matriz de coeficientes del sistema.

Asimismo, puede verse que un valor de w que tiende al infinito, da como resultado el método de Gauss-Newton estándar, que por lo tanto viene a ser un caso especial del método de Levenberg-Marquardt, el cual puede interpretarse como un método de *mínimos cuadrados con atenuación.*

Para ilustrar la dependencia de este nuevo método respecto al parámetro w, la representación geométrica del problema de minimización de $\bar{\Phi}_w$ resulta ser de gran utilidad. Con ese fin, considérense simultáneamente la familia de isosuperficies esféricas concéntricas en $\boldsymbol{b}^{(s)}$ y la de isosuperficies elipsoidales concéntricas en $\boldsymbol{b}^{(s)}+\boldsymbol{\delta}^*$, correspondientes a Q y $\langle\Phi\rangle$ respectivamente. Nótese además que los valores de Q y $\langle\Phi\rangle$ aumentan conforme las isosuperficies se van alejando de sus centros. Asimismo, las isosuperficies de $w\langle\Phi\rangle$ son las mismas que las de $\langle\Phi\rangle$ para todos los valores de w.

Sea Q_0 una isosuperficie cualquiera de Q, y G el grupo de isosuperficies de $\langle\Phi\rangle$ que la intersectan. Para determinar el punto de Q_0 en el que $\bar{\Phi}_w$ toma el menor valor, basta con identificar en G la isosuperficie $\langle\Phi\rangle_0$ en la que $\langle\Phi\rangle$ es menor a las demás, ya que el valor de Q es el mismo en todos los puntos de Q_0. Tal $\langle\Phi\rangle_0$ es, de acuerdo a lo expresado en el párrafo anterior, la isosuperficie en G más cercana al centro común. En este punto cabe observar que si uno se desplaza secuencialmente por las isosuperficies de G de manera que estas se vayan acercando progresivamente al mínimo, se tiene que el límite de dicha secuencia es $\langle\Phi\rangle_0$. De esto se concluye que $\langle\Phi\rangle_0$ hace contacto con Q_0 en un único punto, en el cual además ambas son tangentes.

El anterior análisis cualitativo puede corroborarse analíticamente de ma-

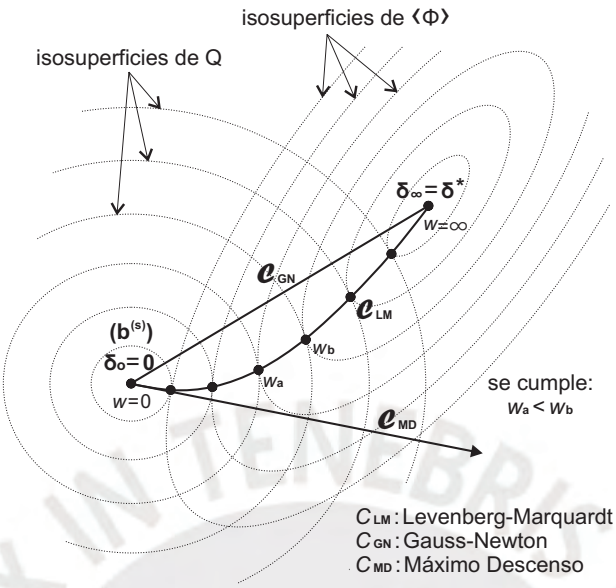


Figura 4.1: Camino de minimización $\mathcal{C}_{LM}$ del método de Levenberg-Marquardt, formado por los puntos que minimizan $\bar{\Phi}_w$, para todos los valores de w. El camino $\mathcal{C}_{LM}$ parte de $\boldsymbol{\delta}_0$ y termina en $\boldsymbol{\delta}_\infty$ pasando por los puntos donde las isosuperficies de Q y $\langle\Phi\rangle$ son tangentes. Se puede afirmar que $\mathcal{C}_{LM}$ es en cierto sentido una interpolación entre los caminos $\mathcal{C}_{GN}$ y $\mathcal{C}_{MD}$ correspondientes a los métodos de Gauss-Newton y del máximo descenso respectivamente.

nera simple. Sea $\boldsymbol{\delta}_w$ el punto donde $\bar{\Phi}_w = w\langle\Phi(\boldsymbol{\delta})\rangle + Q(\boldsymbol{\delta})$ es mínimo. Consecuentemente,

$$w\nabla\langle\Phi(\boldsymbol{\delta}_w)\rangle = -\nabla Q(\boldsymbol{\delta}_w). \tag{4.9}$$

Las gradientes de $\langle\Phi\rangle$ y Q son entonces paralelas en $\boldsymbol{\delta}_w$ para todos los valores de w, lo que implica que las isosuperficies de $\langle\Phi\rangle$ y Q que pasan por el mínimo de $\bar{\Phi}_w$, para cualquier w, son tangentes.

La línea que forman los puntos mínimos $\boldsymbol{\delta}_w$ correspondientes a cada valor de w, es el *camino de minimización* de el cual el método de Levenberg-Marquardt selecciona el vector de paso $\boldsymbol{\delta}^{(s)}$. En este caso particular, dicho camino es una curva que une los puntos $\boldsymbol{\delta}_0 = \mathbf{0}$ ($\boldsymbol{b}^{(s)}$ en el sistema de coordenadas original) y $\boldsymbol{\delta}_\infty = \boldsymbol{\delta}^*$ correspondientes a $w = 0$ y $w = \infty$ respectivamente. Por otro lado, los caminos de minimización de los métodos de Gauss-Newton y del máximo descenso son líneas rectas.

La figura 4.1 muestra la construcción del camino de minimización $\mathcal{C}_{LM}$, correspondiente al método de Levenberg-Marquard, a partir de los puntos de tangencia entre las isosuperficies de $\langle\Phi\rangle$ y Q, tal como se detalló ante-

riormente. Asimismo se muestran, a efectos de comparación, los caminos de minimización del método de Gauss-Newton ($\mathcal{C}_{GN}$) y del método del máximo descenso ($\mathcal{C}_{MD}$). Nótese que en el método de Levenberg-Marquardt, el parámetro w que controla la longitud del paso, define al mismo tiempo su dirección. Dicha dirección coincide con la dirección de minimización del método del máximo descenso cuando w tiende a cero, y varía continuamente hasta coincidir con la dirección de minimización del método de Gauss-Newton cuando w tiende al infinito.

En la figura se puede también observar que para cualquier valor de w finito, el método de Levenberg-Marquardt satisface los dos objetivos propuestos. Por un lado mejora la solución $\boldsymbol{\delta}_\infty$ del método de Gauss-Newton estándar por medio del vector $\boldsymbol{\delta}_w$ con menor módulo, y por otro consigue disminuir el valor inicial de $\langle\Phi\rangle$, esto es, $\langle\Phi(\boldsymbol{\delta}_w)\rangle < \langle\Phi(\boldsymbol{\delta}_0)\rangle$. Otra cualidad que se desprende de la figura es que el módulo de $\boldsymbol{\delta}_w$ disminuye conforme w disminuye.

4.1.2. Demostración de la Utilidad del Método

Las conclusiones recién obtenidas por inspección visual carecen evidentemente de rigor matemático. Por lo tanto, es necesario presentar las demostraciones analíticas que garanticen que el método cumple los objetivos de minimización propuestos.

Como punto de partida, nótese que en el caso trivial $\boldsymbol{\delta}_w = \boldsymbol{\delta}_0 = \boldsymbol{\delta}_\infty = \mathbf{0}$ (con w arbitrario), el vector solución del método se mantiene en el punto inicial, por lo que no se produce minimización alguna. Este caso trivial se presenta si y solo si los términos $[\boldsymbol{c_1 c_0}], [\boldsymbol{c_2 c_0}], \ldots, [\boldsymbol{c_k c_0}]$ en (4.4) son todos cero. Sin embargo, dichos términos nunca son todos cero puesto que vienen a ser las componentes de la gradiente de Φ en el punto $\boldsymbol{b}^{(s)}$, la cual se asumió no estacionaria en dicho punto. De esta manera queda excluído el único caso en el que no aplican las demostraciones dadas a continuación.

Por definición se tiene

$$w\langle\Phi(\boldsymbol{\delta}_w)\rangle + Q(\boldsymbol{\delta}_w) = \bar{\Phi}_w(\boldsymbol{\delta}_w) < \bar{\Phi}_w(\boldsymbol{\delta}_\infty) = w\langle\Phi(\boldsymbol{\delta}_\infty)\rangle + Q(\boldsymbol{\delta}_\infty),$$

y además, dado que $\boldsymbol{\delta}_\infty$ es el mínimo de $\langle\Phi\rangle$, se tiene

$$w\langle\Phi(\boldsymbol{\delta}_\infty)\rangle + Q(\boldsymbol{\delta}_\infty) < w\langle\Phi(\boldsymbol{\delta}_w)\rangle + Q(\boldsymbol{\delta}_\infty),$$

de forma que por transitividad resulta

$$Q(\boldsymbol{\delta}_w) < Q(\boldsymbol{\delta}_\infty). \tag{4.10}$$

Esta desigualdad muestra que para cualquier w, el método de Levenberg-Marquardt produce un vector de incrementos $\boldsymbol{\delta}_w$ cuyo módulo es menor a su correspondiente obtenido por el método de Gauss-Newton estándar.

Asimismo, por definición se cumple

$$w\langle\Phi(\boldsymbol{\delta}_w)\rangle < w\langle\Phi(\boldsymbol{\delta}_w)\rangle + Q(\boldsymbol{\delta}_w) = \bar{\Phi}_w(\boldsymbol{\delta}_w) < \bar{\Phi}_w(\boldsymbol{\delta}_0) = w\langle\Phi(\boldsymbol{\delta}_0)\rangle + Q(\boldsymbol{\delta}_0),$$

y puesto que $\boldsymbol{\delta}_0 = \mathbf{0}$, se cumple además

$$w\langle\Phi(\boldsymbol{\delta}_0)\rangle + Q(\boldsymbol{\delta}_0) = w\langle\Phi(\boldsymbol{\delta}_0)\rangle,$$

con lo cual se obtiene

$$\langle\Phi(\boldsymbol{\delta}_w)\rangle < \langle\Phi(\boldsymbol{\delta}_0)\rangle. \tag{4.11}$$

Esta expresión indica que para cualquier w, el vector de incrementos $\boldsymbol{\delta}_w$ obtenido como resultado del método de Levenberg-Marquardt disminuye el valor inicial de $\langle\Phi\rangle$.

Sin embargo, para garantizar su utilidad, es necesario demostrar que el método es capaz de reducir el valor que toma Φ en el punto inicial $\boldsymbol{b}^{(s)}$ representado por el vector de incrementos $\boldsymbol{\delta} = \boldsymbol{\delta}_0 = \mathbf{0}$. Con ese fin, es conveniente primero resolver las ecuaciones (4.8). Así, utilizando la técnica de los determinantes vista en la sección 2.2.1, resulta para la primera componente

$$(\delta_w)_1 = \frac{-[\boldsymbol{c_1}\boldsymbol{c_0}]w^{1-k} + \cdots}{w^{-k} + \cdots} = -[\boldsymbol{c_1}\boldsymbol{c_0}]w + \cdots, \tag{4.12}$$

con lo cual, en la vecindad $w = 0$, se tiene

$$\left.\frac{\mathrm{d}(\delta_w)_1}{\mathrm{d}w}\right|_{w=0} = -[\boldsymbol{c_1}\boldsymbol{c_0}]. \tag{4.13}$$

Para el resto de componentes se sigue este mismo procedimiento.

Por otra parte, dada la relación definida entre un vector de parámetros $\boldsymbol{b}$

y su respectivo vector de incrementos $\boldsymbol{\delta}$, es posible demostrar que

$$\frac{\mathrm{d}\Phi}{\mathrm{d}w} = \frac{\partial\Phi}{\partial b_1}\cdot\frac{\mathrm{d}(\delta_w)_1}{\mathrm{d}w} + \frac{\partial\Phi}{\partial b_2}\cdot\frac{\mathrm{d}(\delta_w)_2}{\mathrm{d}w} + \cdots + \frac{\partial\Phi}{\partial b_k}\cdot\frac{\mathrm{d}(\delta_w)_k}{\mathrm{d}w}; \tag{4.14}$$

y de la definición del símbolo de sumatoria [ ], es posible igualmente demostrar que

$$\left.\frac{\partial\Phi}{\partial b_1}\right|_{\boldsymbol{b}=\boldsymbol{b}^{(s)}} = 2[\boldsymbol{c_1 c_0}], \qquad \left.\frac{\partial\Phi}{\partial b_2}\right|_{\boldsymbol{b}=\boldsymbol{b}^{(s)}} = 2[\boldsymbol{c_2 c_0}], \quad \ldots \quad \left.\frac{\partial\Phi}{\partial b_k}\right|_{\boldsymbol{b}=\boldsymbol{b}^{(s)}} = 2[\boldsymbol{c_k c_0}]. \tag{4.15}$$

Por consiguiente, sustituyendo (4.13) y (4.15) en (4.14) resulta

$$\left.\frac{\mathrm{d}\Phi}{\mathrm{d}w}\right|_{w=0} = -2\left([\boldsymbol{c_1 c_0}]^2 + [\boldsymbol{c_2 c_0}]^2 + \ldots + [\boldsymbol{c_k c_0}]^2\right). \tag{4.16}$$

Esta derivada es negativa puesto que las derivadas parciales en (4.15) no son todas cero de acuerdo al supuesto de que Φ no es estacionaria en $\boldsymbol{b}^{(s)}$. De este modo, Φ en función de w es decreciente en $w = 0$, asegurando así que existan valores de w para los que el valor de Φ se reduce respecto al valor $\Phi(\boldsymbol{b}^{(s)})$.

Existe otra forma de llegar a esta última conclusión. Por definición, la gradiente de $\bar{\Phi}_{\Delta w}$ es cero en el punto $\boldsymbol{\delta}_{\Delta w}$, esto es

$$\Delta w \nabla\langle\Phi(\boldsymbol{\delta}_{\Delta w})\rangle + \nabla Q(\boldsymbol{\delta}_{\Delta w}) = 0, \tag{4.17}$$

y dado que $\boldsymbol{\delta}_0 = \mathbf{0}$, dicho punto puede expresarse como

$$\boldsymbol{\delta}_{\Delta w} = \left[\frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w}(0) + (\Delta w)\right]\Delta w, \tag{4.18}$$

donde (Δw) es un vector cuyo límite es cero cuando Δw tiende a cero. Por otro lado, de (4.6) se desprende

$$\nabla Q(\boldsymbol{\delta}_{\Delta w}) = 2\,\boldsymbol{\delta}_{\Delta w}, \tag{4.19}$$

de modo que introduciendo (4.18) en (4.19), y este a su vez en (4.17), resulta

$$\frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w}(0) = -\frac{1}{2}\nabla\langle\Phi(\boldsymbol{\delta}_{\Delta w})\rangle - (\Delta w). \tag{4.20}$$

Puesto que esta expresión es válida para cualquier valor positivo de Δw, entonces también lo será en el límite cuando tiende a cero, con lo cual se obtiene

$$\frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w}(0) = -\frac{1}{2}\nabla\langle\Phi(\mathbf{0})\rangle, \tag{4.21}$$

que no es otra cosa que (4.13).

Si bien ya se demostró que

$$0 = \|\boldsymbol{\delta}_0\| \,<\, \|\boldsymbol{\delta}_w\| \,<\, \|\boldsymbol{\delta}_\infty\| = \|\boldsymbol{\delta}^*\|,$$

es conveniente analizar cómo varía $\|\boldsymbol{\delta}_w\|$ según se incrementa w. Con ese fin, tómese como punto de partida la siguiente expresión obtenida por definición:

$$(w+\Delta w)\nabla\langle\Phi(\boldsymbol{\delta}_{w+\Delta w})\rangle + \nabla Q(\boldsymbol{\delta}_{w+\Delta w}) = \mathbf{0}, \tag{4.22}$$

la cual para valores infinitamente pequeños de Δw toma la forma

$$(w+\Delta w)\nabla\Big\langle\Phi\Big(\boldsymbol{\delta}_w + \frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w}\Delta w\Big)\Big\rangle + \nabla Q\Big(\boldsymbol{\delta}_w + \frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w}\Delta w\Big) = \mathbf{0},$$

que es a la vez equivalente a

$$(w+\Delta w)\left\{\nabla\langle\Phi(\boldsymbol{\delta}_w)\rangle + \Big[\nabla\nabla^T\langle\Phi(\boldsymbol{\delta}_w)\rangle\Big]\frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w}\Delta w\right\} + \left\{\nabla Q(\boldsymbol{\delta}_w) + \Big[\nabla\nabla^T Q(\boldsymbol{\delta}_w)\Big]\frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w}\Delta w\right\} = \mathbf{0}.$$

Introduciendo (4.6) y (4.9) en esta última ecuación, y llevándola al límite $\Delta w = 0$, se obtiene

$$\Big[w\nabla\nabla^T\bar{\Phi}_w(\boldsymbol{\delta}_w)\Big]\,\frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w} - \nabla Q(\boldsymbol{\delta}_w) = \mathbf{0},$$

que, tras aplicarle el producto interno con $\mathrm{d}\boldsymbol{\delta}_w/\mathrm{d}w$ a ambos miembros, se transforma en

$$\frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w}\cdot\nabla Q(\boldsymbol{\delta}_w) = \frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w}\cdot\left[\bar{A}\,\frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w}\right], \qquad \text{donde}: \bar{A} = w\nabla\nabla^T\bar{\Phi}_w(\boldsymbol{\delta}_w). \tag{4.23}$$

La matriz $\bar{A}$ es el producto entre el factor positivo w y el hessiano de $\bar{\Phi}_w$

evaluado en un mínimo.

Por otro lado, la segunda derivada de una función multivariable arbitraria G en un punto P y una dirección $\hat{\boldsymbol{v}}$ es

$$\frac{\mathrm{d}^2 G}{\mathrm{d}v^2}(P) = \hat{\boldsymbol{v}}^T H(P)\hat{\boldsymbol{v}}, \tag{4.24}$$

donde H es el hessiano de G. Por lo tanto, si en P la función G tiene un mínimo, necesariamente (4.24) es positiva para cualquier vector dirección $\hat{\boldsymbol{v}}$.

Cabe aquí indicar que a toda matriz M que cumple la condición

$$\hat{\boldsymbol{v}}^T M\, \hat{\boldsymbol{v}} > 0,$$

para cualquier $\hat{\boldsymbol{v}}$, se le denomina matriz positiva definida(positive definite) [11]. Luego, como el producto de una matriz positiva definida y una constante positiva sigue siendo una matriz positiva definida, se concluye que $\bar{A}$ es positiva definida, de manera que (4.23) implica

$$\frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w} \cdot \nabla Q(\boldsymbol{\delta}_w) > 0. \tag{4.25}$$

Para determinar la variación de $\|\boldsymbol{\delta}_w\|$ respecto a w, basta con analizar el comportamiento de su derivada a lo largo de su dominio. Una manera de simplicar el cálculo de la misma, consiste en expresarla en función de la derivada de $\|\boldsymbol{\delta}_w\|^2$, así se tiene

$$\frac{\mathrm{d}\|\boldsymbol{\delta}_w\|^2}{\mathrm{d}w} = 2\,\|\boldsymbol{\delta}_w\|\frac{\mathrm{d}\|\boldsymbol{\delta}_w\|}{\mathrm{d}w}, \tag{4.26}$$

y puesto que

$$\|\boldsymbol{\delta}_w\|^2 = \boldsymbol{\delta}_w \cdot \boldsymbol{\delta}_w, \tag{4.27}$$

resulta

$$\frac{\mathrm{d}\|\boldsymbol{\delta}_w\|^2}{\mathrm{d}w} = 2\,\frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w} \cdot \boldsymbol{\delta}_w. \tag{4.28}$$

Introduciendo (4.28) en (4.26) se obtiene la expresión buscada

$$\frac{\mathrm{d}\|\boldsymbol{\delta}_w\|}{\mathrm{d}w} = \frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w} \cdot \hat{\boldsymbol{\delta}}_w, \tag{4.29}$$

la cual, tras considerar que $\nabla Q(\boldsymbol{\delta}_w)$ es paralela a $\hat{\boldsymbol{\delta}}_w$, toma la forma

$$\frac{\mathrm{d}\|\boldsymbol{\delta}_w\|}{\mathrm{d}w} = \|\nabla Q(\boldsymbol{\delta}_w)\|^{-1} \frac{\mathrm{d}\boldsymbol{\delta}_w}{\mathrm{d}w} \cdot \nabla Q(\boldsymbol{\delta}_w). \tag{4.29a}$$

De este resultado y de (4.25) se deduce

$$\frac{\mathrm{d}\|\boldsymbol{\delta}_w\|}{\mathrm{d}w} > 0, \tag{4.30}$$

para todo w positivo. Con esto queda finalmente demostrado que el módulo del vector $\boldsymbol{\delta}_w$ es estrictamente creciente en w.

4.1.3. Cálculo del Vector de Paso

Que el método logre reducir el valor de Φ en cada iteración depende directamente del valor de w que se elija. Por esta razón es importante analizar qué valores de w son los adecuados.

Teóricamente, el mejor valor de w a elegir es aquel que satisface

$$\frac{\mathrm{d}\Phi}{\mathrm{d}w}(\boldsymbol{b}_w) = 0 \qquad \text{donde} : \boldsymbol{b}_w = \boldsymbol{b}^{(s)} + \boldsymbol{\delta}_w; \tag{4.31}$$

sin embargo, resolver esta ecuación en la práctica es un problema complejo.

Una estrategia, utilizada por Cauchy [3] para abordar este tipo de problemas, consiste en plantear la aproximación

$$\Phi(\boldsymbol{b}_w) \approx \Phi(\boldsymbol{b}^{(s)}) + w \left(\frac{\mathrm{d}\Phi}{\mathrm{d}w}\right)\Bigg|_{w=0}. \tag{4.32}$$

Bajo el supuesto de que $\boldsymbol{b}^{(s)}$ es tal que origina que el valor de $\Phi(\boldsymbol{b}_w)$ sea pequeño, el miembro de la izquierda en (4.32) puede considerarse igual a cero, con lo cual resulta la fórmula

$$w \cong -\left.\frac{\Phi(\boldsymbol{b}^{(s)})}{(\mathrm{d}\Phi/\mathrm{d}w)}\right|_{w=0} = \frac{1}{2} \frac{\Phi(\boldsymbol{b}^{(s)})}{[\boldsymbol{c_1 c_0}]^2 + [\boldsymbol{c_2 c_0}]^2 + \ldots + [\boldsymbol{c_k c_0}]^2}, \tag{4.33}$$

que constituye una buena opción para w bajo las condiciones impuestas.

4.2. Enfoque de Marquardt

Posterior a Levenberg y de manera independiente, Marquardt concibió prácticamente el mismo método propuesto por Levenberg, aunque siguiendo una línea de razonamiento distinta.

El desarrollo de esta sección es hasta cierto punto una reproducción del trabajo original de Marquardt [21]. Cabe señalar que la notación adoptada en este y en los dos capítulos anteriores fue planeada con el fin de lograr la mayor compatibilidad posible con la notación utilizada por Marquardt.

4.2.1. Base Teórica del Método

Los siguientes tres teoremas constituyen el fundamento teórico en el que se basa el método. En ellos se busca hacer énfasis en las relaciones geométricas involucradas.

Teorema 1. *Si λ es un real positivo arbitrario y $\boldsymbol{\delta}_0$ satisface la ecuación*

$$(A + \lambda I)\boldsymbol{\delta} = \boldsymbol{g}, \tag{4.34}$$

entonces $\boldsymbol{\delta}_0$ minimiza a $\langle \Phi \rangle$ en la esfera cuyo radio $\|\boldsymbol{\delta}\|$ satisface

$$\|\boldsymbol{\delta}\|^2 = \|\boldsymbol{\delta}_0\|^2. \tag{4.35}$$

Demostración. Antes de abordar la demostración de este caso específico es conveniente analizar el problema más general. Este consiste en determinar el punto que minimiza la función $G(\boldsymbol{\delta})$ sobre la isosuperficie $H(\boldsymbol{\delta}) = 0$, donde $G(\boldsymbol{\delta}) : \mathbf{R}^k \to \mathbf{R}$ y $H(\boldsymbol{\delta}) : \mathbf{R}^k \to \mathbf{R}$ son doblemente diferenciables.

Condición necesaria para el punto mínimo buscado es que la gradiente de G en dicho punto sea ortogonal a la isosuperficie de H en él o, expresado de otra manera, que sea paralela a la gradiente de H en tal punto. Por lo tanto, el punto mínimo debe ser solución de

$$\nabla G(\boldsymbol{\delta}) + \lambda \nabla H(\boldsymbol{\delta}) = 0 \qquad \text{y} \qquad H(\boldsymbol{\delta}) = 0,$$

donde el valor particular del parámetro λ debe seleccionarse de manera que implique el cumplimiento de la segunda ecuación de restricción. Para calcular tal valor se obtiene $\boldsymbol{\delta}$ en función de λ a partir de la primera ecuación.

Este resultado se introduce en la ecuación de restricción que toma la forma $H(\boldsymbol{\delta}(\lambda)) = 0$, de donde finalmente puede despejarse λ.

A este procedimiento de minimización con restricción se le denomina *método de los multiplicadores de Lagrange* y al parámetro λ, *multiplicador de Lagrange*.

Aplicando este método al caso específico en el que la función a minimizar es

$$\langle \Phi(\boldsymbol{\delta}) \rangle = \|\boldsymbol{y} - \boldsymbol{f}_0 - P\boldsymbol{\delta}\|^2, \tag{4.36}$$

bajo la restricción

$$\|\boldsymbol{\delta}\|^2 = \|\boldsymbol{\delta}_0\|^2, \tag{4.37}$$

se obtiene para el punto mínimo la condición necesaria

$$\frac{\partial \tilde{\Phi}}{\partial \delta_1} = \frac{\partial \tilde{\Phi}}{\partial \delta_2} = \cdots = \frac{\partial \tilde{\Phi}}{\partial \delta_k} = 0, \qquad \frac{\partial \tilde{\Phi}}{\partial \lambda} = 0, \tag{4.38}$$

donde

$$\tilde{\Phi}(\boldsymbol{\delta}, \lambda) = \|\boldsymbol{y} - \boldsymbol{f}_0 - P\boldsymbol{\delta}\|^2 + \lambda(\|\boldsymbol{\delta}\|^2 - \|\boldsymbol{\delta}_0\|^2). \tag{4.39}$$

Tomando las derivadas indicadas se tiene

$$-[P^T(\boldsymbol{y} - \boldsymbol{f}_0) - P^T P\boldsymbol{\delta}] + \lambda\boldsymbol{\delta} = \mathbf{0} \tag{4.40}$$

y

$$\|\boldsymbol{\delta}\|^2 - \|\boldsymbol{\delta}_0\|^2 = 0. \tag{4.41}$$

Si se denota la solución de (4.40) como $\boldsymbol{\delta}_0$, entonces se cumple

$$(P^T P + \lambda I)\boldsymbol{\delta}_0 = P^T(\boldsymbol{y} - \boldsymbol{f}_0), \tag{4.42}$$

que es equivalente a

$$(A + \lambda I)\boldsymbol{\delta}_0 = \boldsymbol{g}. \tag{4.43}$$

Demostrar que $\boldsymbol{\delta}_0$ cumple (4.41) es trivial.

Finalmente, el que el punto estacionario $\boldsymbol{\delta}_0$ sea efectivamente un mínimo se concluye del hecho que A es positiva definida, por ser proporcional al hessiano de $\langle \Phi \rangle$, y $\lambda \geq 0$ (véase pág. 78). □

Teorema 2. *Si $\boldsymbol{\delta}(\lambda)$ es la solución de* (4.34) *para un valor de λ dado, en-*

tonces $\|\boldsymbol{\delta}(\lambda)\|^2$ *es una función monótona decreciente continua de* λ*, tal que cuando* $\lambda \to \infty$*,* $\|\boldsymbol{\delta}(\lambda)\|^2 \to 0$*.*

Demostración. Siempre que los eigenvectores de una matriz A sean linealmente independientes, existe una transformación de similaridad (similarity transformation) que reduce A a una matriz diagonal D. Esto es, existe una matriz no singular M cuyas columnas son precisamente los eigenvectores de A, de manera que

$$M^{-1}AM = D,$$

donde los elementos de D son los eigenvalores de A [30]. Puesto que en el presente caso A es simétrica, sus eigenvalores son distintos entre sí, originando que la matriz M sea ortonormal y represente una rotación de coordenadas, con lo cual $M^TM = I$. Más aún, al ser A positiva definida, se cumple que todos los elementos de D son positivos.

Premultiplicando (4.34) por M^{-1} resulta

$$M^{-1}(A + \lambda I)MM^{-1}\boldsymbol{\delta} = M^{-1}\boldsymbol{g},$$

de modo que

$$\boldsymbol{\delta} = M(D + \lambda I)^{-1}M^{-1}\boldsymbol{g}. \tag{4.44}$$

Luego, definiendo $\boldsymbol{v} = M^{-1}\boldsymbol{g}$, se tiene

$$\begin{aligned} \|\boldsymbol{\delta}(\lambda)\|^2 &= \boldsymbol{v}^T[(D + \lambda I)^{-1}]^T M^T M(D + \lambda I)^{-1}\boldsymbol{v} \\ &= \boldsymbol{v}^T[(D + \lambda I)^2]^{-1}\boldsymbol{v} \\ &= \sum_{j=1}^{k} \frac{v_j^2}{(D_j + \lambda)^2}, \end{aligned} \tag{4.45}$$

la cual claramente es una función decreciente en λ, (con $\lambda \geq 0$), tal que cuando $\lambda \to \infty$, $\|\boldsymbol{\delta}(\lambda)\|^2 \to 0$.

La transformación ortonormal hacia una matriz diagonal ha sido explícitamente mostrada con el fin de facilitar la prueba del siguiente teorema. □

Teorema 3. *Si* γ *es el ángulo entre* $\boldsymbol{\delta}(\lambda)$ *y* $-\nabla\Phi$*, entonces* γ *es una función monótona decreciente continua de* λ*, tal que cuando* $\lambda \to \infty$*,* $\gamma \to 0$*. Puesto que* $-\nabla\Phi$ *es independiente de* λ*, significa que* $\boldsymbol{\delta}(\lambda)$ *rota hacia* $-\nabla\Phi$ *conforme* $\lambda \to \infty$*.*

Demostración. Obsérvese que $-\nabla\Phi = 2\boldsymbol{g}$, de manera que γ es igualmente el ángulo que forman $\boldsymbol{\delta}(\lambda)$ y $\boldsymbol{g}$.

Por definición se tiene

$$\begin{aligned}\cos\gamma &= \frac{\boldsymbol{\delta}(\lambda)^T\boldsymbol{g}}{(\|\boldsymbol{\delta}\|)(\|\boldsymbol{g}\|)} \\ &= \frac{\boldsymbol{v}^T(D+\lambda I)^{-1}\boldsymbol{v}}{(\boldsymbol{v}^T[(D+\lambda I)^2]^{-1}\boldsymbol{v})^{1/2}(\boldsymbol{g}^T\boldsymbol{g})^{1/2}} \\ &= \frac{\displaystyle\sum_{j=1}^{k}\frac{v_j^2}{(D_j+\lambda)}}{\left[\displaystyle\sum_{j=1}^{k}\frac{v_j^2}{(D_j+\lambda)^2}\right]^{1/2}(\boldsymbol{g}^T\boldsymbol{g})^{1/2}}.\end{aligned} \tag{4.46}$$

Derivando y simplificando, resulta

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\cos\gamma = \frac{\left[\displaystyle\sum_{j=1}^{k}\frac{v_j^2}{(D_j+\lambda)}\right]\left[\displaystyle\sum_{j=1}^{k}\frac{v_j^2}{(D_j+\lambda)^3}\right] - \left[\displaystyle\sum_{j=1}^{k}\frac{v_j^2}{(D_j+\lambda)^2}\right]^2}{\left[\displaystyle\sum_{j=1}^{k}\frac{v_j^2}{(D_j+\lambda)^2}\right]^{3/2}(\boldsymbol{g}^T\boldsymbol{g})^{1/2}} \tag{4.47}$$

$$= \frac{\left[\displaystyle\sum_{j=1}^{k}v_j^2\prod{}_{1j}\right]\left[\displaystyle\sum_{j=1}^{k}v_j^2\prod{}_{3j}\right] - \left[\displaystyle\sum_{j=1}^{k}v_j^2\prod{}_{2j}\right]^2}{\left[\displaystyle\sum_{j=1}^{k}\frac{v_j^2}{(D_j+\lambda)^2}\right]^{3/2}\left[\displaystyle\prod_{j=1}^{k}(D_j+\lambda)^2\right](\boldsymbol{g}^T\boldsymbol{g})^{1/2}}, \tag{4.48}$$

donde

$$\prod{}_{1j} = \prod_{\substack{h=1\\h\neq j}}^{k}(D_h+\lambda), \qquad \prod{}_{2j} = \prod_{\substack{h=1\\h\neq j}}^{k}(D_h+\lambda)^2 \qquad \prod{}_{3j} = \prod_{\substack{h=1\\h\neq j}}^{k}(D_h+\lambda)^3.$$

Puesto que el denominador en (4.48) está compuesto por factores positivos, el signo de $\mathrm{d}(\cos\gamma)/\mathrm{d}\lambda$ es entonces el signo del numerador. Notando que

$$\prod{}_{1j}\prod{}_{3j} = (\prod{}_{2j})^2,$$

el numerador se puede escribir como

$$\left[\sum_{j=1}^{k}(v_j\prod{}_{1j}^{1/2})^2\right]\left[\sum_{j=1}^{k}(v_j\prod{}_{3j}^{1/2})^2\right]-\left[\sum_{j=1}^{k}(v_j\prod{}_{1j}^{1/2})(v_j\prod{}_{3j}^{1/2})\right]^2. \qquad (4.49)$$

De la desigualdad de Schwarz se desprende que (4.49) es mayor que cero, por lo tanto $\mathrm{d}(\cos\gamma)/\mathrm{d}\lambda$ es siempre positiva para cualquier $\lambda > 0$. De este resultado se concluye que γ es una función monótona decreciente de λ.

Por otro lado, para valores muy grandes de λ, la matriz $(A + \lambda I)$ es dominada por la diagonal λI. Por lo tanto de (4.34) se observa que cuando $\lambda \to \infty$, $\boldsymbol{\delta}(\lambda) \to \boldsymbol{g}/\lambda$, de manera que en el límite el ángulo entre $\boldsymbol{\delta}$ y $\boldsymbol{g}$ tiende a cero. Para el caso en que $\lambda = 0$ en (4.34), los vectores $\boldsymbol{\delta}$ y $\boldsymbol{g}$ forman un ángulo $0 < \gamma < \pi/2$, excepto en el caso trivial cuando A es la matriz identidad. □

4.2.2. Escalamiento del Espacio de Parámetros

Un aspecto numérico importante a considerar en el procedimiento de solución de sistemas de ecuaciones lineales del tipo $A\,\boldsymbol{\delta} = \boldsymbol{g}$ es el grado de sensibilidad que poseen las matrices A y $\boldsymbol{g}$ respecto a perturbaciones en sus elementos. En ese sentido se dice que A y $\boldsymbol{g}$ son muy sensibles, o están pobrememnte acondicionadas, cuando una pequeña perturbación en ellas produce una gran desviación en el vector solución del sistema $\boldsymbol{\delta}$.

Con el fin de lograr que el vector de parámetros $\boldsymbol{\delta}$ obtenido en cada iteración posea un buen nivel de inmunidad a las posibles perturbaciones producidas por errores presentes en las matrices del sistema A y $\boldsymbol{g}$, y puesto que el método es invariante respecto a transformaciones lineales de coordenadas, es conveniente llevar a cabo un escalamiento adecuado del espacio de parámetros $\boldsymbol{b}$.

En particular, escalando cada componente b_j de $\boldsymbol{b}$ en unidades de la desviación estándar de las derivadas $\partial f_i/\partial b_j$ tomadas sobre todas las muestras $(i = 1, 2, \ldots, n)$, se obtiene una transformación que convierte a la matriz A en la matriz de coeficientes de correlación normalizados de los $\partial f_i/\partial b_j$. La ventaja de este escalamiento es que minimiza la sensibilidad de A y $\boldsymbol{g}$, consiguiendo así un acondicionamiento óptimo del sistema de ecuaciones.

En el nuevo espacio de parámetros $\boldsymbol{b}^\star$ se tiene que la matriz escalada $A^\star$

y el vector escalado $\boldsymbol{g}^{\star}$ poseen la forma

$$A^{\star} = (a^{\star}_{jj'}) = \left(\frac{a_{jj'}}{\sqrt{a_{jj}}\sqrt{a_{j'j'}}} \right), \tag{4.50}$$

$$\boldsymbol{g}^{\star} = (g^{\star}_{j}) = \left(\frac{g_{j}}{\sqrt{a_{jj}}} \right), \tag{4.51}$$

de manera que el vector de paso $\boldsymbol{\delta}^{\star}$ según el método de Gauss-Newton en el nuevo espacio de parámetros se obtiene resolviendo

$$A^{\star}\, \boldsymbol{\delta}^{\star} = \boldsymbol{g}^{\star}. \tag{4.52}$$

El correspondiente vector de paso $\boldsymbol{\delta}$ en el espacio de parámetros original se obtiene a partir de su versión $\boldsymbol{\delta}^{\star}$ mediante la relación

$$\delta_{j} = \frac{\delta^{\star}_{j}}{\sqrt{a_{jj}}}. \tag{4.53}$$

4.2.3. Construcción del Algoritmo

El bosquejo general del algoritmo para implementar el método queda finalmente delineado. Específicamente, para la s-ésima iteración se construye la ecuación

$$(A^{\star(s)} + \lambda^{(s)} I)\, \boldsymbol{\delta}^{\star(s)} = \boldsymbol{g}^{\star(s)}, \tag{4.54}$$

la cual equivale a minimizar $\langle \Phi \rangle$ en el espacio de parámetros original, pero esta vez no sobre la esfera dada en (4.35), sino sobre una elipse de la forma

$$\sum_{j=1}^{k} a_{jj} \delta_{j}^{2} = \text{cte.} \tag{4.55}$$

A partir del $\boldsymbol{\delta}^{(s)}$ obtenido al introducir en (4.53) la solución $\boldsymbol{\delta}^{\star(s)}$ de la ecuación (4.54), se genera el nuevo punto tentativo

$$\boldsymbol{b}^{(s+1)} = \boldsymbol{b}^{(s)} + \boldsymbol{\delta}^{(s)}, \tag{4.56}$$

el cual conduce a un nuevo valor de $\Phi^{(s+1)}$. Cabe mencionar que es esencial seleccionar un valor de $\lambda^{(s)}$ que conduzca al resultado

$$\Phi^{(s+1)} < \Phi^{(s)}. \tag{4.57}$$

De la teoría precedente se puede afirmar que siempre existirá un $\lambda^{(s)}$ suficientemente grande como para satisfacer (4.57), a menos que $\boldsymbol{b}^{(s)}$ sea ya un mínimo de Φ. Para determinar el valor de $\lambda^{(s)}$ que satisfaga (4.57), y que a la vez produzca una rápida convergencia del algoritmo, se requiere un tanteo del tipo ensayo error.

En cada iteración se desea minimizar Φ dentro de la máxima vecindad posible sobre la cual la versión linealizada del modelo ofrezca una adecuada representación de la función no lineal. Por consiguiente, la estrategia para elegir $\lambda^{(s)}$ debe buscar utilizar un valor pequeño de $\lambda^{(s)}$ siempre que se den la condiciones que permitan una buena convergencia del método estándar de Gauss-Newton. Esto es especialmente relevante en los últimos pasos del procedimiento de convergencia, cuando los puntos tentativos están en la vecindad inmediata del mínimo $\boldsymbol{b}_{min}$, donde los contornos de Φ son asintóticamente elípticos, y la expansión lineal del modelo necesita ser una buena aproximación solo sobre una pequeña región.

Por otro lado, valores grandes de $\lambda^{(s)}$ solo deben ser utilizados cuando sean necesarios para satisfacer (4.57). Si bien es cierto que $\Phi^{(s+1)}$ como función de λ posee un mínimo, y que la elección de dicho λ en la $(s+1)-$ésima iteración maximiza $(\Phi^{(s)} - \Phi^{(s+1)})$, tal elección óptima localmente resulta ser una estrategia pobre globalmente, ya que esta por lo general requiere un valor substancialmente mayor de λ del que es necesario para satisfacer (4.57). Esta estrategia hereda por ende la propiedad del método del máximo descenso de tener un progreso inicial rápido seguido de uno progresivamente más lento.

En consecuencia, la estrategia se define de la siguiente manera:

1. Definir el factor de incremento/decremento $\nu > 1$.
2. Asignar un valor inicial a λ; por ejemplo: $\lambda^{(0)} = 10^{-2}$.
3. Para la $(s+1)-$ésima iteración:
   a) Calcular $\Phi(\lambda^{(s)})$ y $\Phi(\lambda^{(s)}/\nu)$.

b) Si $\Phi(\lambda^{(s)}/\nu) \leq \Phi^{(s)}$, tomar $\lambda^{(s+1)} = \lambda^{(s)}/\nu$.

c) Si $\Phi(\lambda^{(s)}/\nu) > \Phi^{(s)}$ y $\Phi(\lambda^{(s)}) \leq \Phi^{(s)}$, tomar $\lambda^{(s+1)} = \lambda^{(s)}$.

d) Si $\Phi(\lambda^{(s)}/\nu) > \Phi^{(s)}$ y $\Phi(\lambda^{(s)}) > \Phi^{(s)}$, incrementar λ multiplicándolo sucesivamente por ν hasta que para el menor u se cumpla: $\Phi(\lambda^{(s)}\nu^u) \leq \Phi^{(s)}$. Tomar $\lambda^{(s+1)} = \lambda^{(s)}\nu^u$.

En relación al punto *d*), es necesario hacer algunos comentarios complementarios.

En ocasiones, en problemas en los que la correlación entre los sucesivos puntos tentativos $\boldsymbol{b}^{(s)}$ es extremadamente alta ($> 0{,}99$), puede ocurrir que el valor de λ se incremente hasta valores inaceptablemente altos. Para tales casos es preferible modificar la prueba en *d*) como se detalla a continuación.

Sea

$$\boldsymbol{b}^{(s+1)} = \boldsymbol{b}^{(s)} + K^{(s)}\boldsymbol{\delta}^{(s)}, \qquad K^{(s)} \leq 1. \tag{4.58}$$

Notando que el ángulo $\gamma^{(s)}$ (definido en la pág 82) es función decreciente de $\lambda^{(s)}$, se selecciona un ángulo umbral $\gamma_0 < \pi/2$ y se fija

$$K^{(s)} = 1, \qquad \text{si} \qquad \gamma^{(s)} \geq \gamma_0. \tag{4.59}$$

Si no es posible superar la prueba *d*) a pesar de haber incrementado $\lambda^{(s)}$ hasta haber conseguido que se cumpla $\gamma^{(s)} < \gamma_0$, entonces se deja de incrementar $\lambda^{(s)}$ y se toma un $K^{(s)}$ lo suficientemente pequeño como para que se cumpla $\Phi^{(s+1)} < \Phi^{(s)}$. Esto es siempre posible puesto que $\gamma^{(s)} < \gamma_0 < \pi/2$. Una buena elección para el ángulo umbral es $\gamma_0 = \pi/4$.

Nótese que el hecho que $\cos\gamma$ sea positivo cuando $\lambda = 0$ solo se puede garantizar cuando la matriz $A^\star$ es positiva definida. En presencia de niveles de correlación muy grandes, el grado en que $A^\star$ es positiva definida es tenue y puede degradarse notoriamente debido a errores de redondeo presentes en las operaciones de punto flotante realizadas en los cálculos por computadora. En tales casos, el método puede llegar a divergir indistintamente del valor de $K^{(s)}$ utilizado. Un beneficio asociado al hecho de que el método le sume λ a la diagonal de $A^\star$, es que se garantiza que $\cos\gamma$ sea positivo aun en el caso que $A^\star$ sea débilmente positiva definida.

A pesar de que no siempre es posible prever la ocurrencia de altas correlaciones entre los parámetros tentativos $\boldsymbol{b}^{(s)}$ relacionados a modelos no

lineales, con frecuencia sí es posible reducir substancialmente dichos niveles de correlación mediante un mejor diseño experimental en la producción de los datos medidos.

Numéricamente, se considera que el proceso de iteración ha convergido cuando se cumple la condición

$$\frac{|\delta_j^{(s)}|}{\tau + |b_j^{(s)}|} < \epsilon, \qquad j = 1, 2, \ldots, k \tag{4.60}$$

para algún $\epsilon > 0$ suficientemente pequeño (como 10^{-5}) y algún τ adecuado (como 10^{-3}). Si bien la elección de ν es arbitraria, se ha encontrado en la práctica que $\nu = 10$ es una buena elección.

4.3. Similitud y Diferencia entre los Enfoques de Levenberg y Marquardt

Si bien ambos enfoques fueron concebidos mediante razonamientos distintos, los dos son versiones de un mismo método, lo cual se evidencia al considerar la equivalencia $w^{-1} = \lambda$ entre los parámetros de magnificación-atenuación 'w' y 'λ' utlizados por Levenberg y Marquardt respectivamente.

Asimismo, la diferencia entre ambos enfoques radica en las estrategias distintas utilizadas en la determinación de los parámetros w y λ, con el fin de minimizar Φ en cada iteración. Esto produce que la razón de convergencia no sea la misma para cada caso.

En el enfoque de Levenberg, la estrategia consiste en minimizar Φ localmente como una función del parámetro w^{-1} que se le suma a la diagonal, con lo cual se consigue una razón de convergencia lineal

$$\|\boldsymbol{b}^{(s+1)} - \boldsymbol{b}_{min}\| \leq C\|\boldsymbol{b}^{(s)} - \boldsymbol{b}_{min}\|, \qquad C < 1, \tag{4.61}$$

como la que presenta el método del máximo descenso.

Por otro lado, en el enfoque de Marquardt, la estrategia consiste en minimizar Φ en la mayor vecindad posible sobre la cual la aproximaxión lineal sea una representación adecuada. Esta estrategia busca por ende el menor valor de λ que satisfaga (4.57), con lo cual se obtiene una razón de convergencia

cuadrática

$$\|\boldsymbol{b}^{(s+1)} - \boldsymbol{b}_{min}\| \leq C\|\boldsymbol{b}^{(s)} - \boldsymbol{b}_{min}\|^2, \qquad C < 1, \tag{4.62}$$

similar a la que posee el método de Gauss-Newton.

Capítulo 5

Estimación de Parámetros y Conclusiones

El presente capítulo muestra los resultados y conclusiones de la aplicación del método de Levenberg–Marquardt al caso específico del modelo TGM introducido en el primer capítulo para la estimación de sus parámetros de mejor ajuste según el criterio de los mínimos cuadrados.

5.1. Procedimiento de Estimación de Parámetros

El procedimiento de estimación de parámetros se lleva a cabo en dos etapas.

En la primera etapa se ajusta la impedancia mecánica del modelo a los datos experimentales. El algoritmo utilizado para ello es una versión optimizada del método de Levenberg-Marquardt [23], [8], analizado en el capítulo anterior.

De (1.5) se puede reconocer que los parámetros r_2 y c_2 introducidos por el modelo TGM tienen una mayor influencia en la parte real de la impedancia mecánica a bajas frecuencias. Tomando en cuenta este hecho, los parámetros r_1, r_2, y c_2 son estimados ajustando el modelo de la parte real de la impedancia mecánica a los datos medidos en el rango que va de los 5 Hz hasta dos veces la frecuencia de resonancia. Los otros dos parámetros m y c_1 se obtienen ajustando el modelo de la parte imaginaria de la impedancia mecánica a los

datos medidos en el intervalo de frecuencias cuyos límites resultan de dividir y multiplicar la frecuencia de resonancia por un factor de 1.3. Este angosto rango de frecuencias alrededor de la frecuencia de resonancia se encontró que era el óptimo para determinar estos parámetros, de forma tal que el punto de cruce por cero del modelo de la parte imaginaria de la impedancia mecánica coinciden con la frecuencia de resonancia real.

En la segunda etapa se refinan los parámetros obtenidos en la etapa anterior con el fin de conseguir que el modelo coincida con los datos medidos en ciertos puntos claves de la curva de magnitud de la mobilidad mecánica (definida como la inversa de la magnitud de la impedancia mecánica), tales como el punto de resonancia o de máxima amplitud y los puntos que están 3 dB por debajo de este.

Para este propósito se define la función

$$F_i = \frac{1}{\sqrt{\left(r_1 + \frac{r_2}{1+(\omega_i c_2 r_2)^2}\right)^2 + \left((\omega_i m - \frac{1}{\omega_i c_1}) - \frac{\omega_i c_2 r_2^2}{1+(\omega_i c_2 r_2)^2}\right)^2}} - x_i, \tag{5.1}$$

donde el par (ω_i, x_i) compuesto por la magnitud de la mobilidad x_i medida a la frecuencia ω_i representa un punto de coincidencia. Los parámetros para lograr la coincidencia son entonces ceros de la función F_i. Ya que se tienen tres puntos de coincidencia y cinco parámetros, calcular estos últimos consiste en resolver un sistema de ecuaciones subdeterminado y en general no lineal.

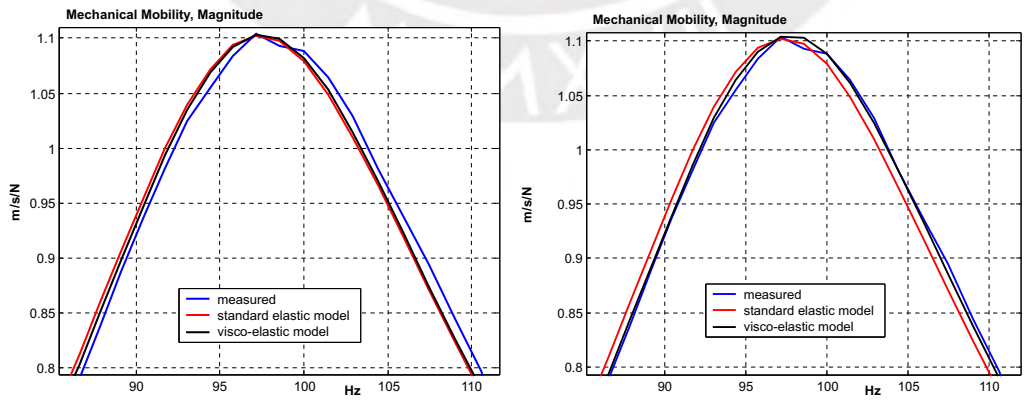


Figura 5.1: Mobilidad mecánica medida y simulada alrededor de la frecuencia de resonancia. Izquierda: sin proceso de refinamiento de parámetros. Derecha: con parámetros refinados

Para reducir el número de variables es conveniente excluir aquellas cuyas variaciones producen un menor efecto en (5.1). Tras una serie de pruebas, se determinó que la mayor sensibilidad se presentaba ante cambios en los parámetros m, c_1 y c_2.

Al final se determinan los ceros mediante un algoritmo basado en el método de Newton–Raphson.

La mejora alcanzada en esta etapa se muestra en la figura 5.1.

La implementación práctica del método descrito en este trabajo se llevó a cabo mediante un programa especialmente desarrollado para tal fin. El programa controla el analizador de señales, automatiza el proceso de medición, procesa los datos medidos, realiza el ajuste según el Método de Levenberg–Marquardt, y muestra los parámetros resultantes.

En la figura 5.2 se pueden ver los resultados medidos y simulados de un altavoz JBL 2206, en los que se observa una muy buena concordancia desde las bajas frecuencias hasta una frecuencia de aproximadamente 1,3 veces la frecuencia de resonancia. Esto demuestra que el modelo TGM es adecuado.

Los resultados, tal y como los muestra el programa, pueden verse en la figura 5.3. Nótese que el recuadro superior corresponde a los parámetros Thiele–Small definidos por el modelo tradicional (también calculados por el programa), mientras que el inferior corresponde a los parámetros definidos por el modelo TGM.

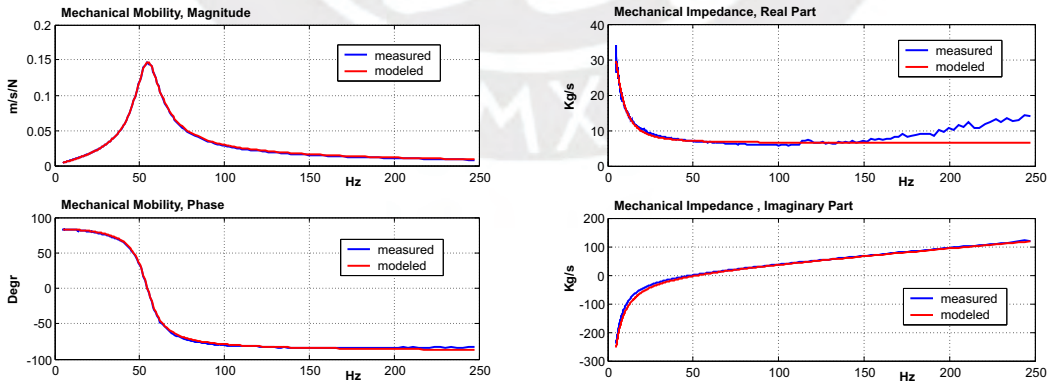


Figura 5.2: Mobilidad mecánica medida y simulada de un altavoz JBL 2206. Izquierda: magnitud y fase. Derecha: parte real y parte imaginaria.

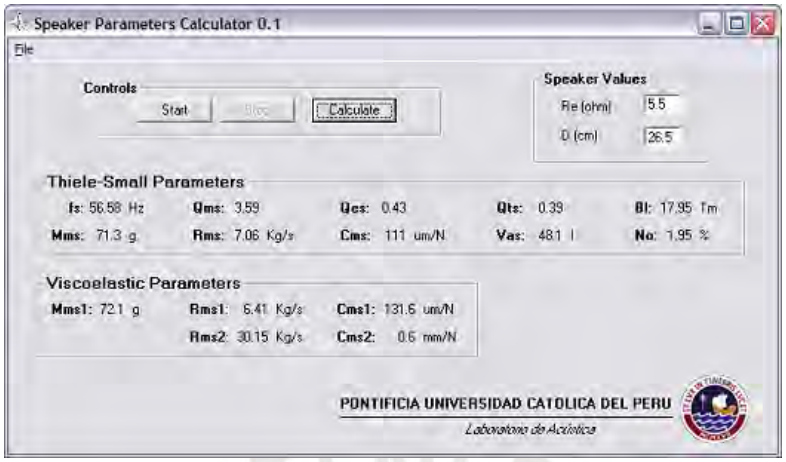


Figura 5.3: Parámetros estimados de un altavoz JBL 2206.

5.2. Conclusiones

A manera de resumen y epílogo, se pueden mencionar las siguientes conclusiones:

- El modelo TGM de la suspensión mecánica de un altavoz es una extensión del modelo tradicional que describe el incremento que exhibe la resistencia mecánica de un altavoz conforme disminuye la frecuencia en el rango de las frecuencias bajas, logrando con ello un mejor ajuste a los datos medidos que el que se obtiene con la resistencia mecánica constante del modelo tradicional.

- El criterio de los mínimos cuadrados cuantifica el grado de discrepancia entre el modelo y los n datos medidos mediante el cuadrado de la norma del vector de residuos (diferencia entre los datos modelados y medidos) tal como esta se define en el espacio euclideano $\mathbf{R}^n$. Bajo ciertos supuestos es posible demostrar que este criterio es el óptimo cuando los datos medidos están contaminados con errores de medición. Existen varios métodos para la estimación de parámetros según este criterio.

- El método de Gauss–Newton realiza una aproximación de primer orden de los residuos y estima iterativamente los parámetros que minimizan la sumatoria de los cuadrados de los residuos linealizados. La ventaja de este método es su rápida convergencia una vez alcanzada la vecindad

de los parámetros de ajuste óptimo y su desventaja es la divergencia que presenta cuando los parámetros iniciales están fuera de la región de convergencia.

- El método del Máximo Descenso utiliza la dirección de la gradiente en el espacio de parámetros para aproximar iterativamente los parámetros de ajuste óptimo. La ventaja de este método es su convergencia segura aun cuando los parámetros iniciales estén fuera de la región de convergencia y su desventaja es la lenta convergencia que presenta una vez alcanzada la vecindad de los parámetros de ajuste óptimo.

- El método de Levenberg–Marquardt realiza una interpolación óptima entre el método de Gauss–Newton y el método del Máximo Descenso. Esta interpolación está basada en la máxima vecindad en la cual la aproximación lineal introducida en el método de Gauss–Newton constituye una buena representación del problema no lineal original.

- La aplicación del método de Levenberg–Marquardt al problema particular de la estimación de parámetros mecánicos según el modelo TGM demostró ser adecuada por el buen ajuste obtenido (figura 5.2) y eficiente por el tiempo, en el orden de los milisegundos, necesario para llevar a cabo la estimación.

- Tal como se puede observar en la figura 1.5, el incremento que exhibe la resistencia mecánica del altavoz conforme aumenta la frecuencia, en el rango de las frecuencias altas, se debe a la contribución de la impedancia acústica generada por las cavidades en el circuito magnético del altavoz. El desarrollo de una extensión del modelo que reproduzca dicho efecto es materia de posteriores investigaciones.

Bibliografía

[1] L. L. Beranek, "Acoustics," American Institute of Physics, New York, 1986, p. 185.

[2] R. R. Brown, J. B. Dennis, y C. Kingsley, "Design of Systems Using Digital Computers," WADC Tech. Note 56-3383, Servo. Lab., M.I.T., 1956.

[3] A. L. Cauchy, "Méthode Générale pour la Résolution des Systèmes D'Équations Simultanées," Comptes Rendus Acad. Sci. Paris, vol. 25, pp. 536–538, 1847.

[4] R. Courant y D. Hilbert, "Methoden der mathematischen Physik II," 2da ed., Springer-Verlag, Berlin, 1968, p. 13.

[5] H. B. Curry, "The Method of Steepest Descent for Non-Linear Minimization Problems," Quarterly of Applied Mathematics, vol. 2, pp. 258–261, 1944.

[6] B. J. Elliot, "On the Measurement of Low-Frequency Parameters of Moving-Coil Piston Transducers," presentado en la AES 58th convention, 1977, preprint 1299.

[7] R. L. Ellis, "On the Method of Least Squares," Trans. Camb. Phil. Soc., vol. 8, pp. 204–219, 1894.

[8] R. Fletcher, "A Modified Marquardt Subroutine for Non-Linear Least Squares," Report AERE - R.6799, U.K. Atomic Energy Authority, 1971.

[9] C. F. Gauss, "Theorie der Bewegung der Himmelskörper welche in Kegelschnitten die Sonne Umlaufen," Carl Meyer, Hannover, 1865.

[10] C. F. Gauss, "Theorie der den kleinsten Fehlern unterworfenen Combination der Beobachtungen," en A. Börsch y P. Simon (eds.), "Abhandlungen zur Methode der kleinsten Quadrate", Stankiewicz, Berlin, 1887, pp. 1–53.

[11] G. H. Golub y C. F. Van Loan, "Matrix Computations," 3ra ed., The Johns Hopkins University Press, Baltimore, 1996, p. 140.

[12] H. O. Hartley, "The Modified Gauss-Newton Method for the Fitting of Non-Linear Regression Functions by Least Squares," Technometrics, vol. 3, pp. 269–280, 1961.

[13] S. Jønsson, J. N. Moreno, y H. Bøg "Measurement of Closed Box Loudspeaker System Parameters Using a Laser Velocity Transducer and an Audio Analyzer," presentado en la AES 92nd convention, 1992, preprint 3325.

[14] M. H. Knudsen y J. G. Jensen, "Low-Frequency Loudspeaker Models that Include Suspension Creep," J. Audio Eng. Soc., vol. 41, pp.3–18, 1993.

[15] A. N. Kolmogorov, "Justification of the Method of Least Squares," en A. N. Shiryayev (ed.), "Selected Works of A. N. Kolmogorov", Kluwer Academis Publishers, Dordrecht, 1992, vol. II, pp. 285–302.

[16] A. N. Kolmogorov, "Foundations of the Theory of Probability," 2da ed., Chelsea Publishing Company, New York, 1956, p. 6.

[17] P. S. Laplace, "Théorie Analytique des Probabilités," Courcier, Paris, 1812, livre II, chap. IV.

[18] W. M. Leach, R. W. Schafer, y T. P. Barnwell, "Time-Domain Measurements of Loudspeaker Driver Parameter," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-1 27, pp. 734–739, 1979.

[19] A. M. Legendre, "Sur la Méthode des Moindres Quarrés," en A. M. Legendre, "Nouvelles Méthodes pour la Détermination des Orbites des Comètes", Didot, Paris, 1805, pp. 72–80.

[20] K. Levenberg, “A Method for the Solution of Certain Non-Linear Problems in Least Squares,” Quarterly of Applied Mathematics, vol. 2, pp. 164–168, 1944.

[21] D. W. Marquardt, “An Algorithm for Least-Squares Estimation of Nonlinear Parameters,” J. Soc. Indust. Appl. Math., vol. 11, pp. 431–441, 1963.

[22] D. W. Marquardt, “Solution of Nonlinear Chemical Engineering Models,” Chemical Engineering Progress, vol. 55, pp. 65–70, 1959.

[23] J. Moré, “Levenberg-Marquardt Algorithm: Implementation and Theory,” en G. A. Watson (ed.), “Numerical Analysis”, Springer-Verlag, Berlin, 1978.

[24] J. N. Moreno, “Measurement of Loudspeaker Parameters Using a Laser Velocity Transducer and Two-Channel FFT Analysis,” J. Audio Eng. Soc., vol. 39, pp. 243–249, 1991.

[25] J. N. Moreno y V. R. Medina, “Measurement of Loudspeaker Parameters Considering a Better Fitting for the Mechanical Impedance,” presentado en la AES 51st international conference, Helsinki, 2013.

[26] P. M. Morse y H. Feshbach, “Methods of Theoretical Physics,” McGraw-Hill, New York, 1953, p. 494.

[27] A. Pringsheim, “Zur Geschichte des Taylorschen Lehrsatzes,” Bibliotheca Mathematica, ser. 3, vol. 1, pp. 433–479, 1900.

[28] R. H. Small, “Simplified Loudspeaker Measurements at Low Frequencies,” J. Audio Eng. Soc., vol. 20, pp. 28–33, 1972.

[29] A. N. Thiele, “Measurement of the Thiele/Small Parameters of Tweeters,” J. Elec. and Electron. Eng., Australia, vol. 9, pp. 186–199, 1989.

[30] J. H. Wilkinson, “Algebraic Eigenvalue Problem,” Oxford University Press, Oxford, 1965, p. 2.