Universidad Michoacana de San Nicolás de Hidalgo

Facultad de Ciencias Físico Matemáticas
"Mat. Luis Manuel Rivera Gutiérrez"
Posgrado en Ciencias en Ingeniería Física

Uso de esquemas en diferencias finitas generalizadas para el modelado de dispersión de contaminantes bajo el efecto de islas de calor urbano en forma de viento mesoescala

T E S I S

que para obtener el grado de

Doctor en Ciencias en Ingeniería Física

presenta

Maestro en Ciencias (Matemáticas)
Juan Salvador Lucas Martínez

Asesor:

Doctor en Ciencias (Matemáticas)
Francisco Javier Domínguez-Mota

Morelia, Michoacán, México
Febrero, 2021

Para Ana Laura y David Alejandro,
mis motivos para seguir adelante cada día.

La investigación de este trabajo se realizó gracias al apoyo económico de las siguientes entidades:

- Consejo Nacional de Ciencia y Tecnología (CONACyT), a través del Programa de Becas Nacionales para programas de posgrado reconocidas en el Programa Nacional de Posgrados de Calidad (PNPC).

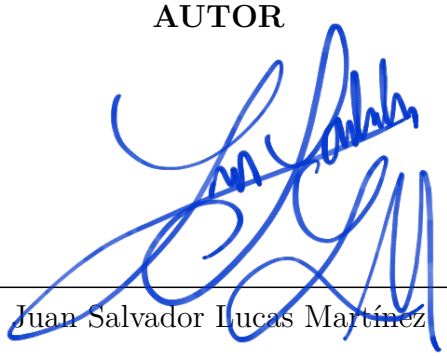
Además, las siguientes instituciones apoyaron a la investigación con sus instalaciones:

- Facultad de Ciencias Físico Matemáticas “Mat. Luis Manuel Rivera Gutiérrez”, adscrita a la Universidad Michoacana de San Nicolás de Hidalgo.

Declaración de Responsabilidad

- Declaro que este trabajo de tesis, denominado **"Uso de esquemas en diferencias finitas generalizadas para el modelado de dispersión de contaminantes bajo el efecto de islas de calor urbano en forma de viento mesoescala"**, es un trabajo original de mi investigación y ha sido escrito en su totalidad por mí.
- Declaro que esta tesis de investigación no constituye en ninguna de sus partes un plagio del trabajo material o intelectual de ningún otro autor.
- Declaro que el trabajo reportado ha sido desarrollado como resultado de mi trabajo, en colaboración con mi asesor y mi grupo de trabajo. Asimismo, las fuentes de conocimiento desarrolladas por otros autores han sido debidamente citadas y referenciadas en el escrito de esta tesis.
- Declaro que esta tesis, así como el trabajo en ella desarrollado y reportado, no está siendo utilizado para obtener otro grado académico, título o diploma en otra dependencia de la Universidad Michoacana de San Nicolás de Hidalgo, o cualquier otra institución nacional o extranjera.
- Declaro que se han proporcionado las debidas referencias sobre la literatura y los recursos de apoyo, respetando íntegramente el contenido científico de los trabajos citados, y que la presente tesis es original y no se deriva de ningún trabajo citado.
- Declaro que los artículos derivados de esta tesis fueron elaborados por el autor, con el apoyo del tutor principal y el grupo de trabajo involucrado.

AUTOR



Juan Salvador Lucas Martínez

Índice general

Uso de esquemas en diferencias finitas generalizadas para el modelado de dispersión de contaminantes bajo el efecto de islas de calor urbano en forma de viento mesoescala

Juan Salvador Lucas Martínez

Resumen

En este proyecto presento algunos esquemas en diferencias finitas generalizadas, los cuales fueron empleados en el estudio de fenómenos estacionarios de difusión-advección afines a la dispersión de contaminantes en la atmósfera en presencia de islas de calor urbanas. Estas discretizaciones se usaron en problemas de tipo Poisson, en particular, para estudiar un problema estacionario, el cual tiene la peculiaridad de que una de sus condiciones de frontera provoca un crecimiento abrupto en el gradiente de la solución a lo largo de dicha frontera. Se discute la estructura matricial del esquema propuesto, en el cual se incluye un análisis de los valores propios asociados al operador de Laplace obtenidos mediante estas discretizaciones. Los esquemas que se proponen son sencillos en su planteamiento y permiten obtener resultados numéricos con una precisión bastante aceptable.

Palabras Clave: diferencias finitas generalizadas, difusión, advección, Poisson, determinación de parámetros.

Abstract

In this project some schemes in generalized finite differences are presented, which were used in the study of steady-state, diffusion-advection phenomena, related to pollution dispersal in the atmosphere due to urban heat islands. These discrete models were used in Poisson-type problems, particularly, to study a steady-state problem, which has the particular feature that one of its boundary conditions provokes a high increase in the gradient of the solution along such boundary. Matrix structure of the proposed scheme is discussed, including an analysis of the eigenvalues associated to the Laplace operator obtained using these discrete models. Our proposed schemes are simple in their approach and they allowed us to obtain numerical results with a quite acceptable precision.

Keywords: generalized finite differences, diffusion, advection, Poisson, parameter identification.

Agradecimientos

Quiero extender de manera particular mi más sincero agradecimiento a mis padres por todo el apoyo recibido a lo largo de toda mi formación académica. Sin su apoyo y consejos, no habrá llegado a este punto. Gracias por apoyarme y permitirme ser quien quiero ser.

Agradezco también a todos mis compañeros de generación por haberme acompañado a lo largo de esta maravillosa experiencia. Gracias Marco, Beto, Tzitlali, Eliezer, los dos Gerardos, Temo, Dani y Lety. No solo disfruté mucho de su compañía, ustedes me ayudaron a crecer también en el ámbito profesional. Sin ustedes, no habría disfrutado el Doctorado tanto como lo hice.

Siempre estaré eternamente agradecido con mi querida Facultad de Ciencias Físico Matemáticas, que ha sido mi casa por casi quince años, a ella le debo toda mi formación académica. Esta gran Facultad ha sido testigo de toda mi evolución en los ámbitos personal y profesional, fue entre estos muros donde descubrí lo que soy y lo que me apasiona en la vida. Espero que en un futuro no muy lejano tenga la oportunidad de volver a pisar sus aulas.

No puedo dejar de agradecer el apoyo que recibí de mi familia a lo largo de estos años de mi Doctorado. Ana Laura, gracias por todo tu apoyo, tu comprensión y tu paciencia. Sé que todo este tiempo has estado esperando a que levante la mirada de los libros, de la computadora, y compartamos algo de nuestro tiempo juntos. Quiero que sepas que todo este tiempo estuve muy concentrado en mi trabajo porque te prometí que un día todo esto iba a valer la pena. Y a David Alejandro, no tengo palabras para expresar mi agradecimiento. A pesar de tu corta edad y a tu pequeña estatura, eres la persona más grande que conozco. Tú eres la razón por la que todas las mañanas me levanto a dar lo mejor de mí. Tú me enseñaste a dejar de hacer las cosas por mí mismo, y a dar todo lo que soy para honrar a mi familia.

Por último, pero no menos importante, quiero dar mi más profundo y sincero agradecimiento a mi tutor, Dr. Francisco Javier Domínguez Mota, por todo su apoyo, sus enseñanzas, pero sobre todo, por toda su paciencia a lo largo de estos años. A pesar de que no he sido ni el más brillante ni el más eficiente de sus estudiantes, él siempre estuvo ahí para explicarme con la misma paciencia, desde la primera hasta la 10^n-ésima vez. Nunca olvidaré su cara cuando le pedí dirigir mis estudios en este Posgrado. Gracias

por enseñarme las maravillas de las Matemáticas que yo desconocía hasta que inicié mis estudios de Doctorado. Sé que aún hay mucho por hacer y tanto por aprender, pero siempre le estaré agradecido por haber abierto esa ventana.

Hay mucha gente que no menciono, hay personas que conocí a lo largo de estos años, y personas que estuvieron ahí desde mucho antes. No hay suficiente espacio para nombrarlos a todos, pero supongo que todas las personas a las que me refiero saben quiénes son, pues nunca ignoro una oportunidad de agradecer su compañía, ya que es gracias a todas estas personas que disfruto tanto hacer lo que hago.

Capítulo 1

Ecuaciones de transporte de contaminantes

1.1 Antecedentes

La Física es el lenguaje del universo, y las Matemáticas son su alfabeto. Estas disciplinas han permitido a la humanidad analizar el mundo a su alrededor y provee de una vía para buscar una explicación a los fenómenos físicos que observamos en nuestra realidad.

A lo largo de los siglos se han desarrollado muchos conceptos teóricos, hipótesis, formulaciones y teorías en el ámbito de las Matemáticas, muchas de las cuales han tenido su inspiración en problemas físicos. Entre todos estos conceptos, considero que uno de los más importantes es el operador de Laplace ∇, el cual aparece con frecuencia en el estudio de problemas físicos, en una amplia variedad de problemas. En el ámbito de las ecuaciones diferenciales parciales, este operador está presente en los principales tipos de EDP's: en problemas de tipo elíptico en la formulación de problemas de tipo Laplace $\nabla^2 u = 0$, o de manera más general, en problemas de tipo Poisson $\nabla^2 u = f$, así como la ecuación de Helmholtz $\nabla^2 u = ku$; en problemas de tipo parabólico, como en fenómenos de difusión $\frac{\partial u}{\partial t} = k\nabla^2 u$; o en problemas de tipo hiperbólico, como la ecuación de onda $\frac{\partial u}{\partial t} = c^2\nabla^2 u$, o la ecuación de advección $\frac{\partial u}{\partial t} + k\nabla^2 u = 0$, la cual se discutirá con más detalle. Los resultados asociados a este operador resultan de gran interés para varios campos afines a la Física y las Matemáticas, como la Teoría de problemas de valores a la frontera para ecuaciones elípticas o la Hidrodinámica, solo por citar algunos ejemplos. En el caso particular de este proyecto, nuestra motivación es la ecuación de difusión-advección

$$u\frac{\partial c}{\partial x} + v\frac{\partial c}{\partial y} + w\frac{\partial c}{\partial z} = \frac{\partial}{\partial x}\left(K_x\frac{\partial c}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y\frac{\partial c}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z\frac{\partial c}{\partial z}\right), \tag{1.1.1}$$

donde c es la concentración de contaminante (g/m^3) en el punto (x, y, z); u, v y w son las componentes del viento (m/s) en direcciones hacia abajo (x), de costado (y) y vertical (z), respectivamente; K_x, K_y y K_z son los coeficientes de difusividad en las

direcciones correspondientes.

Las soluciones exactas en problemas donde interviene el operador de Laplace, o de manera más general, expresiones de tipo elíptico, son no sólamente bienvenidas, sino también muy deseadas. Daniela Buske *et al* [12], calcularon las soluciones exactas para la ecuación de advección-difusión en tres dimensiones, cuya existencia y unicidad está garantizada por el Teorema de Cauchy-Kowalewski. D.M. Moreira *et al* [67], obtuvieron una solución a la ecuación de advección-difusión en dos dimensiones usando transformadas de Laplace.

Existen soluciones exactas para problemas de tipo Poisson o problemas elípticos muy particulares. Una de mis referencias favoritas para el estudio de estos problemas es [65], donde se describe con amplio detalle esta clase de problemas, y las condiciones particulares en las que existe una solución exacta. Sin embargo, muchas de estas propuestas requieren soluciones numéricas de ecuaciones trascendentes, como en el trabajo de D.M. Moreira [67] o la propuesta de M. Stynes [79]. Bajo condiciones estándar (condiciones particulares de Dirichlet o Neumann a la frontera), la existencia y unicidad de la solución está garantizada por el Teorema de Cauchy-Kowalewski. Li y Lu [62] estudiaron problemas de tipo elíptico con singularidades a la fontera, entre los que se destaca el problema de Motz [68].

Se han desarrollado diferentes enfoques y técnicas para obtener soluciones a estos problemas. La técnica particular de interés para el desarrollo de este proyecto es la que se conoce en la literatura como *esquemas en diferencias finitas generalizadas*, la cual consiste, a *grosso modo*, en proponer una forma discreta tanto para el dominio como para los operadores involucrados en un problema físico, a fin de aprovechar dichas estructuras discretas para calcular una solución aproximada al problema. En contraste con los métodos en diferencias finitas clásicas, nuestra propuesta tiene la ventaja de tener una mayor adaptabilidad para modelar la situación planteada por los problemas físicos que discutiremos.

Esta técnica ha sido estudiada por varios autores a lo largo de varias décadas, y a partir de diferentes enfoques. Nuestras referencias se remontan a 1972, cuando P.S. Jensen [47] estudió diversas maneras de adaptar mallados regulares a dominios con una geometría irregular.

En 2003, Jörg Kuhnert y Sudarshan Tiwari [54] publicaron su propuesta titulada *Finite Pointset Method*, la cual es una técnica basada en un método libre de malla, la cual emplearon para estudiar fenómenos relacionados a las ecuaciones de Navier-Stokes. Más tarde, en 2014, Edgar Reséndis-Flores e Irma García-Calvillo [73] usaron el Finite Pointset Method para estudiar fenómenos de conducción de calor.

La siguiente referencia importante para nuestro proyecto fueron los trabajos de J.J. Benito-Muñoz, F. Ureña-Prieto y L. Gavete-Corvinos [82, 83, 85, 86], los cuales han publicado, entre 2001 y 2017, varios estudios relacionados al método de diferencias fini-

tas generalizadas, entre los que se incluyen problemas de tipo parabólico e hiperbólico, su solución a la ecuación de advección-difusión, así como el uso de este método para estudiar la propagación de ondas sísmicas.

En 2018, Yan Gu, *et al* [37], publicaron su trabajo donde emplearon un método en diferencias finitas generalizadas para resolver la ecuación de calor en $3+1$ dimensiones. Si bien su trabajo lleva en el título el método de diferencias finitas generalizadas, su propuesta se basa en un método libre de malla, donde la distribución de los puntos en el dominio se hace en base a métodos de colocación.

De 2014 a la fecha, F.J. Domínguez-Mota [27, 24, 19, 25, 26], ha realizado varios estudios relacionados a los esquemas en diferencias finitas generalizadas, entre los que destacan sus aplicaciones a la ecuación de calor, fenómenos de advección-difusión, y más recientemente, la formulación de un esquema en DFG para la solución al problema de Motz, un problema de *benchmark* en métodos numéricos característico por sus condiciones a la frontera.

Debo destacar de manera particular el trabajo de Manju Agarwal y Abhinav Tandon [4], quienes en 2009 propusieron un esquema en diferencias finitas para el modelado de islas de calor urbano en forma de viento mesoescala y sus efectos en la dispersión de contaminantes en la atmósfera en zonas urbanas. Unos de los objetivos de este proyecto es comparar nuestras soluciones con las reportadas por Agarwal y Tandon, cuyo trabajo fue una de las principales motivaciones para este proyecto, motivo por el cual creo que se merece una descripción más detallada.

Para concluir esta sección, quiero brindar una perspectiva general sobre la esctructura de este trabajo. A continuación, se brinda una breve descripción del trabajo de Agarwal y Tandon. En lo que resta de este capítulo, haremos una breve descripción de las ecuaciones de transporte que se utilizan para los fines de este proyecto. En el Capítulo 2 se describen los métodos para resolver el problema discreto, del cual hablaremos con detalle más adelante. En el Capítulo 3 se describen los esquemas en diferencias finitas generalizadas que hemos empleado para nuestras discretizaciones. Finalmente, el Capítulo 4 lo dedicaremos a plantear el problema inverso que estudiamos en este proyecto.

1.1.1 El modelo de Agarwal y Tandon

El modelo propuesto por estos autores se basa en la ecuación de conservación de masa. La ecuación que propusieron para la concentración de contaminantes en estado estático promedio es

$$u\frac{\partial c}{\partial x} + v\frac{\partial c}{\partial y} + w\frac{\partial c}{\partial z} = \frac{\partial}{\partial x}\left(K_x\frac{\partial c}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y\frac{\partial c}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z\frac{\partial c}{\partial z}\right) + R. \tag{1.1.2}$$

donde c es la concentración de contaminante en el punto (x, y, z); R es el término de

remoción/reacción; u, v y w son las componentes del viento en direcciones hacia abajo (x), de costado (y) y vertical (z), respectivamente; K_x, K_y y K_z son los coeficientes de difusividad en las direcciones correspondientes.

El problema físico que estudiaron estos autores consiste en un área fuente que se esparce sobre la superficie de una ciudad a una distancia finita. Se asume que los contaminantes se emiten a una tasa constante desde el área fuente y se esparcen dentro de la capa de mezcla adyacente a la superficie de la tierra donde toma lugar la mezcla de sustancias como resultado de la turbulencia y el movimiento de convección. Esta capa de mezcla se extiende desde la superficie hasta una altura donde todos los flujos y divergencias resultantes de la acción de la superficie virtualmente se anulan. Los autores consideraron una región fuente dentro del área urbana que se estiende desde el origen hasta una distancia l en dirección x hacia abajo ($0 \leqslant x \leqslant l$) y un área libre de fuente ($l < x \leqslant X$) más allá de l, donde X es la distancia deseada para calcular la distribución de concentración. Asumiendo la homogeneidad del terreno, la concentración principal de contaminante se considera constante a lo largo de la dirección transversal y, lo cual reduce el problema a dos dimensiones, y por tanto se eliminan los gradientes en la dirección y. Más aún, asumiendo que la dirección x de difusión es despreciable en comparación con la direción x de advección, la ecuación (1.1.2) se simplifica a:

$$u\frac{\partial c}{\partial x} + w\frac{\partial c}{\partial z} = \frac{\partial}{\partial z}\left(K_z\frac{\partial c}{\partial z}\right) + R. \tag{1.1.3}$$

El contaminante, emitido desde el área fuente, es transportado horizontalmente por el viento a gran escala, el cual se toma como función de la altitud (distancia vertical), y adicionalmente, en las direcciones horizontal y vertical por el viento mesoescala, el cual se elige como representante del viento local causado por una fuente de calor. Así, la ecuación (1.1.3) se escribe como:

$$(u + u_e)\frac{\partial c}{\partial x} + w_e\frac{\partial c}{\partial z} = \frac{\partial}{\partial z}\left(K_z\frac{\partial c}{\partial z}\right) + R. \tag{1.1.4}$$

donde u es el viento a gran escala en la dirección horizontal x, u_e y w_e son las componentes del viento mesoescala en las direcciones x y z respectivamente.

El efecto de islas de calor en una ciudad causan levantamientos de aire sobre el centro de la isla de calor. Este levantamiento de aire causa una afluencia de aire en los alrededores, por lo que se crea una gran corriente convectiva inducida térmicamente, lo que origina el viento mesoescala. Este comportamiento se puede modelar como:

$$u_e \propto -x, \quad w_e \propto z.$$

lo que refleja el comportamiento básico de este tipo de viento, es decir, una afluencia superficial y una corriente que se levanta sobre el origen. El viento a gran escala (u) y la difusividad vertical (K_z) se parametrizan como funciones de la altura vertical z como:

$$\begin{aligned} u &= u(z) = u_r \left(\frac{z}{z_r}\right)^{\alpha}, \\ K_z &= K_z(z) = K_r \left(\frac{z}{z_r}\right)^{\beta}, \end{aligned}$$

donde $u_r = u(z_r)$ y $K_r = K_z(z_r)$ son la velocida medida del viento y la difusividad vertical a una altura de referencia z_r y α, β son las constantes que dependen de la estabilidad atmosférica y la aspereza de la superficie. Los modelos para las componentes horizontal y vertical del viento mesoescala dentro del rango de validez se toman como:

$$\begin{aligned} u_e &= -ax \left(\frac{z}{z_r}\right)^{\alpha}, \\ w_e &= \frac{az}{(\alpha+1)} \left(\frac{z}{z_r}\right)^{\alpha}, \end{aligned}$$

donde a es una constante de proporcionalidad.

Tomando en cuenta los mecanismos de remoción tales como las reacciones químicas, la lluvia y los mecanismos artificiales que prevalecen en la atmósfera, la ecuación (1.1.4) se escribe como:

$$(u+u_e)\frac{\partial c}{\partial x} + w_e \frac{\partial c}{\partial z} = \frac{\partial}{\partial z}\left(K_z \frac{\partial c}{\partial z}\right) - \lambda c, \tag{1.1.5}$$

donde λ es un parámetro constante de reducción de primer orden que define la pérdida fraccionaria de contaminante por unidad de tiempo a través de varios procesos de deposición que existen en la atmósfera.

Se asume que hay algún contaminante de concentración c_0 entrando en $x = 0$ en el dominio de interés, por lo que

$$c(0,z) = c_0 \quad \text{en } x = 0 \text{ para } 0 \leqslant z \leqslant H,$$

donde H es la altura de la capa de mezcla.

Se asume que los reactivos químicos contaminantes del aire se emiten a una tasa constante desde el nivel del suelo y son removidos de manera simultánea desde la atmósfera por la absorción del suelo (deposición en seco), que puede ser expresado como:

$$K_z \frac{\partial c}{\partial z} = \begin{cases} v_d c - Q & \text{en } z = 0,\ 0 \leqslant x \leqslant l, \\ v_d c & \text{en } z = 0,\ l < x \leqslant X, \end{cases}$$

donde Q es la tasa de emisión de contaminante, l es la longitud de la fuente en dirección hacia abajo, X es la distancia total considerada en dirección hacia abajo y v_d es la

velocidad de deposición en seco. Los contaminantes son confinados dentro de la altura de mezcla y no son capaces de penetrar la capa de mezcla, por lo que

$$K_z \frac{\partial c}{\partial z} = 0 \quad \text{en } z = H,\ x > 0.$$

1.2 Transporte

Los fenómenos de transporte son muy comunes en la naturaleza, ya sea que este término se aplique al calor o a cierta sustancia en particular, estos fenónemos se encuentran prácticamente en todas partes del planeta, en la hidrósfera, la atmósfera y la pedósfera.

En este contexto, el término *transporte* se emplea para denotar los procesos que determinan la distribución de agentes biológicos, geológicos, químicos o de calor en el medio ambiente. Para los fines de este trabajo, se entenserá al término *transporte* como una interacción entre procesos físicos con un efecto en especies o componentes, o en el calor. En esta definición se dejarán de lado otros procesos relacionados al fenómeno de transporte, tales como la absorción, la degradación, el decaimiento o las reacciones de diferentes tipos.

En este capítulo se consideran dos tipos de procesos de transporte: la advección y la difusión o dispersión. La advección se refiere al transporte en su sentido más básico: una partícula siendo movida de un lugar a otro por la acción de un campo de flujo. La difusión es un fenómeno que se origina cuando están presentes ciertas diferencias en la concentración, y que se debe a la tendencia en los sistemas físicos a igualar los gradientes de concentración. Si dentro de un sistema físico hay especies con la posibilidad de moverse de un sitio a otro, habrá una red difusiva o un flujo dispersivo desde un punto con una concentración alta hacia puntos con menor concentración. Las ecuaciones de transporte se derivan de principios de conservación de masa y de la Ley de Fick.

1.3 Ecuación de continuidad para la masa

Para la deducción de la ecuación de conservación de masa, considere un volumen de control en forma de prisma rectangular, cuyas dimensiones son Δx, Δy y Δz. La idea es calcular la diferencia de masa dentro del volumen de control durante un pequeño intervalo de tiempo Δt, para lo cual se deben de considerar las seis caras del volumen de control.

El volumen de control se muestra en la figura (1.1), el cual contiene cierta cantidad de masa al inicio del intervalo de tiempo Δt, y contiene una cantidad diferente al final. Durante este intervalo de tiempo hay un flujo que entra por una de las caras del volumen de control, así como un flujo de salida a través de otra de las caras. Las masas al inicio y al final del intervalo Δt están dadas por:

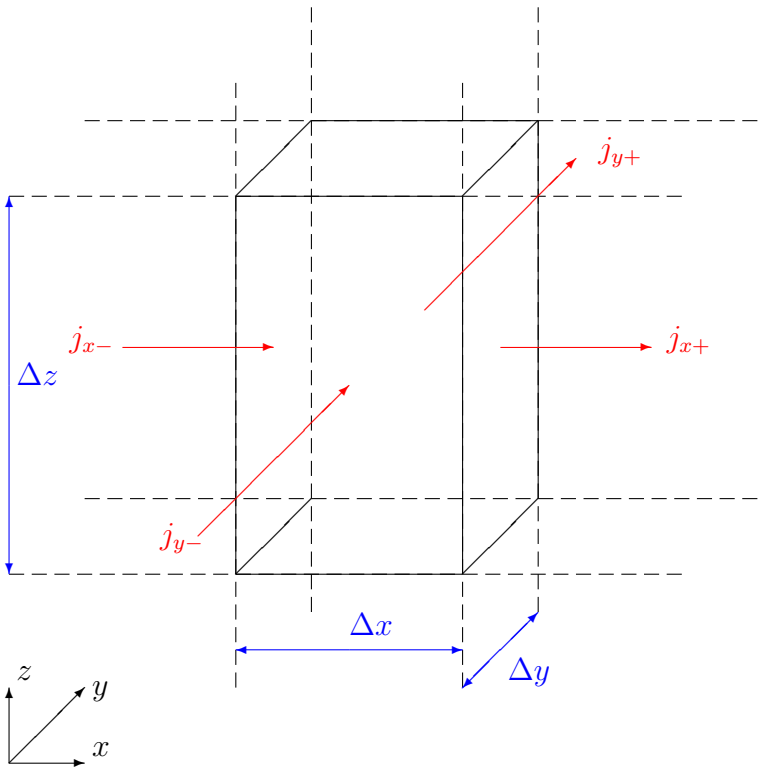


Figura 1.1: Volumen de control.

$$\theta \cdot c(x,t) \cdot \Delta x \Delta y \Delta z \quad \text{y} \quad \theta \cdot c(x,t+\Delta t) \cdot \Delta x \Delta y \Delta z,$$

donde el parámetro θ denota la contribución en el volumen total. Para medios porosos saturados, θ denota la porosidad del material. En la zona no saturada, dentro del suelo por ejemplo, θ denota la saturación volumétrica de agua cuando se refiere a la parte acuosa. En la situación en que dos fluidos comparten el espacio (por ejemplo agua y aceite), la contribución de cada fase se debe de tomar en cuenta. El parámetro c denota la concentración, la cual se mide en [masa/volumen], y $\Delta x \Delta y \Delta z$ es el volumen del elemento de control. El cambio en la masa con respecto al tiempo está dado por:

$$\theta \cdot \frac{c(x,t+\Delta t) - c(x,t)}{\Delta t} \cdot \Delta x \Delta y \Delta z.$$

Los flujos en la dirección x del volumen de control están dados por:

$$\theta j_{x-}(x,t) \Delta y \Delta z \quad \text{y} \quad \theta j_{x+}(x,t) \Delta y \Delta z,$$

donde j_{x-} denota el flujo de masa por unidad de área a través de la cara izquierda del volumen de control, en la dirección negativa x. De la misma manera, j_{x+} denota el flujo de masa por unidad de área en la dirección x a través de la cara derecha, en la dirección positiva x (ver figura (1.1)). Los flujos pueden variar espacial o temporalmente. Si el flujo agrega masa al volumen de control, entonces dicho flujo se considera positivo; de la misma manera, si el flujo resta masa del volumen de control, dicho flujo se considera

negativo. El flujo de masa se mide en $[M/(L^2 \cdot T)]$. El término $\theta \Delta y \Delta z$ denota el área por la que atraviesa el flujo.[1]

El balance entre ambos flujos está dado por:

$$\theta(j_{x-}(x,t) - j_{x+}(x,t))\Delta y \Delta z.$$

Por simplicidad, se han omitido los flujos a través de las otras cuatro caras del volumen de control, asumiendo que el flujo en las direcciones y y z son nulos. Como se mencionó anteriormente, ambas formulaciones miden el cambio en la masa, y por tanto, deben coincidir:

$$\theta \cdot \frac{c(x,t+\Delta t) - c(x,t)}{\Delta t} \cdot \Delta x \Delta y \Delta z = \theta(j_{x-}(x,t) - j_{x+}(x,t))\Delta y \Delta z. \tag{1.3.1}$$

Dividiendo entre el volumen $\Delta x \Delta y \Delta z$ y el parámetro θ:

$$\frac{c(x,t+\Delta t) - c(x,t)}{\Delta t} = -\frac{j_{x+}(x,t) - j_{x-}(x,t)}{\Delta x}. \tag{1.3.2}$$

Obteniendo el límite cuando $\Delta x \to 0$ y $\Delta t \to 0$ se obtiene la ecuación diferencial:

$$\frac{\partial c}{\partial t} = -\frac{\partial}{\partial x} j_x, \tag{1.3.3}$$

que es una formulación para el principio de conservación de la masa. En esta deducción se asume que las funciones c y j_x son suficientemente suaves, o mejor dicho, suficientemente diferenciables, matemáticamente hablando. La ecuación (1.3.3) es válida para transporte en una dimensión y es la base para el análisis de procesos de transporte. La unidad de esta ecuación es $[M/L^3 \cdot T]$.

La ecuación (1.3.3) es válida si no hay fuentes o sumideros para las sustancias en cuestión. Esta formulación se puede extender para considerar fuentes o sumideros adicionales. Las fuentes o sumideros en cuestión se pueden describir usando un término $q(x,t)$, medido en $[M/L^3 \cdot T]$, el cual puede variar espacial y temporalmente. El término integral correspondiente

$$\int_{\Delta x} \int_{\Delta t} q(x,t) dt dx,$$

se agrega en el lado derecho de las ecuaciones (1.3.1) y (1.3.2). Este término es positivo cuando añade masa (fuente) y negativo cuando remueve masa (sumidero). En la deducción de la ecuación (1.3.3) el término integral debe ser diferenciado, lo que lleva a la ecuación general de transporte en una dimensión:

[1] En general se asume que la contribución volumétrica y la contribución en el área, comparadas con el volumen total o el área total, respectivamente, se cuantifican con el parámetro θ, lo cual en general no es cierto. En sistemas particulares, tales como los filtros, ambas razones pueden variar significativamente.

$$\theta\frac{\partial c}{\partial t} = -\frac{\partial}{\partial x}\theta j_x + q. \tag{1.3.4}$$

Los flujos en las componentes y y z también se peuden tomar en cuenta, haciendo deducciones análogas. Los flujos j_{y-}, j_{y+}, j_{z-} y j_{z+} se introducen, se balancean y se agregan al lado derecho de (1.3.1) y (1.3.2). Tomando los límites cuando $\Delta y \to 0$ y $\Delta z \to 0$ se obtiene:

$$\theta\frac{\partial c}{\partial t} = -\left(\frac{\partial}{\partial x}\theta j_x + \frac{\partial}{\partial y}\theta j_y + \frac{\partial}{\partial z}\theta j_z\right) + q, \tag{1.3.5}$$

que es la formulación general para la conservación de masa en tres dimensiones. Usando el operador nabla ∇, esta ecuación se puede escribir de forma más compacta como:

$$\theta\frac{\partial c}{\partial t} = -\nabla \bullet \theta\mathbf{j} + q. \tag{1.3.6}$$

La ecuación de conservación de masa que se acaba de obtener es insuficientemente por sí misma para una formulación matemática completa, ya que involucra demasiadas variables, tales como la concentración c y los componentes del vector de flujo $\mathbf{j}$. Para reducir la cantidad de variables, se recomienda usar una formulación que relacione al vector de flujo con la concentración, para así llegar a una ecuación en la que la concentración sea la única variable desconocida.

El flujo de advección está dado por el producto de la concentración por la velocidad. En el caso tridimensional, las componentes del flujo son:

$$j_x = v_x c, \quad j_y = v_y c, \quad j_z = v_z c.$$

Estas componentes se pueden escribir en notación vectorial como

$$\mathbf{j} = \mathbf{v}c = c\mathbf{v}.$$

1.4 El principio de conservación

La conservación de una variable A, la cual puede representar masa, momento o energía, la cual es dependiente del tiempo t y de las dimensiones espaciales x, y y z, se expresa de manera cuantitativa por la ecuación diferencial:

$$\frac{\partial}{\partial t}A = \frac{\partial}{\partial x}j_{A_x} + \frac{\partial}{\partial y}j_{A_y} + \frac{\partial}{\partial z}j_{A_z} + Q,$$

donde j_{A_x}, j_{A_y} y j_{A_z} son las componentes del vector de flujo $\mathbf{j}_A$ en las correspondientes direcciones espaciales. Todos los términos involucrados en esta ecuación son dependientes del tiempo t y de las componentes espaciales x, y y z. El término Q reune a todas las fuentes y sumideros. Si $Q(x, y, z, t)$ es positivo, entonces hay una fuente al tiempo t en la posición $\mathbf{r} = (x, y, z)$; si $Q(x, y, z, t)$ es negativo, entonces hay un sumidero en

la posición y tiempo correspondientes.

De acuerdo con la ecuación de continuidad, el cambio en la variable A en el tiempo es igual al costo local del flujo. La ecuación de contiuidad se deduce del costo en un volumen de control, como se hizo en la sección (1.3).

En el intervalo de tiempo Δt, la cantidad de A por unidad de volumen cambia de $A(x, y, z, t)$ a $A(x, y, z, t + \Delta t)$. El cambio total en el volumen de control es entonces

$$\left(A(x, y, z, t + \Delta t) - A(x, y, z, t)\right) \Delta x \Delta y \Delta z.$$

Por otro lado, el costo total se expresa en términos de los flujos, las fuentes y los sumideros. En cada dimensión espacial hay dos caras, a través de las cuales la variable A puede entrar o salir, de acuerdo a la componente del flujo. En la dirección x los flujos a través de las dos caras están dados por:

$$\left(j_{A_x}(x + \frac{\Delta x}{2}, y, z, t) - j_{A_x}(x - \frac{\Delta x}{2}, y, z, t)\right) \Delta y \Delta z \Delta t.$$

Observe que en este paso se omite el efecto de las fuentes o sumideros. Este hecho se basa en la suposición de que el tamaño del paso Δt es suficientemente pequeño como para ignorar el efecto de dichas fuentes o sumideros.

Ambas expresiones del cambio dentro del volumen de control dentro del intervalo Δt deben coincidir, lo que lleva a la ecuación:

$$\begin{aligned}
& \left(A(x, y, z, t + \Delta t) - A(x, y, z, t)\right) \Delta x \Delta y \Delta z \\
= \; & \left(j_{A_x}(x + \frac{\Delta x}{2}, y, z, t) - j_{A_x}(x - \frac{\Delta x}{2}, y, z, t)\right) \Delta y \Delta z \Delta t \\
+ \; & \left(j_{A_y}(x, y + \frac{\Delta y}{2}, z, t) - j_{A_y}(x, y - \frac{\Delta y}{2}, z, t)\right) \Delta x \Delta z \Delta t \\
+ \; & \left(j_{A_z}(x, y, z + \frac{\Delta z}{2}, t) - j_{A_z}(x, y, z - \frac{\Delta z}{2}, t)\right) \Delta x \Delta y \Delta t \\
+ \; & Q \Delta x \Delta y \Delta z \Delta t.
\end{aligned}$$

Para simplificar esta ecuación, se divide cada término por el producto $\Delta x \Delta y \Delta z \Delta t$, para obtener:

$$\begin{aligned}
\frac{A(x, y, z, t + \Delta t) - A(x, y, z, t)}{\Delta t} = \; & \frac{j_{A_x}(x + \frac{\Delta x}{2}, y, z, t) - j_{A_x}(x - \frac{\Delta x}{2}, y, z, t)}{\Delta x} \\
+ \; & \frac{j_{A_y}(x, y + \frac{\Delta y}{2}, z, t) - j_{A_y}(x, y - \frac{\Delta y}{2}, z, t)}{\Delta y} \\
+ \; & \frac{j_{A_z}(x, y, z + \frac{\Delta z}{2}, t) - j_{A_z}(x, y, z - \frac{\Delta z}{2}, t)}{\Delta z} + Q.
\end{aligned}$$

Ahora se toman los límites cuando $\Delta x \to \partial x$, $\Delta y \to \partial y$, $\Delta z \to \partial z$ y $\Delta t \to \partial t$, para obtener la ecuación diferencial:

$$\frac{\partial A}{\partial t} = \nabla \bullet \mathbf{j}_A + Q. \tag{1.4.1}$$

Esta ecuación relaciona el flujo como una función de la concentración y es válida en el caso del transporte advectivo. Para el caso del transporte difusivo, es necesario introducir relaciones adicionales, por ejemplo, la Ley de Fick.

1.5 Ecuaciones de difusión

Por difusión se entiende el fenómeno que causa la tendencia natural en un sistema a balancear las diferencias de concentración. Si en un sistema existe una alta concentración en un punto inicial y una baja concentración en un punto terminal, se genera una red de flujo difusivo de componente desde el punto con mayor concentración hacia el punto con menor concentración. A escala molescular, la difusión es un movimiento aleatorio de moléculas en todas las direcciones. En un sistema donde no existen diferencias en los niveles de conentración, todos los caminos mantienen juntos el mismo nivel de concentración. Pero si el nivel de concentración no es constante, existirá entonces una red de flujo en una dirección, desde los niveles con alta concetración hacia los niveles con menor concentración.

Un sistema con diferencias iniciales de concentración alcanzará eventualmente un nivel constante de concentración si no intervienen procesos externos, los cuales podrían estabilizar el gradiente de concentración. El flujo de difusión puede entonces ser balanceado mediante procesos externos que mantengan una entrada y salida de flujo constante, pero aún en este caso, existe un flujo de difusión que acompaña al gradiente de concentración.

La (primera) Ley de Fick[2] es una fórmula empírica de la cuantificación de un flujo de difusión, cuya unidad de medida es [M/A · T]. El enunciado de la ley para fluidos es:

$$\mathbf{j} = -D\nabla c. \tag{1.5.1}$$

Esta ley afirma que el flujo de difusión es proporcional al gradiente negativo de la concentración. El signo menos se debe a que la dirección del flujo va desde puntos con

[2]Adolf Eugen Fick (1829-1901) fue un fisiólogo alemán. La segunda Ley de Fick es una ley de conservación para la masa en un medio de una fase:

$$\frac{\partial}{\partial t}c = D\frac{\partial^2}{\partial x^2}c.$$

Esta fórmula se obtiene cuando se reemplaza el flujo de (1.5.1) en la ecuación (1.4.1)

altas concentraciones a puntos con bajas concentraciones. La constante de proporcionalidad es la *constante de difusión* o *difusividad*, cuya unidad es [A/T].

Cabe mencionar que la difusividad D depende en general del fluido y del componente transportado, así como de la temperatura, la presión y del medio geoquímico. Todas las sustancias tienen una difusividad en gases, que varía en general de su difusividad en líquidos, y que depende a su vez del tipo de líquido o gas.

La difusividad, como se definió en (1.5.1), es válida para sistemas de una sola fase, y depende de las moléculas involuradas, es decir, de los componentes y del medio. Por esta razón es común referirse al parámetro D como *difusividad molecular*. A partir de este momento, y para tomar en cuenta esta observación, se introduce el parámetro D_{mol} para referirse a la difusividad molecular, mientras que se reserva el parámetro D para la difusividad en general.

Para formular la Ley de Fick para sistemas multifase, por ejemplo, para medios porosos, se deben agregar algunas observaciones. Primero se debe de tomar en cuenta que el área que atraviesa el flujo difusivo podría ser solo una parte del área total. Es común asumir que el área se reduce por el mismo factor que el volumen. La contribución volumétrica del espacio poroso, la porosidad, se toma entonces como el factor que mide la contribución del área activa. Es por ello que el factor θ aparece en el lado derecho de (1.5.1) cuando se aplica a medios porosos. En medios porosos no saturados θ representa el contenido volumétrico de agua.

La segunda observación necesaria a tomar en cuenta es que las trayectorias o los caminos que sigue el flujo de difusión son necesariamente más largos cuando están presentes varias fases. En un sistema de una sola fase, el flujo de difusión puede moverse a través del camino más corto disponible, pero en un sistema de varias fases pueden presentarse obstáculos a lo largo de dicha trayectoria. Como las trayectorias son más largas en sistemas multifase, el flujo difusivo en esos sistemas es más pequeño que en sistemas de una sola fase. Si las trayectorias del flujo son más prolongadas, es conveniente introducir un factor $\vartheta > 1$ en el denominador del gradiente de concentración.

Este incremento en la trayectoria del flujo debe de ser considerada en el cálculo del flujo $\mathbf{j}$. El flujo en dirección normal es más pequeño que el flujo que sigue una trayectoria más larga de la normal. El efecto combinado de ambas correcciones con el *factor de prolongación de caminos* ϑ conduce a la ecuación:

$$\mathbf{j} = -\frac{1}{\vartheta^2} D_{\text{mol}} \nabla c. \tag{1.5.2}$$

Algunos autores de la dinámica de fluidos prefieren escribir esta formulación de la Ley de Fick introduciendo el término denominado *tortuosidad*, denotado τ, cuyo valor oscila entre cero y uno. La tortuosidad se relaciona con el factor de prolongación de caminos mediante la fórmula $\tau = 1/\vartheta^2$, con lo que la ecuación (1.5.2) se puede escribir en términos de τ como:

$$\mathbf{j} = -\tau D_{\text{mol}} \nabla c. \tag{1.5.3}$$

De esta manera, el coeficiente del gradiente consta de tres términos, y se le suele llamar difusividad efectiva D_{eff}:

$$D_{\text{eff}} = \theta \tau D_{\text{mol}}. \tag{1.5.4}$$

Cabe señalar que la interpretación del término *efectiva* puede estar sujeta a interpretación. En ocasiones se puede entender por 'efectiva' al producto de la difusividad por la tortuosidad, sin incluir la porosidad. A veces el término 'efectivo' se omite. En contraste con la difusividad efectiva, a la difusividad en una fase se le suele llamar *difusividad molecular*.

Para una discusión más detallada sobre estas ecuaciones, recomiendo referirse a [41, 20, 28].

1.5.1 Solución analítica a la ecuación de difusión en una dimensión para difusividad constante

Las soluciones generales de la ecuación de difusión se pueden obtener para una variedad de condiciones iniciales y de frontera si se asume que la difusividad es constante. En esta sección se presenta una forma de calcular una solución analítica para la ecuación de difusión en una dimensión con difusividad constante. Considere la ecuación

$$\frac{\partial c}{\partial t} = D \frac{\partial^2 c}{\partial x^2}. \tag{1.5.5}$$

Es posible obtener una solución utilizando el método de separación de variables. Se asume que la función de concentración se puede escribir como:

$$c = X(x)T(t).$$

Entonces se sustituye esta solución en (1.5.5) para obtener

$$X \frac{dT}{dt} = DT \frac{d^2 X}{dx^2},$$

lo cual se puede reescribir como

$$\frac{1}{T} \frac{dT}{dt} = \frac{D}{X} \frac{d^2 X}{dx^2},$$

en esta ecuación, el lado izquierdo de la igualdad depende unicamente de t, mientras que el lado derecho depende solamente de x. Como ambas expresiones coinciden, ambas deben ser iguales a una cierta constante, la cual, para fines de simplicidad, será llamada $-\lambda^2 D$. Se tienen entonces dos ecuaciones diferenciales ordinarias:

$$\begin{aligned} \frac{1}{T}\frac{dT}{dt} &= -\lambda^2 D, \\ \frac{1}{X}\frac{d^2X}{dx^2} &= -\lambda^2, \end{aligned}$$

cuyas soluciones son

$$\begin{aligned} T(t) &= e^{-\lambda^2 Dt}, \\ X(x) &= A\sin(\lambda x) + B\cos(\lambda x), \end{aligned}$$

lo cual lleva a una solución para (1.5.5) de la forma

$$c = (A\sin(\lambda x) + B\cos(\lambda x))\, e^{-\lambda^2 Dt}, \tag{1.5.6}$$

donde A y B son constantes de integración. Dado que la ecuación (1.5.5) es lineal, la solución general se obtiene sumando soluciones del tipo (1.5.6), con lo que se obtiene

$$c = \sum_{m=1}^{\infty} (A_m \sin(\lambda_m x) + B_m \cos(\lambda_m x))\, e^{-\lambda_m^2 Dt}, \tag{1.5.7}$$

donde las constantes A_m, B_m y λ_m son determinadas por las condiciones iniciales y de frontera del problema físico en particular. De este modo, si se estudia el problema de difusión fuera de una hoja plana de ancho l, a través de la cual la sustancia que se esparce está inicialmente distribuida de manera uniforme y cuyas superficies se mantienen a un nivel cero de concentración, las condiciones son:

$$c = c_0, \quad 0 < x < l, \quad t = 0, \tag{1.5.8}$$

$$c = 0, \quad x = 0, \quad x = l \quad t > 0. \tag{1.5.9}$$

Las condiciones de frontera (1.5.9) implican que:

$$B_m = 0, \quad \lambda_m = \frac{m\pi}{l},$$

y por tanto la condición inicial (1.5.8) se convierte en

$$c_0 = \sum_{m=1}^{\infty} A_m \sin\left(\frac{m\pi x}{l}\right), \quad 0 < x < l. \tag{1.5.10}$$

Al multiplicar ambos lados de (1.5.10) por $\sin(p\pi x/l)$ e integrando desde 0 hasta l usando las relaciones

$$\int_0^l \sin\left(\frac{p\pi x}{l}\right) \sin\left(\frac{m\pi x}{l}\right) dx = \begin{cases} 0, & m \neq p, \\ \frac{l}{2}, & m = p, \end{cases}$$

se tiene que para m par, este término se anula, mientras que para m impar se tiene

$$A_m = \frac{4c_0}{m\pi}, \quad m \text{ impar.} \tag{1.5.11}$$

Por tanto, la solución es

$$c = \frac{4C_0}{\pi} \sum_{n=0}^{\infty} \frac{1}{2n+1} e^{-D(2n+1)^2\pi^2 t/l^2} \sin \frac{(2n+1)\pi x}{l}, \tag{1.5.12}$$

Esta solución en series trigonométricas es convergente para tiempos moderados y largos, por lo que se emplea en evaluación numérica en lugar de soluciones analíticas que involucran a la función de error.[3]

En (1.5.10) la distribución inicial se expresa como una suma de funciones seno. Esto revela el significado físico de las series trigonométricas en (1.5.12), donde cada término se corresponde con un término en la serie de Fourier (1.5.10) que representa a la distribución inicial.

1.6 Ecuaciones de dispersión

El factor de proporcionalidad D en la Ley de Fick (1.5.1) en general no es constante, por ejemplo, cuando se considera un fluido a través de un medio poroso homogéneo. Es por ello que se requiere un nuevo ajuste a la Ley de Fick si se considera el fenómeno de advección, como se puede consultar en [41, 28].

En estos fenómenos donde la difusividad no es constante, se sigue presentando una fuerte dependencia con la velocidad del fluido. A este fenómeno se le conoce como *dispersión*, el cual es un fenómeno general que incluye a la difusión como caso particular. Para el caso unidimensional, este fenómeno se describe como:

$$D = \tau D_{\text{mol}} + \alpha_L v. \tag{1.6.1}$$

La dispersividad efectiva, la cual se usa en la Ley de Fick, consiste de dos partes. La primera se origina en la difusión molecular y la segunda en los flujos en medios porosos. En los flujos a altas velocidades predomina la segunda parte, que es la situación más común en las aguas subterráneas, aunque los flujos en los acuíferos son más bien lentos en comparación con otros departamentos hidrólogicos. El factor de proporcionalidad entre la dispersión y la velocidad a lo largo de una trayectoria del flujo está dado por el parámetro α_L, que tiene dimensión de [longitud]. También se suele emplear el término 'longitud de dispersión' o *dispersividad longitudinal*. El subíndice se refiere a longitudinal, ya que solo es válido en la dirección del flujo.

[3] La función de error $\operatorname{erf} z$ es

$$\operatorname{erf} z = \frac{2}{\sqrt{\pi}} \int_0^z e^{-\eta^2}\, d\eta.$$

En el caso de dos o tres dimensiones, es necesario generalizar el concepto de dispersión. El factor α_L como factor de proporcionalidad entre la dispersividad efectiva y la velocidad de filtración ya no es válido cuando se considera el movimiento en dirección transversal a la dirección del flujo. Es por ello que se define la *dispersividad transversal* α_T. La formulación analítica se vuelve un poco más compleja, ya que el factor escalar D_{eff} debe ser reemplazado por el *tensor de dispersión* $\mathbf{D}$:

$$\mathbf{D} = (\tau D_{\text{mol}} + \alpha_T v)\mathbf{I} + \frac{\alpha_L - \alpha_T}{v}\mathbf{v}\mathbf{v}^T. \tag{1.6.2}$$

donde $\mathbf{I}$ es la matriz identidad. La matriz $\mathbf{v}\mathbf{v}^T$ contiene a los productos de los componentes de la velocidad. Esta formulación toma en cuenta que la mezcla constante en la dirección de la velocidad es diferente a la mezcla constante transversal a la dirección de la velocidad y es válida para vectores $\mathbf{v}$ arbitrarios, los cuales pueden variar espacial y temporalmente. Con el tensor de dispersión, el flujo de dispersión se escribe como:

$$\mathbf{j} = -\mathbf{D}\nabla c. \tag{1.6.3}$$

1.7 La ecuación de transporte de masa

Cuando se consideran la advección junto con la difusión o dispersión, el vector de flujo en la dirección x es la suma de ambas contribuciones:

$$j_x = -D\frac{\partial c}{\partial x} + vc. \tag{1.7.1}$$

Se deben de tomar en cuenta diferentes contribuciones en la difusividad D: la escala molecular, la tortuosidad y la dispersión a escala regional. Las componentes y y z del flujo se escriben de manera similar, por lo que el flujo se escribe en notación vectorial como:

$$\mathbf{j} = -\mathbf{D}\nabla c + \mathbf{v}c. \tag{1.7.2}$$

En esta ecuación se considera la difusividad $\mathbf{D}$ en su forma más general (1.6.2). Este resultado se sustituye en la ecuación de conservación de masa (1.3.4), que en una dimesión lleva a:

$$\theta\frac{\partial c}{\partial t} = \frac{\partial}{\partial x}\theta\left(D\frac{\partial c}{\partial x} - vc\right) + q.$$

Para el caso de un flujo a velocidad constante, de esta ecuación se obtiene una de las formulaciones más comunes para la ecuación de transporte:

$$\theta\frac{\partial c}{\partial t} = \frac{\partial}{\partial x}\left(\theta D\frac{\partial c}{\partial x}\right) - \theta v\frac{\partial c}{\partial x} + q.$$

Para el caso de difusividad constante, esta ecuación se escribe como:

$$\theta\frac{\partial c}{\partial t} = \theta D\frac{\partial^2 c}{\partial x^2} - \theta v\frac{\partial c}{\partial x} + q.$$

Para el caso general, esta ecuación se escribe como:

$$\theta \frac{\partial c}{\partial t} = \nabla \bullet \theta(\mathbf{D}\nabla c - \mathbf{v}c) + q. \tag{1.7.3}$$

La ecuación (1.7.3) se conoce como *ecuación de transporte de masa*, y es válida para toda clase de agentes biológicos, geológicos y químicos. Se trata de una ecuación parabólica, que resulta lineal si sus coeficientes son constantes.

En el caso de un flujo incompresible, o libre de divergencia, la ecuación (1.7.3) se simplifica como:

$$\theta \frac{\partial c}{\partial t} = \nabla \bullet (\theta \mathbf{D}\nabla c) - \theta \mathbf{v} \bullet \nabla c + q. \tag{1.7.4}$$

Para una discusión más detallada de la ecuación de transporte de masa, recomiendo consultar [41, 28].

Capítulo 2

Métodos iterativos para optimización

A primera vista, la expresión *problema inverso* puede resultar algo ambigua, matemáticamente hablando, ya que formalidad de la disciplina matemática a menudo exige una definición riguroza para poder denominar a dos conceptos como "inversos" uno del otro. Algunos autores han sugerido denominar a dos problemas como "inversos" si la formulación de un problema involucra al otro. Históricamente se ha denominado como *el problema directo* al problema más simple o al que ha sido estudiado primero, mientras que al segundo problema se le ha denominado como *problema inverso*. Sin embargo, cuando se refiere a problemas físicos, en la mayoría de los casos se sigue una nomenclatura estándar. Por ejemplo, si se desea predecir el comportamiento de un sistema físico a partir de su estado presente y de las leyes físicas involucradas, usualmente se denomina a este problema como "problema directo". Por otro lado, podrían considerarse como "problemas inversos" la determinación de un estado presente a partir de futuras observaciones, o la identifiación de parámetros físicos a partir de la observación de la evolución del sistema.

Desde el punto de vista de las aplicaciones, se pueden distinguir dos motivaciones para estudiar problemas inversos: en primer lugar, se desea conocer estados pasados o parámetros de un sistema físico. En segundo lugar, se desea conocer cómo influenciar a un sistema físico por medio de su estado presente o a través de ciertos parámetros para dirigirlo hacia un estado futuro deseado. Entonces se podría decir que los problemas inversos se ocupan de determinar las causas de un estado o efecto deseado.

2.1 Problemas bien planteados

Para reducir la ambigüedad al definir un problema inverso, el matemático francés Jacques Salomon Hadamard propuso la siguiente definición para que un problema esté bien planteado.

Definición 2.1.1. Se dice que un modelo matemático para un fenómeno físico está **bien planteado** si satisface que:

1. Para todos los datos admisibles, existe una solución.
2. Para todos los datos admisibles, la solución es única.
3. La solución depende continuamente de los datos.

Por supuesto que, matemáticamente hablando, esta no es una definición riguroza, ya que no se especifica la noción de solución, ni se especifica qué datos se consideran admisibles, ni la topología empleada para medir la continuidad. Esta propuesta se emplea de manera adaptativa para cada problema en particular. Si un problema falla en al menos uno de los puntos de la definición (2.1.1), se dice que el problema está **mal planteado**.

En relación a los tres puntos propuestos en (2.1.1), en un problema mal planteado, muchas ocasiones en la práctica es posible relajar el primer punto. Si bien es muy importante contar con al menos una solución para un problema inverso, en ocasiones es posible relajar la noción de solución al problema para encontrar así un modelo adecuado al problema físico.

El incumplimiento del segundo punto de (2.1.1) es un asunto más delicado, ya que si se encuentran varias soluciones para un mismo problema, es necesario definir criterios adicionales para decidir qué solución es más adecuada para el problema en cuestión. Se debe verificar la completez del modelo, y de ser posible, agregar información adicional. La cuestión de la unicidad es relevante en problemas inversos donde se busca una causa para un efecto observado. Si solo se desea encontrar una causa para un efecto deseado, puede que sea conveniente contar con una variedad de posibles soluciones, a partir de las cuales se puede elegir una a través de criterios adicionales.

En relación al punto 2 de (2.1.1), cabe señalar que aún si esta condición se cumple para todos los datos medidos en un problema físico, la disretización requerida para el análisis del problema hace que sea más complicado conseguir la unicidad para una solución. La discretización de un problema físico o cualquier otra aproximación numérica conducen eventualmente a problemas en dimensión finita, por ejemplo, sistemas de ecuaciones lineales cuya solución probablemente no es única.

Para ilustrar el concepto de problemas mal planteados, considere el siguiente ejemplo:

Ejemplo 2.1.2. Dos problemas inversos clásicos en matemáticas son la diferenciación y la integración. En principio, no es claro cuál de estos dos problemas debería ser considerado como problema directo y cuál como problema inverso. Sin embargo, el proceso de diferenciación posee propiedades de un problema mal planteado.

Sea $f \in C^1[0,1]$ una función, $\delta \in (0,1)$, $n \in \mathbb{N}$ $(n \geqslant 2)$ arbitrario. Defina

$$f_{\delta,n}(x) := f(x) + \delta \sin\left(\frac{nx}{\delta}\right) \quad x \in [0,1]. \tag{2.1.3}$$

Entonces

$$f'_{\delta,n}(x) := f'(x) + n \cos\left(\frac{nx}{\delta}\right) \quad x \in [0,1]. \tag{2.1.4}$$

Entonces, en la norma uniforme,

$$||f - f_{\delta,n}||_\infty = \delta,$$

pero

$$||f' - f'_{\delta,n}||_\infty = n.$$

De esta manera, si se considera a f y $f_{\delta,n}$ como los datos exactos y perturbados, respectivamente, entonces para un error δ arbitrariamente pequeño, el error en el resultado, es decir en la derivada, puede ser arbitrariamente grande. Por tanto, la derivada no depende de manera continua de los datos, con respecto a la norma uniforme, fallando así el tercer punto de (2.1.1). Por supuesto que se podría forzar la dependencia de manera continua midiendo el error en los datos con la norma C^1, pero esto sería como hacer trampa, ya que en ese caso se pensaría que el error en los datos es pequeño si el error en los valores de la función y en los valores de la derivada, que es lo que se quiere calcular, fuese pequeño.

Por otro lado, observe que f' es solución a la ecuación integral simple del primer tipo

$$(Sx)(s) := \int_0^t x(s)\, ds = f(t) - f(0),$$

la cual tiene solución en $C[0,1]$ solo si $f \in C^1[0,1]$, por supuesto. El problema directo correspondiente sería calcular f a partir de x, es decir, la integración, el cual es un proceso estable en $C[0,1]$. Note que la integración suaviza, por ejemplo, errores altamente oscilatorios en x (por ejemplo, como los de (2.1.4)) son amortiguados (a $\delta \sin(nx/\delta)$) y tienen un efecto muy pequeño en los datos para el problema inverso. Este alizamiento es responsable del hecho que errores de amplitud pequeña pero alta frecuencia, creen oscilaciones grandes en la solución al problema inverso. Estas consideraciones no se restringen a este problema en concreto: cada vez que un problema directo tiene propiedades de suavizamiento o alizamiento, se espera la aparición de oscilaciones que se originan en perturbaciones pequeñas en los datos de la solución del problema inverso. Este efecto se acentúa más a medida que se suaviza más la solución al problema directo.

2.2 Identificación de parámetros

La identificación de parámetros se refiere a la estimación de los coeficientes en una ecuación diferencial a partir de observaciones de la solución a dicha ecuación. A estos coeficientes se les conoce como *parámetros del sistema*, y la solución y sus derivadas constituyen las *variables de estado*. El problema directo consiste en calcular las variables de estado dados los parámetros del sistema y condiciones de frontera adecuadas. El problema directo es típicamente bien planteado, mientras que la identificación de parámetros, el problema inverso de interés en este proyecto, es usualmente un problema mal planteado (en el sentido de (2.1.1)). Más aún, si el problema directo es lineal en las variables de estado, la identificación de parámetros es en general un problema no lineal, como se muestra en [69, 87]. Para ilustrar este punto, considere los siguientes ejemplos:

Ejemplo 2.2.1. Considere la ecuación del oscilador armónico amortiguado

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = f(t), \quad t > 0, \tag{2.2.2}$$

con condiciones iniciales $x(0) = x_0$, $\frac{dx}{dt}(0) = v_0$. $x(t)$ representa el desplazamiento de la masa al tiempo t, y $\frac{dx}{dt}$ representa su velocidad. x y $\frac{dx}{dt}$ son las variables de estado, mientras que la masa m, el coeficiente de amortiguamiento c, la constante del resorte k y la función de fuerza externa $f(t)$ conforman los parámetros del sistema. El problema directo consiste en determinar las variables de estado, dados los parámetros del sistema y el estado inicial (x_0, v_0). Dependiendo de la información disponible, se pueden formular varios problemas inversos. Por ejemplo, se pueden despreciar la fuerza y el amortiguamiento. Se podría desplazar el sistema y tratar de determinar m y k a partir de la observación del movimiento resultante. Este problema está mal planteado en el sentido de que no se puede determinar de manera única m y k a partir de los datos observados. La solución a la ecuación (2.2.2) con $c = f = v_0 = 0$ es

$$x(t) = x_0 \cos(\omega t), \quad \omega = \sqrt{\frac{k}{m}}, \tag{2.2.3}$$

donde ω representa la frecuencia de oscilación del sistema. A partir de ciertas observaciones del estado, por ejemplo, del desplazamiento $x(t)$ sobre un intervalo de longitud $2\pi/\omega$, es posible determinar de manera única a ω. Sin embargo, a partir de ω sólo se puede determinar el cociente k/m. A partir de (2.2.3) se puede calcular

$$\omega = \frac{1}{t}\cos^{-1}\left(\frac{x(t)}{x_0}\right).$$

Esto establece que la dependencia de ω con $x(t)$ no es lineal, y por tanto la dependencia de k u ω (siempre que se conozca a uno de ellos) con ω es no lineal.

Ejemplo 2.2.4. Considere la ecuación de difusión estacionaria en una dimensión

$$-\frac{d}{dx}\left(D(x)\frac{du}{dx}\right) = f(x), \quad 0 < x < 1, \tag{2.2.5}$$

con condiciones de frontera

$$u(0) = u_L, \quad u(1) = u_R.$$

Esta ecuación modela la distribución de temperatura en estado estacionario dentro de una barra metálica delgada. En este caso, la variable de estado es la distribución de temperatura $u(x)$, $0 < x < 1$. Los parámetros del sistema son el coeficiente de difusión $D(x)$ y el término correspondiente a la fuente de calor $f(x)$. Como estas son funciones y no escalares, la ecuación (2.2.5) se conoce como sistema de parámetros distribuidos. El problema directo (2.2.5), junto con sus condiciones de frontera de Dirichlet, es bien planteado cuando la función $D(x)$ es suficientemente diferenciable y está suficientemente lejos del cero.

Ahora considere el siguiente problema inverso: dados $f(x)$ y $u(x)$ para $0 < x < 1$, estime el valor de $D(x)$. Formalmente,

$$D(x) = \int_{y=0}^{x} \frac{f(y)\,dy}{u_x}. \tag{2.2.6}$$

Donde $u_x = \frac{du}{dx}$. De esta expresión se puede notar que la dependencia de D con $\frac{du}{dx}$ (y por tanto con u) es no lineal. Esta expresión también ejemplifica el mal planteamiento del problema. Si $\frac{du}{dx} = 0$ dentro de un subintervalo de $(0, 1)$, entonces $D(x)$ no estaría determinado de manera única dentro de dicho subintervalo. La falta de continuidad en la dependencia con la variable de estado es un detalle más sutil. Considere la perturbación parametrizada $\delta u = \varepsilon \sin(x/\varepsilon^2)$. Esta perturbación se anula cuando $\varepsilon \to 0$. La perturbación corrrespondiente en la derivada, $\frac{d}{dx}\delta u = \cos(x/\varepsilon^2)/\varepsilon$, crece arbitrariamente a medida que $\varepsilon \to 0$. De (2.2.6) se observa que la perturbación correspondiente en $D(x)$ crece arbitrariamente a medida que $\varepsilon \to 0$.

2.3 Mínimos cuadrados: una perspectiva abstracta

Una técnica empleada en la identificación de parámetros es la formulación usando mínimos cuadrados. A continuación se presenta una breve introducción con algunas generalidades sobre esta formulación, las cuales se presentan con mayor detalle en [69, 87, 18]. Considere el sistema abstracto de parámetros distribuidos:

$$A(q)u = f, \tag{2.3.1}$$

donde q representa la distribución de parámetros que se desea estimar, $A(q)$ representa un operador que depende de los parámetros, y u representa la correspondiente variable de estado. A la ecuación (2.3.1) se le denomina ecuación de estado. En el ejemplo (2.2.4), q representa el coeficiente de difusión D, y $A(D) = -\frac{d}{dx}(D(x)\frac{d}{dx}(\cdot))$ es el operador de difusión de (2.2.5).

Suponga que los datos observados se pueden expresar como

$$d = Cu + \eta \tag{2.3.2}$$

donde η representa ruido en los datos. C se denomina el mapeo de estado a observación. Por ejemplo, si la variable de estado se mide en n puntos discretos x_i, entonces

$$[Cu]_i = u(x_i), \quad i = 1, \ldots, n.$$

Para fines de los cálculos, se requerirán varios espacios abstractos de funciones. Sea $\mathcal{Q}$ el espacio de parámetros, el cual contiene al parámetro q, sea $\mathcal{U}$ el espacio de estado, el cual contiene a la variable u, y sea $\mathcal{V}$ el espacio de observación, el cual contiene a los datos observados d. Por simplicidad, se asume que estos tres espacios son espacios de Hilbert.

El problema inverso consiste en estimar q en (2.3.1) dados los datos d en (2.3.2). Una forma de atacar este problema es resolver el problema de minimización de mínimos cuadrados regularizados restringidos

$$\min_{u\in\mathcal{U},q\in\mathcal{Q}} \frac{1}{2}||Cu-d||_{\mathcal{V}}^2 + \alpha J(q), \quad \text{sujeto a} \quad A(q)u = f. \tag{2.3.3}$$

Aquí $J(q)$ es un funcional de regularización, incorporado con la finalidad de imponer estabilidad, información a priori o ambas, y α es un parámetro positivo de regularización. A este enfoque se le conoce como salida de mínimos cuadrados regularizados.

Suponga que el problema directo, resulver para u en (2.3.1), es bien plantedo, y denote la solución como

$$u = A(q)^{-1}f. \tag{2.3.4}$$

A partir de (2.3.3) se puede obtener el problema de minimización de mínimos cuadrados regularizados no restringidos

$$\min_{q\in\mathcal{Q}} T(q), \quad T(q) = \frac{1}{2}||F(q)-d||_{\mathcal{V}}^2 + \alpha J(q), \tag{2.3.5}$$

donde ahora

$$F(q) = CA(q)^{-1}f. \tag{2.3.6}$$

El mapeo $F : \mathcal{Q} \to \mathcal{V}$ se conoce usualmente como mapeo de parámetro a observación.

Existen muchos algoritmos para resolver computacionalmente el problema (2.3.5); sin embargo, muchos de esos algoritmos se enfocan en funciones suficientemente diferenciables, digamos, de clase C^2. Como puede esperarse, muchos problemas físicos involucran funciones que no son suficientemente diferenciables, o incluso discontinuas, para las cuales no es posible en general identificar a un minimizador.

Si la función en cuestión fuese continua en todo su dominio y diferenciable por piezas, se podría identificar una solución examinando el subgradiente o el gradiente generalizado, que son generalizaciones del concepto de gradiente para funciones que no son suficientemente diferenciables. Como un caso en particular, se podrían estudiar las funciones

$$f(x) = ||r(x)||_1, \quad f(x) = ||r(x)||_\infty,$$

donde $r(x)$ es una función vectorial, como un problema de optimización para ciertas funciones que no sean lo suficientemente diferenciables.

Para fines de estudiar el problema (2.3.5), se sugieren algunos métodos iterativos, para lo cual se sugiere comenzar por considerar las estrategias de búsqueda en línea y región de confianza.

2.4 Búsqueda en línea y región de confianza

En la optimización no restringida, se busca minimizar una función objetivo que depende de variables reales, las cuales no están sujetas a restricciones. La formulación básica es

$$\min_x f(x),$$

donde $x \in \mathbf{R}^n$ con $n \geqslant 1$ y $f : \mathbf{R}^n \to \mathbf{R}$ es una función (preferentemente suave).

En la búsqueda en línea, el algoritmo elige una dirección p_k y busca a lo largo de esta dirección desde la iteración actual x_k hacia una nueva iteración con un valor menor de la función. La distancia a recorrer a lo largo de p_k se puede encontrar aproximando una solución al prolema de optimización en una dimensión para encontrar el tamaño de paso α:

$$\min_{\alpha>0} f(x_k + \alpha p_k). \tag{2.4.1}$$

Al resolver (2.4.1) de manera exacta se tendría el máximo beneficio en la dirección p_k, pero dicha minimización exacta puede resultar costosa, y usualmente es innecesaria. En su lugar, el algoritmo de búsqueda en línea genera un número limitado de longitudes de prueba hasta encontrar una que aproxime lo suficiente una solución para (2.4.1). En la siguiente iteración, se calculan una nueva dirección y un nuevo tamaño de paso, y el proceso se repite.

En el algoritmo conocido como región de confianza, la información reunida sobre f se usa para construir una función modelo m_k cuyo comportamiento cerca del punto actual x_k sea similar al de la función objetivo f. Debido a que el modelo m_k podría no ser una buena aproximación a f cuando x está lejos de x_k, se restringe la búsqueda del minimizador de m_k a alguna vecindad de x_k. Dicho de otra forma, se puede encontrar un candidato de paso p al resolver aproximadamente el problema

$$\min_p m_k(x_k + p), \tag{2.4.2}$$

donde $x_k + p$ está en la región de confianza. Si el candidato a solución no produce suficiente decrecimiento en f, se concluye que la región es demasiado grande, así que se reduce el radio de la región y se resuelve nuevamente (2.4.2). Usualmente, la región de confianza es una bola definida por $||p||_2 \leqslant \Delta$, donde el escalar $\Delta > 0$ se denomina *radio de la región de confianza*. Tambien se pueden usar regiones de confianza en forma de cuadrados o elipses.

El modelo m_k de (2.4.2) usualmente se define como una función cuadrática de la forma

$$m_k(x_k + p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p, \tag{2.4.3}$$

donde f_k, ∇f_k y B_k son escalar, vector y matriz, respectivamente. Como lo indica la notación, f_k y ∇f_k se eligen como los valores de la función y del gradiente evaluados en

el punto x_k, por lo que m_k y f coinciden hasta el primer orden en la iteración actual x_k. La matriz B_k es o bien el Hessiano $\nabla^2 f_k$ o alguna aproximación a él.

Los enfoques de la búsqueda en línea y la región de confianza difieren en el orden en que eligen la dirección y la distancia de la siguiente iteración. La búsqueda en línea inicia arreglando la dirección p_k y luego identifica una distancia apropiada, el tamaño de paso α_k. En la región de confianza, primero se elige la máxima distancia - el radio de la región de confianza Δ_k - y luego se busca una dirección y un paso que alcance la mejor aproximación posible sujeto a la restricción de distancia. Si el paso no es satisfactorio, se reduce el radio Δ_k y se intenta de nuevo.

2.4.1 Búsqueda de direcciones para métodos de búsqueda en línea

La dirección de descenso más pronunciado $-\nabla f_k$ es la elección más obvia para métodos de búsqueda en línea. Es intuitiva; de entre todas las direcciones en que se podría mover a partir de x_k, es la dirección en la que f decrece más rápidamente, como lo confirma la expansión de Taylor, la cual dice que para cualquier dirección p y cualquier tamaño de paso α, se tiene

$$f(x_k + \alpha p) = f(x_k) + \alpha p^T \nabla f_k + \frac{1}{2}\alpha^2 p^T \nabla^2 f(x_k + tp)p, \quad \text{para alguna } t \in (0, \alpha).$$

La tasa de cambio de f a lo largo de la dirección de p en x_k es simplemente el coeficiente de α, o sea $p^T \nabla f_k$. Por tanto, la dirección unitaria p de más rápido descenso es la solución al problema

$$\min_p p^T \nabla f_k, \quad \text{sujeto a } ||p|| = 1.$$

Dado que $p^T \nabla f_k = ||p||\,||\nabla f_k|| \cos\theta = ||\nabla f_k|| \cos\theta$, donde θ es el ángulo entre p y ∇f_k, es fácil ver que el minimizador se alcanza cuando $\cos\theta = -1$ y

$$p = -\frac{\nabla f_k}{||\nabla f_k||},$$

como se afirmó.

El método de descenso más pronunciado es un método de búsqueda en línea que se mueve a lo largo de $p_k = -\nabla f_k$ en cada paso. Puede elegir el tamaño de paso α_k de varias maneras. Una ventaja de la dirección de descenso más pronunciado es que requiere calcular el gradiente ∇f_k pero no segundas derivadas. Sin embargo, este método puede ser insoportablemente lento en problemas complejos.

Los métodos de búsqueda en línea pueden usar cualquier dirección distinta al descenso más pronunciado. En general, cualquier dirección descendiente - una que forme un ángulo de menos de $\pi/2$ radianes con $-\nabla f_k$ - garantiza producir un decremento

en f, siempre que el tamaño de paso sea suficientemente pequeño, lo cual se puede verificar usando el Teorema de Taylor:

$$f(x_k + \varepsilon p_k) = f(x_k) + \varepsilon p_k^T \nabla f_k + O(\varepsilon^2).$$

Cuando p_k es una dirección descendiente, el ángulo θ_k entre p_k y ∇f_k tiene $\cos\theta_k < 0$, por lo que

$$p_k^T \nabla f_k = ||p_k||\,||\nabla f_k|| \cos\theta_k < 0.$$

Se sigue que $f(x_k + \varepsilon p_k) < f(x_k)$ para todos los valores positivos ε suficientemente pequeños.

Otra dirección de búsqueda importante - tal vez la más importante de todas - es la *dirección de Newton*. Esta dirección se obtiene de la expansión de Taylor hasta segundo orden

$$f(x_k + p) \approx f_k + p^T \nabla f_k + \frac{1}{2} p^T \nabla^2 f_k p := m_k(p). \tag{2.4.4}$$

Si se asume de momento que $\nabla^2 f_k$ es positiva definida, se obtiene la dirección de Newton al encontrar el vector p que minimiza $m_k(p)$. Al igualar a cero la derivada de $m_k(p)$, se obtiene:

$$p_k^N = -(\nabla^2 f_k)^{-1} \nabla f_k. \tag{2.4.5}$$

La dirección de Newton es confiable cuando la diferencia entre $f(x_k+p)$ y $m_k(p)$ no es muy grande. Al comparar (2.4.5) con la expansión de Taylor, se observa que la única diferencia entre estas funciones es que la matriz $\nabla^2 f(x_k + tp)$ en el tercer término de la expansión fue remplazado por $\nabla^2 f_k$. Si $\nabla^2 f$ es suficientemente suave, esta diferencia introduce una perturbación de sólo $O(||p||^3)$ dentro de la expansión, así que si $||p||$ es pequeño, la aproximación $f(x^k + p) \approx m_k(p)$ es bastante precisa.

La dirección de Newton puede ser usada en métodos de búsqueda en línea cuando $\nabla^2 f_k$ es positiva definida, ya que en dicho caso se tiene

$$\nabla f_k^T p_k^N = -p_k^N \nabla^2 f_k p_k^N \leqslant -\sigma_k ||p_k^N||^2$$

para algún $\sigma_k > 0$. A menos de que el gradiente ∇f_k (y por tanto el paso p_k^N) sea cero, se tiene que $\nabla f_k^T p_k^N < 0$, y por tanto la dirección de Newton es descendente.

A diferencia de la dirección de descenso más pronunciado, existe un tamaño de paso "natural" $\alpha = 1$ asociado a la dirección de Newton, que es usado en la mayoría de sus implementaciones.

Si $\nabla^2 f_k$ no es positiva definida, puede ser que la dirección de Newton no esté definida, dado que $(\nabla^2 f_k)^{-1}$ pudiera no existir. Aún cuando esté definida, podría no satisfacer la propiedad descendente $\nabla f_k^T p_k^N < 0$, en cuyo caso no es una dirección

conveniente. En estas situaciones, los métodos de búsqueda en línea modifican la definición de p_k para hacerla que satisfaga la condición descendente mientras conservan el beneficio de la información de segundo orden contenida en $\nabla^2 f_k$. Por otro lado, si $A = \nabla^2 f_k$ es una matríz con entradas no negativas fuera de la diagonal y al menos una entrada negativa en la diagonal, de tal suerte que A no es positiva definida, se puede considerar la traslación $\tilde{A} = A - \text{mín}(a_{ii})I$, donde $\text{mín}(a_{ii}) < 0$. $\tilde{A}$ es una matríz positiva cuyo espectro es el mismo de la matriz A pero trasladado hacia la derecha (ver [9, 8]).

Los métodos que usan la dirección de Newton tienen una tasa rápida de convergencia local, típicamente cuadrática. Después de que se llega a una vecindad de la solución, la convergencia con alta precisión ocurre en sólo unas pocas iteraciones. La desventaja principal de la dirección de Newton es la necesidad del Hessiano $\nabla^2 f(x)$. El cálculo explícito de esta matriz puede resultar en un proceso incómodo y costoso.

Los métodos Cuasi-Newton de búsqueda en línea proporcionan una alternativa atractiva al método de Newton sin tener que calcular el Hessiano y aún así alcanzan una tasa superlineal de convergencia. En lugar del Hessiano $\nabla^2 f_k$, se usa una aproximación B_k, la cual se actualiza en cada paso para tomar en cuenta la información adquirida en cada paso. Las actualizaciones hacen uso del hecho de que los cambios en el gradiente proporcionan información sobre la segunda derivada de f a lo largo de la dirección de búsqueda. A partir del desarrollo de Taylor, se tiene que

$$\nabla f(x+p) = \nabla f(x) + \nabla^2 f(x)p + \int_0^1 [\nabla^2 f(x+tp) - \nabla^2 f(x)]p\, dt.$$

Dado que $\nabla f(\cdot)$ es continuo, el tamaño de la integral es $o(||p||)$. Al hacer $x = x_k$ y $p = x_{k+1} - x_k$, se obtiene

$$\nabla f_{k+1} = \nabla f_k + \nabla^2 f_k(x_{k+1} - x_k) + o(||x_{k+1} - x_k||).$$

Cuando x_k y x_{k+1} están en una vecindad de la solución x^*, dentro de la cual $\nabla^2 f$ es positiva definida, el término final in esta expansión es eventualmente dominado por el término $\nabla^2 f_k(x_{k+1} - x_k)$, y entonces se puede escribir

$$\nabla^2 f_k(x_{k+1} - x_k) \approx \nabla f_{k+1} - \nabla f_k. \tag{2.4.6}$$

Se puede elegir una nueva aproximación al Hessiano B_{k+1} que imite la propiedad (2.4.6) del verdadero Hessiano, o sea, que cumpla la siguiente condición, conocida como la *ecuación secante*:

$$B_{k+1}s_k = y_k, \quad s_k = x_{k+1} - x_k, \quad y_k = \nabla f_{k+1} - \nabla f_k. \tag{2.4.7}$$

Usualmente se imponen condiciones adicionales a B_{k+1}, tales como simetría (motivado por el Hessiano verdadero), y el requisito de que la diferencia entre aproximaciones sucesivas B_k y B_{k+1} tenga un rango bajo.

Dos de las más populares formulaciones para actualizar la aproximación al Hessiano B_k son la *fórmula simétrica de rango uno* (SR1 por sus siglas en inglés), definida por:

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T s_k}, \tag{2.4.8}$$

y la *fórmula BFGS*, llamada así por sus inventores, Broyden, Fletcher, Goldfarb y Shanno, la cual se define como:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}. \tag{2.4.9}$$

La diferencia entre las matrices B_k y B_{k+1} es una matriz de rango uno en el caso de (2.4.8) y una matriz de rango dos en el caso de (2.4.9). Ambas actualizaciones satisfacen la ecuación secante y ambas mantienen la simetría. Se puede demostrar que la actualización (2.4.9) genera actualizaciones positivas definidas siempre que la aproximación inicial B_0 sea positiva definida y $s_k^T y_k > 0$.

En los métodos cuasi-Newton, la dirección de búsqueda se obtiene usando B_k en lugar del Hessiano en (2.4.5), es decir

$$p_k = -B_k^{-1} \nabla f_k. \tag{2.4.10}$$

Algunas implementaciones de los métodos cuasi-Newton evitan la necesidad de factorizar B_k en cada iteración al actualizar el inverso de B_k en lugar de B_k. De hecho, la fórmula equivalente a (2.4.8) y (2.4.9), aplicada a la aproximación de la inversa $H_k := B_k^{-1}$ es

$$H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T, \quad \rho_k = \frac{1}{y_k^T s_k}. \tag{2.4.11}$$

De esta manera, p_k se puede calcular haciendo $p_k = -H_k \nabla f_k$.

Otra clase de direcciones de búsqueda son las generadas por los *métodos no lineales de gradiente conjugado*, los cuales tienen la forma:

$$p_k = -\nabla f(x_k) + \beta_k p_{k-1},$$

donde β_k es un escalar que asegura que p_k y p_{k-1} sean conjugados. Los métodos de gradiente conjugado fueron diseñados originalmente para resolver sistemas de ecuaciones lineales $Ax = b$, donde la matriz de coeficientes A es simétrica y definida positiva. Este problema es equivalente al problema de minimizar la función cuaadrática convexa definida por

$$\phi(x) = \frac{1}{2} x^T A x - b^T x,$$

así que era natural investigar extensiones de estos algoritmos para tipos más generales de problemas de optimización no restringidos. En general, los métodos no lineales de gradiente conjugado proporcionan direcciones de búsqueda que son mucho más efectivas que la dirección de descenso más empinada y son casi igual de simples de calcular.

Estos métodos no alcanzan las tasas de convergencia de los métodos de Newton o cuasi-Newton, pero tienen la ventaja de que no requieren almacenar matrices.

Todas las direcciones de búsqueda discutidas hasta ahora se pueden usar en una búsqueda en línea, y dan origen a los métodos de búsqueda en línea de descenso más pronunciado, Newton, cuasi-Newton y gradiente conjugado. Con excepción del gradiente conjugado, todos los métodos tienen un análogo en los métodos de región de confianza.

2.4.2 Modelos para métodos de región de confianza

Si se hace $B_k = 0$ en (2.4.3) y se define la región de confianza usando la norma euclideana, el problema (2.4.2) se convierte en:

$$\min_p f_k + p^T \nabla f_k \quad \text{sujeto a } ||p||_2 \leqslant \Delta_k.$$

Se puede escribir la solución a este problema como:

$$p_k = -\frac{\Delta_k \nabla f_k}{||\nabla f_k||}.$$

Esto es un paso en la dirección más pronunciada donde el tamaño de paso se determina por el radio de la región de confianza; la región de confianza y la búsqueda en línea plantean básicamente el mismo enfoque en este caso.

Un algoritmo de región de confianza más interesante se obtiene al elegir B_k como el Hessiano exacto $\nabla^2 f_k$ en (2.4.3). Debido a la restricción de la región de confianza $||p||_2 \leqslant \Delta_k$, se garantiza que el problema (2.4.2) tiene solución aún cuando $\nabla^2 f_k$ no es positiva definida. El método de Newton para región de confianza ha probado ser muy efectivo en la práctica.

Si la matriz B_k en la función cuadrática modelo m_k de (2.4.3) se define usando una aproximación de tipo cuasi-Newton, se obtiene un método cuasi-Newton de región de confianza.

Uno de los métodos empleados en este proyecto para resolver el problema (2.3.5) está basado en un método de región de confianza, y se le conoce en la literatura como *Método de pata de perro*.

2.4.3 Método de pata de perro

En esta sección, se asume que la función modelo m_k que se usa en cada iteración x_k es cuadrática, y se basa en el desarrollo de Taylor de f alrededor de x_k:

$$f(x_k + p) = f_k + g_k^T p + \frac{1}{2} p^T \nabla^2 f(x_k + tp) p, \tag{2.4.12}$$

donde $f_k = f(x_k)$, $g_k = \nabla f(x_k)$ y $t \in (0,1)$. Usando una aproximación de segundo orden B_k al Hessiano, se define m_k como:

$$m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p, \tag{2.4.13}$$

donde B_k es una matriz simétrica. La diferencia entre $m_k(p)$ y $f(x_k + p)$ es $O(||p||^2)$, la cual es pequeña si p lo es.

Cuando B_k es igual al Hessiano $\nabla^2 f(x_k)$, el error de aproximación en la función modelo m_k es $O(||p||^3)$, por lo que este método es especialmente preciso cuando $||p||$ es pequeño. La elección $B_k = \nabla^2 f(x_k)$ lleva al método de Newton de región de confianza.

Para obtener cada paso, se busca una solución al problema

$$\min_{p \in \mathbf{R}^n} m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p \quad \text{sujeto a } ||p|| \leqslant \Delta_k, \tag{2.4.14}$$

donde $\Delta_k > 0$ es el radio de la región de confianza. Usualmente se usa $||\cdot||$ como la norma Euclideana, por lo que la solución p_k^* de (2.4.14) es el minimizador de m_k en la bola de radio Δ_k. Entonces, el enfoque de región de confianza requiere que se resuelva una sucesión de problemas (2.4.14) en la que la función objetivo y la restricción (que se puede escribir como $p^T p \leqslant \Delta_k^2$) son ambas cuadráticas. Cuando B_k es positiva definida y $||B_k^{-1} g_k|| \leqslant \Delta_k$, la solución de (2.4.14) es fácil de identificar $-$ es símplemente el mínimo $p_k^B = -B_k^{-1} g_k$ de la cuadrática $m_k(p)$. En este caso, p_k^B se denomina *paso completo*. La solución de (2.4.14) no es tan obvia en otros casos, pero se puede encontrar usualmente sin mucho costo computacional. En cualquier caso, se requiere solo una solución aproximada para obtener convergencia y un buen comportamiento.

Uno de los ingredientes principales en los algoritmos de región de confianza es la estrategia para elegir el radio de la región Δ_k en cada iteración. Esta elección se basa en el acuerdo entre la función modelo m_k y la función objetivo f en iteraciones previas. Dado un paso p_k se define el cociente

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}; \tag{2.4.15}$$

al numerador se le llama *reducción real*, y al denominador se le llama *reducción esperada* (o sea, la reducción en f esperada por la función modelo). Observe que como el paso p_k se obtiene al minimizar el modelo m_k sobre una región que incluye $p = 0$, la reducción esperada siempre será no negativa. Por tanto, si ρ_k es negativo, el nuevo valor objetivo $f(x_k + p_k)$ será mayor que el valor actual $f(x_k)$, por lo que este paso debe ser rechazado. Por otro lado, si ρ_k es cercano a uno, hay un buen acuerdo entre el modelo m_k y la función f en ese paso, por lo que es seguro ampliar la región de confianza en la siguiente iteración. Si ρ_k es positivo pero significativamente menor que uno, no se altera la región de confianza, pero si es cercano a cero o negativo, se reduce la región de confianza al reducir Δ_k en la siguiente iteración. Para la implementación de este

algoritmo, es necesario enfocarse en el problema (2.4.14). En la siguiente discusión, se omite el subíndice k para reescribir el problema (2.4.14) de la siguiente manera:

$$\min_{p\in\mathbf{R}^n} m(p) := f + g^T p + \frac{1}{2}p^T B p \quad \text{sujeto a } ||p|| \leqslant \Delta. \tag{2.4.16}$$

Un primer paso para caracterizar las soluciones de (2.4.16) es el siguiente resultado, que demuestra que la solución p^* de (2.4.16) satisface

$$(B + \lambda I)p^* = -g, \tag{2.4.17}$$

para alguna $\lambda \geqslant 0$.

Teorema 2.4.18. *El vector p^* es una solución global al problema de región de confianza*

$$\min_{p\in\mathbf{R}^n} m(p) = f + g^T p + \frac{1}{2}p^T B p, \quad \textit{sujeto a } ||p|| \leqslant \Delta, \tag{2.4.19}$$

sí y sólamente si p^* es factible y existe un escalar $\lambda \geqslant 0$ que satisface las siguientes condiciones:

$$(B + \lambda I)p^* = -g, \tag{2.4.20a}$$
$$\lambda(\Delta - ||p^*||) = 0, \tag{2.4.20b}$$
$$(B + \lambda I) \textit{ es positiva semidefinida.} \tag{2.4.20c}$$

Para una demostración del teorema (2.4.18), consulte [69].

El método de pata de perro es una estrategia para encontrar soluciones aproximadas al problema (2.4.14), con el cual se logra al menos tanta reducción en m_k como con el *punto de Cauchy*[1]. Este método es apropiado cuando el modelo del Hessiano B_k es positiva definida.

Para motivar este método, se comienza examinando el efecto del radio de la región de confianza Δ en la solución $p^*(\Delta)$ del problema (2.4.16). Cuando B es positiva definida, el minimizador sin restricciones de m es $p^B = -B^{-1}g$. Cuando este punto es factible para (2.4.16), obviamente es una solución, por lo que

$$p^*(\Delta) = p^B, \quad \text{cuando } \Delta \geqslant ||p^B||. \tag{2.4.21}$$

Cuando Δ es pequeño en comparación a p^B, la restricción $||p|| \leqslant \Delta$ asegura que el término cuadrático en m tiene un efecto pequeño en la solución de (2.4.16). Para tal Δ, se puede obtener una aproximación a $p(\Delta)$ al simplemente omitir el término cuadrático de (2.4.16) y escribir

$$p^*(\Delta) \approx -\Delta\frac{g}{||g||}, \quad \text{cuando } \Delta \text{ es pequeña.} \tag{2.4.22}$$

El método de pata de perro encuentra una solución aproximada al remplazar la trayectoria curva de $p^*(\Delta)$ con una trayectoria que consiste en dos segmentos rectos.

[1]Este punto es el minimizador de m_k a lo largo de la dirección de descenso más pronunciado g_k.

El primer segmento va del origen al minimizador de m sobre la trayectoria de descenso más pronunciado, la cual es

$$p^U = -\frac{g^T t}{g^T B g} g, \tag{2.4.23}$$

mientras que el segundo segmento va de p^U a p^B. Formalmente, se puede denotar a esta trayectoria como $\tilde{p}(\tau)$ para $\tau \in [0, 2]$, donde

$$\tilde{p}(\tau) = \begin{cases} \tau p^U, & 0 \leqslant \tau \leqslant 1, \\ p^U + (\tau - 1)(p^B - p^U), & 1 \leqslant \tau \leqslant 2 \end{cases} \tag{2.4.24}$$

El método de pata de perro elige p para qie minimice el modelo m a lo largo de esta trayectoria, sujeto a la condición de región de confianza. El siguiente lema muestra que el mínimo sobre la trayectoria de pata de perro se puede encontrar fácilmente.

Lema 2.4.25. *Sea B una matriz positiva definida. Entonces*

1. *$||\tilde{p}(\tau)||$ es una función creciente de τ, y*
2. *$m(\tilde{p}(\tau))$ es una función decreciente de τ.*

La demostración del lema (2.4.25) se puede consultar en [69]. De este lema se sigue que la trayectoria $\tilde{p}(\tau)$ intersecta a la frontera de la región de confianza $||p|| = \Delta$ en exactamente un punto si $||p^B|| \geqslant \Delta$, y en ningún otro punto si $||p^B|| < \Delta$. Como m es decreciente sobre la trayectoria, el valor elegido de p será p^B si $|p^B|| \leqslant \Delta$, y de otra manera en el punto de intersección de la pata de perro y la frontera de la región de confianza. En este último caso, se calcula el valor apropiado de τ resolviendo la siguiente ecuación cuadrática escalar:

$$||p^U + (\tau - 1)(p^B - p^U)||^2 = \Delta^2.$$

Si el Hessiano exacto $\nabla^2 f(x_k)$ se puede usar en el problema (2.4.16), si es positiva definida, se puede hacer $B = \nabla^2 f(x_k)$ (o sea, $p^B = (\nabla^2 f(x_k))^{-1} g_k$) y aplicar el mismo procedimiento para calcular el paso de Newton-pata de perro. De otra manera, se puede definir p^B eligiendo B como una aproximación al Hessiano que sea positiva definida, y proceder para encontrar el paso de pata de perro. Dado que p^B satisface las condiciones de segundo orden, una vez que esté cerca de la solución, esta elección permitirá una convergencia rápida del método de Newton.

El uso de un Hessiano modificado en el método de Newton-pata de perro no es del todo satisfactoria desde un punto de vista intuitivo. Una factorización modificada perturba las diagonales de $\nabla^2 f(x_k)$ de una manera un tanto arbitraria, y los beneficios del enfoque de región de confianza podrían no verse realizados. De hecho, la modificación introducida durante la factorización del Hessiano es redundante en cierto sentido dado que la estrategia de región de confianza introduce su propia modificación. La solución exacta del problema de región de confianza (2.4.14) con

$B_k = \nabla^2 f(x_k)$ es $(\nabla^2 f(x_k) + \lambda I)^{-1} g_k$, donde λ se elige lo suficientemente grande para hacer a $(\nabla^2 f(x_k) + \lambda I)$ positiva definida, y sus valores dependen del radio de la región de confianza Δ_k. Se conluye que el método de Newton-pata de perro es más apropiado cuando la función objetivo es convexa (o sea, $\nabla^2 f(x_k)$ es siempre positiva semidefinida).

El método de pata de perro se puede adaptar para matrices indefinidas B, pero no tiene mucho caso hacerlo ya que el paso completo p^B no es el minimizador sin restricciones de m en este caso.

Ahora que se han estudiado brevemente estos métodos iterativos para minimizar una función objetivo, se estudiarán los métodos utilizados en este trabajo para la resolución del problema (2.3.5), que para los fines de este proyecto, resulta ser un problema no lineal.

2.5 Mínimos cuadrados no lineales

En un problema de mínimos cuadrados, la función objetivo f tiene la forma particular

$$f(x) = \frac{1}{2} \sum_{j=1}^{m} r_j^2(x), \tag{2.5.1}$$

donde cada r_j es una función suave de $\mathbf{R}^n$ a $\mathbf{R}$. A las funciones r_j se les suele denominar *residuales*, y para los fines de esta sección, se asume $m \geqslant n$.

La forma particular de la función objetivo f facilita en general encontrar un minimizador en comparación con los problemas generales de minimización no restringida. El primer paso es ensamblar los componentes individuales r_j de (2.5.1) en un vector residual $r : \mathbf{R}^n \to \mathbf{R}^m$ de la siguiente manera

$$r(x) = (r_1(x), r_2(x), \ldots, r_m(x))^T. \tag{2.5.2}$$

Usando esta notación, se puede escribir a f como $f(x) = \frac{1}{2}||r(x)||_2^2$. Las derivadas de $f(x)$ se pueden escribir en términos del Jacobiano $J(x)$

$$J(x) = \left[\frac{\partial r_j}{\partial x_i}\right]_{\substack{j=1,2,\ldots,m \\ i=1,2,\ldots,n}} = \begin{bmatrix} \nabla r_1(x)^T \\ \nabla r_2(x)^T \\ \vdots \\ \nabla r_m(x)^T \end{bmatrix}, \tag{2.5.3}$$

El gradiente y el Hessiano de f se pueden escribir como

$$\nabla f(x) = \sum_{j=1}^{m} r_j(x)\nabla r_j(x) = J(x)^T r(x), \tag{2.5.4}$$

$$\begin{aligned}\nabla^2 f(x) &= \sum_{j=1}^{m} \nabla r_j(x) r_j(x)^T + \sum_{j=1}^{m} r_j(x)\nabla^2 r_j(x) \\ &= J(x)^T J(x) + \sum_{j=1}^{m} r_j(x)\nabla^2 r_j(x).\end{aligned} \tag{2.5.5}$$

2.5.1 El método de Gauss-Newton

Este método se puede ver como una variante del método de Newton combinado con una búsqueda en línea. En lugar de resolver las ecuaciones de Newton estándar $\nabla^2 f(x_k)p = -\nabla f(x_k)$, se resuelve el siguiente sistema para encontrar la dirección de búsqueda p_k^{GN}:

$$J_k^T J_k p_k^{GN} = -J_k^T r_k. \tag{2.5.6}$$

Esta modificación simple proporciona un número de ventajas sobre el método de Newton. En primer lugar, la aproximación

$$\nabla^2 f_k \approx J_k^T J_k, \tag{2.5.7}$$

evita tener que calcular los Hessianos individuales de los residuales $\nabla^2 r_j, j = 1, 2, \ldots, m$, que se necesitan en el segundo término de (2.5.5). De hecho, si se calculara el Jacobiano J_k en el proceso de evaluar el gradiente $\nabla f_k = J_k^T r_k$, la aproximación (2.5.7) no requiere evaluaciones adicionales, y el ahorro en tiempo de cómputo puede ser muy significativo en algunas aplicaciones. Segundo, hay varias situaciones muy interesantes en las que el primer término $J^T J$ en (2.5.5) domina al segundo término (al menos cerca de la solución x^*), así que $J_k^T J_k$ es una aproximación cercana a $\nabla^2 f_k$ y la tasa de convergencia de Gauss-Newton es similar a la del método de Newton. El primer término de (2.5.5) dominará cuando la norma de cada término de segundo órden (o sea, $|r_j(x)|\,||\nabla^2 r_j(x)||$) sea significativamente más pequeña que los valores propios de $J^T J$. Este comportamiento es común cuando, o bien los residuales r_j son pequeños, o bien cuando son casi afines (y por tanto las $||\nabla^2 r_j||$ son pequeñas). En la práctica, varios problemas de mínimos cuadrados tienen residuales pequeños en la solución, lo que lleva a una rápida convergencia local del método de Gauss-Newton.

Una tercera ventaja del método de Gauss-Newton es que siempre que J_k tiene rango completo y el gradiente ∇f_k no es cero, la dirección p_k^{GN} es una dirección descendente para f, y por tanto es una dirección adecuada para la búsqueda en línea. De (2.5.4) y (2.5.6) se tiene

$$(p_k^{GN})T\nabla f_k = (p_k^{GN})^T J_k^T r_k = -(p_k^{GN})^T J_k^T J_k p_k^{GN} = -||J_k p_k^{GN}||^2 \leqslant 0. \tag{2.5.8}$$

La última desigualdad es estrícta a menos que $J_k p_k^{GN} = 0$, en cuyo caso se tiene por (2.5.6) y el rango completo de J_k que $J_k^T r_k = \nabla f_k = 0$; o sea, x_k es un punto

estacionario. Finalmente, la cuarta ventaja del método de Gauss-Newton surge de la semejanza entre las ecuaciones (2.5.6) y las ecuaciones normales para el problema de mínimos cuadrados lineales. Esta conexión dice que p_k^{GN} es de hecho la solución al problema de mínimos cuadrados lineales

$$\min_p \frac{1}{2}||J_k p + r_k||^2. \tag{2.5.9}$$

Se puede encontrar la dirección de búsqueda aplicando los algoritmos de mínimos cuadrados lineales al problema (2.5.9). De hecho, si se usan algoritmos de factorización QR o de descomposición en valores singulares, no hay necesidad de calcular la aproximación al Hessiano $J_k^T J_k$ en (2.5.6) explícitamente; se puede trabajar directamente con el Jacobiano J_k. Análogamente si se usa una técnica de gradiente conjugado para resolver (2.5.9).

Si el número m de residuales es grande y el número n de variables es relativamente pequeño, podría no ser conveniente almacenar el Jacobiano J explícitamente. Una estrategia alterna podría ser calcular la matriz $J^T J$ y el vector gradiente $J^T r$ al evaluar r_j y ∇r_j sucesivamente para $j = 1, 2, \ldots, m$ y hacer

$$J^T J = \sum_{j=1}^{m} (\nabla r_j)(\nabla r_j)^T, \quad J^T r = \sum_{j=1}^{m} r_j (\nabla r_j). \tag{2.5.10}$$

Los pasos para Gauss-Newton se pueden calcular resolviendo el sistema (2.5.6) de ecuaciones normales directamente.

El problema (2.5.9) sugiere otra motivación para la dirección de búsqueda de Gauss-Newton. Se puede ver esta ecuación como obtenida de un modelo en línea para la función vectorial $r(x_k + p) \approx r_k + J_k p$, sustituida en la función $\frac{1}{2}||\cdot||^2$. En otras palabras, se usa la aproximación

$$f(x_k + p) = \frac{1}{2}||r(x_k + p)||^2 \approx \frac{1}{2}||J_k p + r_k||^2,$$

y se elige p_k^{GN} como el minimizador de esta aproximación.

Las implementaciones del método de Gauss-Newton usualmente realizan una búsqueda en línea en la dirección p_k^{GN}, usando un tamaño de paso α_k para satisfacer ciertas condiciones adicionales. Para más información sobre la convergencia del método de Gauss-Newton, consulte [69].

2.5.2 El método de Levenberg-Marquardt

Este método utiliza la aproximación (2.5.7) al Hessiano, pero reemplaza la búsqueda en línea por una estrategia de región de confianza, lo cual evita una de la debilidades de Gauss-Newton, a saber, su comportamiento cuando el Jacobiano $J(x)$ es de rango bajo. Como en ambos casos se usa las mismas aproximaciones al Hessiano, las propiedades

de convergencia local de ambos métodos son similares.

En el método de Levenberg-Marquardt, el problema a resolver en cada iteración es

$$\min_p \frac{1}{2}||J_k p + r_k||^2, \quad \text{sujeto a } ||p|| \leqslant \Delta_k, \tag{2.5.11}$$

donde $\Delta_k > 0$ es el radio de la región de confianza. Se elige la función modelo $m_k(\cdot)$ de (2.4.14) como

$$m_k(p) = \frac{1}{2}||r_k||^2 + p^T J_k^T r_k + \frac{1}{2} p^T J_k^T J_k p. \tag{2.5.12}$$

Para el resto de esta sección se omite el subíndice k para enfocarnos en el problema (2.5.11), cuya solución se puede caracterizar de la siguiente manera: cuando la solución p^{GN} de las ecuaciones de Gauss-Newton (2.5.6) yacen estríctamente adentro de la región de confianza (o sea, $||p^{GN}|| < \Delta$), entonces el paso p^{GN} también resuelve el problema (2.5.11). De otra manera, existe una $\lambda > 0$ tal que la solución $p = p^{LM}$ de (2.5.11) satisface $||p|| = \Delta$ y

$$(J^T J + \lambda I)p = -J^T r. \tag{2.5.13}$$

Esta afirmación se verifica en el siguiente lema.

Lema 2.5.14. *El vector p^{LM} es una solución al problema de región de confianza*

$$\min_p ||Jp + r||^2, \quad \textit{sujeto a } ||p|| \leqslant \Delta,$$

sí y sólamente si p^{LM} es factible y existe un escalar $\lambda \geqslant 0$ tal que

$$(J^T J + \lambda I)p^{LM} = -J^T r, \tag{2.5.15a}$$

$$\lambda(\Delta - ||p^{LM}||) = 0. \tag{2.5.15b}$$

La demostración del Lema (2.5.14) se puede consultar en [69]. Note que las ecuaciones (2.5.13) son las ecuaciones normales del siguiente problema de mínimos cuadrados lineales:

$$\min_p \frac{1}{2} \left\| \begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix} p + \begin{bmatrix} r \\ 0 \end{bmatrix} \right\|^2. \tag{2.5.16}$$

Similar al caso de Gauss-Newton, la equivalencia entre (2.5.13) y (2.5.16) sugiere una manera de resolver el problema sin calcular el producto de matrices $J^T J$ y su factorización de Cholesky.

Para encontrar un valor de λ que se aproxime a la Δ del lema (2.5.14), se puede usar un algoritmo para calcular raices. Se garantiza la existencia del factor de Cholesky R siempre que la estimación actual de λ sea positiva, dado que la aproximación al Hessiano $B = J^T J$ ya es positiva semidefinida. Debido a la estructura especial de B,

no es necesario calcular la factorización de Cholesky de $B+\lambda I$. En su lugar, se presenta una técnica para encontrar la factorización QR de la matriz coeficiente en (2.5.16):

$$\begin{bmatrix} R_\lambda \\ 0 \end{bmatrix} = Q_\lambda^T \begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix} \tag{2.5.17}$$

donde Q_λ es ortogonal y R_λ es triangular superior. El factor R_λ satisface $R_\lambda^T R_\lambda = (J^T J + \lambda I)$.

Se puede ahorrar tiempo de cómputo en el cálculo de la factorización (2.5.17) usando una combinación de transformaciones de Householder y Givens. Suponga que se usan transformaciones de Householder para calcular la factorización QR de J como

$$J = Q \begin{bmatrix} R \\ 0 \end{bmatrix}. \tag{2.5.18}$$

Entonces se tiene

$$\begin{bmatrix} R \\ 0 \\ \sqrt{\lambda} I \end{bmatrix} = \begin{bmatrix} Q^T & \\ & I \end{bmatrix} \begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix}. \tag{2.5.19}$$

La matriz de la izquierda en esta fórmula es triangular superior excepto por los n términos distintos de cero de la matriz λI. Estos se pueden eliminar con una sucesión de $n(n+1)/2$ rotaciones de Givens, en las que los elementos de la diagonal de la parte triangular superior se usan para eliminar los términos no nulos de λI y rellenar los términos que surgen en el proceso. Los primeros pasos de este proceso son los siguientes:

> intercambiar el renglón n de R con el renglón n de $\sqrt{\lambda} I$ para eliminar la entrada (n, n) de $\sqrt{\lambda} I$;
>
> intercambiar el renglón $n-1$ de R con el renglón $n-1$ de $\sqrt{\lambda} I$ para eliminar la entrada $(n-1, n-1)$ de esta matriz. Este paso rellena la posición $(n-1, n)$ de $\sqrt{\lambda} I$, la cual se elimina al intercambiar el renglón n de R con el renglón $n-1$ de $\sqrt{\lambda} I$, para eliminar la entrada $(n-1, n)$;
>
> intercambiar el renglón $n-2$ de R con el renglón $n-2$ de $\sqrt{\lambda} I$, para eliminar la diagonal $n-2$ de esta matriz. Este paso rellena las entradas $(n-2, n-1)$ y $(n-2, n)$, que se eliminan al...

y así sucesivamente. Si se combinan todas las rotaciones de Givens en una matriz $\overline{Q}_\lambda$, se obtiene de (2.5.19) que

$$\overline{Q}_\lambda^T \begin{bmatrix} R \\ 0 \\ \sqrt{\lambda} I \end{bmatrix} = \begin{bmatrix} R_\lambda \\ 0 \\ 0 \end{bmatrix}.$$

y por tanto se cumple (2.5.17) con

$$Q_\lambda = \begin{bmatrix} Q & \\ & I \end{bmatrix} \overline{Q}_\lambda.$$

La ventaja de este enfoque combinado es que cuando el valor de λ se cambia en el algoritmo de búsqueda de raices, solo se necesita recalcular $\overline{Q}_\lambda$ y no la parte de Householder de la factorización (2.5.19). Esta característica puede ahorrar muchos cálculos si $m \gg n$, dado que solo se necesitan $O(n^3)$ operaciones para recalcular $\overline{Q}_\lambda$ y R_λ para cada valor de λ, después del costo inicial de $O(mn^2)$ operaciones requeridas para calcular Q en (2.5.18).

Los problemas de mínimos cuadrados suelen estar mal escalados. Algunas de las variables podrían tener valores del orden de 10^4, mientras que otras podrían tener valores del orden de 10^{-6}. Si se ignoran tales variaciones, los algoritmos antes descritos podrían encontrar dificultades numéricas o producir soluciones de mala calidad. Una manera de reducir los efectos del mal escalamiento es usar una región de confianza elipsoidal en lugar de la región de confianza esférica definida anteriormente. El paso está restringido a una elipse en la que las longitudes del eje mayor están relacionadas a los valores típicos de las variables correspondientes. De manera analítica, el problema se convierte en

$$\min_p \frac{1}{2}||J_k p + r_k||^2, \quad \text{sujeto a } ||D_k p|| \leqslant \Delta_k, \tag{2.5.20}$$

donde D_k es una matriz diagonal con entradas positivas. En lugar de (2.5.13), la solución de (2.5.20) satisface una ecuación de la forma

$$(J_k^T J_k + \lambda D_k^2) p_k^{LM} = -J_k^T r_k, \tag{2.5.21}$$

y, de manera equivalente, resuelve el problema de mínimos cuadrados lineales

$$\min_p \left\| \begin{bmatrix} J_k \\ \sqrt{\lambda} D_k \end{bmatrix} p + \begin{bmatrix} r_k \\ 0 \end{bmatrix} \right\|^2. \tag{2.5.22}$$

Las diagonales de la matriz de ajuste D_k pueden cambiar de iteración a iteración mientras se reune información sobre el rango de valores de cada componente de x. Si la variación en estos elementos se mantiene dentro de ciertos límites, entonces la convergencia para el caso esférico se mantiene, con modificaciones menores. Más aún, la técnica aquí descrita para calcular R_λ no necesita modificaciones. Seber y Wild [a reference, maybe?] sugieren elegir las diagonales de D_k^2 de tal manera que coincidan con las de $J_k^T J_k$, para hacer al algoritmo invariante bajo reescalamiento diagonal en los componentes de x.

Para problemas en los que m y n son grandes y $J(x)$ es rala, se sugiere resolver (2.5.11) o (2.5.20) de manera aproximada usando el algoritmo CG-Steihaug, con $J_k^T J_k$ reemplazando al Hessiano exacto $\nabla^2 f_k$. La propiedad de la matriz $J_k^T J_k$ de ser positiva semidefinida simplifica este algoritmo, debido a que no puede haber curvatura negativa. No es necesario calcular $J_k^T J_k$ explícitamente para implementar el algoritmo

CG-Steihaug; los productos requeridos en el algoritmo se pueden calcular realizando los productos con J_k y J_k^T por separado.

Para más información sobre la convergencia del método de Levenberg-Marquardt, así como del algoritmo CG-Steihaug, recomiendo referirse a [69].

Hasta ahora se han descrito brevemente los métodos itreativos utilizados en este proyecto para resolver el problema (2.3.5). En la siguiente sección se discuten los métodos de penalización utilizados para complementar este proyecto, los cuales están dirigidos al parámetro α y al funcional $J(q)$ de (2.3.5). En estos métodos, la idea es remplazar el problema original por una sucesión de subproblemas en la que las restricciones del problema se representan por términos añadidos a la función objetivo, esto con la finalidad de imponer estabilidad, añadir información al problema, o en el caso particular de este proyecto, reflejar la física del problema en cuestión.

2.6 Métodos de penalización

El problema de minimizar una función sujeta a ciertas restricciones en sus variables se puede formular como

$$\min_{x\in\mathbf{R}^n} f(x) \quad \text{sujeto a } = \begin{cases} c_i(x) = 0, & i \in \mathcal{E}, \\ c_i(x) \geqslant 0, & i \in \mathcal{I}, \end{cases} \tag{2.6.1}$$

donde f y las funciones c_i son funciones diferenciables y real valuadas, definidas en un dominio de $\mathbf{R}^n$, y $\mathcal{E}$ e $\mathcal{I}$ son dos conjuntos finitos de índices. f se denomina la función objetivo, las c_i's con $i \in \mathcal{E}$ son las restricciones de identidad y las c_i's con $i \in \mathcal{I}$ son las restricciones de desigualdad. Se define el conjunto factible Ω como el cunjunto de puntos que satisfacen las restricciones, es decir,

$$\Omega = \{x : c_i(x) = 0, \quad i \in \mathcal{E}; \quad c_i(x) \geqslant 0, \quad i \in \mathcal{I}\}, \tag{2.6.2}$$

por lo que se puede escribir (2.6.1) como

$$\min_{x\in\Omega} f(x). \tag{2.6.3}$$

En esta sección se estudian las caracterizaciones de las soluciones de (2.6.3). Primero se definirán algunos conceptos y se listarán algunos resultados útiles.

Definición 2.6.4. El conjunto activo $\mathcal{A}(x)$ de un punto factible x consiste del subconjunto de índices de $\mathcal{E}$ correspondientes a la restricción de igualdad, junto con los índices de desigualdad i para los cuales $c_i(x) = 0$, es decir

$$\mathcal{A}(x) = \mathcal{E} \cup \{i \in \mathcal{I} : C_i(x) = 0\}.$$

En un punto factible x, la condición de desigualdad $i \in \mathcal{I}$ se llama activa si $c_i(x) = 0$ e inactiva si se cumple la desigualdad estrícta $c_i(x) > 0$.

Se introduce la *función Lagrangiana*

$$\mathcal{L}(x, \lambda_1) = f(x) - \lambda_1 c_1(x), \tag{2.6.5}$$

donde al escalar λ_1 se le denomina *multiplicador de Lagrange* para la restricción $c_1(x) = 0$. Note que $\nabla_x \mathcal{L}(x, \lambda_1) = \nabla f(x) - \lambda_1 \nabla c_1(x)$.

Definición 2.6.6. (LICQ) Dado un punto $x \in \mathcal{A}(x)$, se dice que se cumple la condición de restricciones linealmente independientes (LICQ por sus siglas en inglés) si el conjunto de gradientes de las restricciones activas $\{\nabla c_i(x) : i \in \mathcal{A}(x)\}$ es linealmente independiente.

Teorema 2.6.7. *(Condiciones necesarias de primer orden) Suponga que x^* es una solución local de (2.6.1), y que las funciones f y c_i de (2.6.1) son continuamente diferenciables, y que x^* cumple la condición de restricciones linealmente independientes. Entonces existe un multiplicador de Lagrange vectorial λ^*, con componentes λ_i^*, $i \in \mathcal{E} \cup \mathcal{I}$ de tal suerte que las siguientes condiciones se cumplen en (x^*, λ^*)*

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = 0, \tag{2.6.8a}$$
$$c_i(x^*) = 0, \quad \textit{para todo } i \in \mathcal{E}, \tag{2.6.8b}$$
$$c_i(x^*) \geqslant 0, \quad \textit{para todo } i \in \mathcal{I}, \tag{2.6.8c}$$
$$\lambda_i^* \geqslant 0, \quad \textit{para todo } i \in \mathcal{I}, \tag{2.6.8d}$$
$$\lambda_i^* c_i(x^*) = 0, \quad \textit{para todo } i \in \mathcal{E} \cup \mathcal{I}. \tag{2.6.8e}$$

Las condiciones (2.6.8) se conocen como *condiciones de Karush-Kuhn-Tucker*, o *condiciones KKT*. Las condiciones (2.6.8e) son condiciones complementarias, éstas implican que, o bien ninguna restricción i es activa, o bien que todos los $\lambda_i^* = 0$, o quizás ambas. En particular, los multiplicadores de Lagrange correspondientes a las restricciones de desigualdad inactivas son nulos, se pueden omitir los términos con índices $i \notin \mathcal{A}(x^*)$ de (2.6.8a) y rescribir esta condición como

$$0 = \nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) - \sum_{i \in \mathcal{A}(x^*)} \lambda_i^* \nabla c_i(x^*). \tag{2.6.9}$$

Definición 2.6.10. Dado un punto factible x y el conjunto activo $\mathcal{A}(x)$, se define el *conjunto de direcciones factibles linealizadas* $\mathcal{F}(x)$ como

$$\mathcal{F}(x) = \left\{ d : \begin{array}{ll} d^T \nabla c_i(x) = 0, & \text{para todo } i \in \mathcal{E}, \\ d^T \nabla c_i(x) \geqslant 0, & \text{para todo } i \in \mathcal{A}(x) \cap \mathcal{I}, \end{array} \right\}.$$

Dado $\mathcal{F}(x^*)$ como de la definición anterior y algún multiplicador vectorial λ^* con las condiciones KKT, se define el *cono crítico* $\mathcal{C}(x^*, \lambda^*)$ como sigue:

$$\mathcal{C}(x^*, \lambda^*) = \{w \in \mathcal{F}(x^*) : \nabla c_i(x^*)^T w = 0, \text{ para toda } i \in \mathcal{A}(x^*) \cap I \text{ con } \lambda_i^* > 0\}.$$

De manera equivalente,

$$w \in \mathcal{C}(x^*, \lambda^*) \Leftrightarrow \begin{cases} \nabla c_i(x^*)^T w = 0, & \text{para todo } i \in \mathcal{E}, \\ \nabla c_i(x^*)^T w = 0, & \text{para todo } i \in \mathcal{A}(x^*) \cap \mathcal{I} \text{ con } \lambda_i^* > 0, \\ \nabla c_i(x^*)^T w \geqslant 0, & \text{para todo } i \in \mathcal{A}(x^*) \cap \mathcal{I} \text{ con } \lambda_i^* = 0. \end{cases} \tag{2.6.11}$$

Para el siguiente resultado se involucran segundas derivadas, por lo que es necesario asumir propiedades más fuertes de suavidad en las funciones. Para este propósito, se asume que f y c_i, $i \in \mathcal{E} \cup \mathcal{I}$ son de clase C^2.

Teorema 2.6.12. *(Condiciones necesarias de segundo orden) Suponga que x^* es una solución local de (2.6.1) que cumple la LICQ. Sea λ^* un multiplicador de Lagrange vectorial con las condiciones KKT. Entonces*

$$w^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) w \geqslant 0, \quad \textit{para todo } w \in \mathcal{C}(x^*, \lambda^*). \tag{2.6.13}$$

Teorema 2.6.14. *(Condiciones suficientes de segundo orden) Suponga que para algún punto factible $x^* \in \mathbf{R}^n$ existe un multiplicador de Lagrange vetorial λ^* de tal suerte que sumple con las condiciones KKT. Suponga también que*

$$w^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) w > 0 \quad \textit{para toda } w \in \mathcal{C}(x^*, \lambda^*), w \neq 0. \tag{2.6.15}$$

Entonces x^* es una solución local estrícta para (2.6.1).

Las demostraciones de todos los resultados listados en esta sección se pueden consultar en [69].

En esta sección se estudian dos métodos para complementar la solución al problema (2.3.5). En primer lugar, el *método de penalización cuadrática*, el cual agrega un múltiplo del cuadrado de la violación a cada restricción de la función objetivo. Este es un método común en la práctica, dada su simplicidad, aunque tiene sus propias desventajas. En segundo lugar, se estudia el *método de penalización exacta no suave*, en el cual un problema sin restricciones (en lugar de una sucesión) reemplaza al problema restringido original. Usando estas funciones de penalización, se podría encontrar una solución usando un método de minimización como los que ya se han mencionado, pero la ausencia de suavidad de la restricción puede complicar las cuentas.

2.6.1 Método de penalización cuadrática

La idea es remplazar un problema de optimización con restricciones con una sola función que consiste de

- la función objetivo original del problema de optimización con restricciones, *además de*
- un término adicional por cada restricción, el cual es positivo cuando el punto actual x viola la restricción, y cero en otro caso.

Muchos enfoques definen una sucesión de estas funciones de penalización, en la cual los términos de penalización para las restricciones violadas se multiplican por un coeficiente positivo. Al hacer más grande a este corficiente, la penalización es más severa, obligando así al minimizador de la función de penalización a acercarse a la región factible del problema con restricciones.

La más sencilla de las funciones de penalización de este tipo es la función de penalización cuadrática, en la que los términos de penalización son los cuadrados de las restricciones violadas. Este enfoque se describe inicialmente en el contexto del problema equitativamente restringido

$$\min_x f(x) \quad \text{sujeto a } c_i(x) = 0, \quad i \in \mathcal{E}. \tag{2.6.16}$$

En esta formulación, la función de penalización cuadrática $Q(x;\mu)$ es

$$Q(x;\mu) := f(x) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} c_i^2(x), \tag{2.6.17}$$

donde $\mu > 0$ es el *parámetro de penalización*. Haciendo crecer a μ de manera arbitraria, se penaliza a las violaciones a las restricciones aumentanto la función considerablemente. Se considera entonces una sucesión $\{\mu_k\}$ con $\mu_k \uparrow \infty$ cuando $k \to \infty$, y se busca aproximar al minimizador x_k de $Q(x;\mu)$ para cada índice k. Dado que los términos en (2.6.17) son suaves, se pueden usar métodos de optimización no restringida para buscar las x_k, para lo cual se pueden emplear los minimizadores x_{k-1}, x_{k-2}, etc., de $Q(\cdot;\mu)$ para valores más pequeños de μ para una iteración inicial. Puede que solo se necesiten unos pocos pasos de minimización no restringida para cada μ_k, si se elige bien la sucesión $\{\mu_k\}$ junto con una buena elección inicial.

Para el problema de optimización general con restricciones

$$\min_x f(x) \quad \text{sujeto a } c_i(x) = 0, \quad i \in \mathcal{E}, \quad c_i(x) \geqslant 0, \quad i \in \mathcal{I} \tag{2.6.18}$$

que contiene restricciones en identidades y desigualdades, se puede definir la función cuadrática de penalización como

$$Q(x;\mu) := f(x) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} c_i^2(x) + \frac{\mu}{2}\sum_{i\in\mathcal{I}} \left([c_i(x)]^-\right)^2, \tag{2.6.19}$$

donde $[y]^- = \max\{-y, 0\}$. En este caso, Q podría resultar menos suave que la función objetivo y que las funciones de restricción. Por ejemplo, si una de las restricciones es $x_1 \geqslant 0$, entonces la función $\min(0, x_1)^2$ tiene una segunda derivada discontinua, por lo que Q no sería de clase C^2.

En la implementación se puede elegir una suceción de parámetros $\{\mu_k\}$ de manera adaptativa basada en la dificultad de minimizar la función de penalización en cada iteración. Cuando el proceso de minimizar $Q(x;\mu_k)$ resulta demasiado costoso para alguna k, se elige μ_{k+1} para que sea modestamente mayor que μ_k; por ejemplo $\mu_{k+1} = 1.5\mu_k$. Si el minimizador de $Q(x;\mu_k)$ se puede aproximar de manera barata, se puede intentar

un incremento más ambicioso, por ejemplo $\mu_{k+1} = 10\mu_k$. La teoría detrás del algoritmo permite latitudes amplias en la elección de tolerancias no negativas τ_k; solo se requiere que $\tau_k \to 0$ para asegurar que la minimización se realiza de manera precisa mientras las iteraciones avanzan.

No hay garantía de que el paro de prueba $||\nabla_x Q(x;\mu_k)|| \leqslant \tau_k$ se cumpla ya que, como se mencionó arriba, las iteraciones podrían alejarse de la región factible cuando el parámetro de penalización no es lo suficientemente grande. Una implementación práctica debe incluir garantías de que dicho parámetro incrementará cuando las violaciones a la restricción no decrezcan lo suficientemente rápido, o cuando las iteraciones parezcan diverger.

Cuando solo hay restricciones de igualdad, $Q(x;\mu_k)$ es suave, por lo que los algoritmos para minimización sin restricciones se pueden usar para identificar a la solución aproximada x_k. Sin embargo, la minimización de $Q(x;\mu_k)$ se vuelve más complicada si μ_k crece, a menos de que se usen técnicas especiales para calcular las direcciones de búsqueda. Por un lado, el Hessiano $\nabla^2_{xx} Q(x;\mu_k)$ podría volverse arbitrariamente mal condicionado cerca del minimizador. Esta propiedad por sí misma es suficiente para hacer que varios algoritmos para minimización sin restricciones, tales como los cuasi-Newton o gradiente conjugado, tengan un desempeño muy pobre. El método de Newton, por otro lado, no es sensible al mal condicionamiento del Hessiano, pero aún así podría encontrar dificultades para μ_k grandes por otras dos razones. Primera, el mal condicionamiento de $\nabla^2_{xx} Q(x;\mu_k)$ podría causar problemas numéricos cuando se resuelven las ecuaciones lineales para calcular el paso de Newton. Sin embargo, estos efectos no son graves y las ecuaciones de Newton se pueden reformular. Segunda, aún cuando x está cerca del minimizador de $Q(x;\mu_k)$, el desarrollo de Taylor cuadrático de $Q(x;\mu_k)$ alrededor de x es una aproximación razonable a la función verdadera sólo en una pequeña vecindad de x. Como el método de Newton se basa en el modelo cuadrático, los pasos que éste genera podrían no avanzar de manera rápida hacia el minimizador de $Q(x;\mu_k)$. Esta dificultad se puede aligerar con una buena elección del punto inicial x^s_{k+1}, o haciendo $x^s_{k+1} = x_k$ y eligiendo μ_{k+1} modestamente mayor a μ_k. Para más información sobre la convergencia del método de penalización cuadrática, consulte [69].

2.6.2 Funciones de penalización no diferenciables

Algunas funciones de penalización son exactas, lo que significa que, para ciertas elecciones de sus parámetros de penalización, una sola minimización con respecto de x puede llevar a la solución exacta del problema de programación no lineal. Esta es una propiedad deseable ya que hace que la ejecución de los métodos de penalización sean menos dependientes de la estrategia para actualizar el parámetro de penalización. La función de penalización cuadrática de la sección anterior no es exacta ya que su minimizador en general no es el mismo de la solución del programa no lineal para cualquier valor positivo de μ. En esta sección se discuten funciones de penalización no diferenciables, las cuales han resultado útiles en varios contextos.

Una función popular de penalización no diferenciable para el problema general de programación no lineal (2.6.18) es la *función de penalización* l_1 definida como

$$\phi_1(x;\mu) = f(x) + \mu \sum_{i\in\mathcal{E}} |c_i(x)| + \mu \sum_{i\in\mathcal{I}} [c_i(x)]^-, \tag{2.6.20}$$

donde $[y]^- = \max\{0, -y\}$. Su nombre se deriva del hecho que el término de penalización es μ por la norma l_1 de la restricción violada. Note que $\phi_1(x;\mu)$ no es diferenciable en algunos puntos, debido a la presencia del valor absoluto y la función $[\cdot]^-$.

El siguiente resultado establece la exactitud de la función de penalización l_1.

Teorema 2.6.21. *Suponga que x^* es una solución local estrícta del problema de programación no lineal (2.6.18) en el que las condiciones necesarias de primer orden del Teorema (2.6.7) se cumplen, con multiplicadores de Lagrange λ_i^*, $i \in \mathcal{E}\cup\mathcal{I}$. Entonces x^* es un minimizador local de $\phi_1(x;\mu)$ para todo $\mu > \mu^*$, donde*

$$\mu^* = ||\lambda^*||_\infty = \max_{i\in\mathcal{E}\cup\mathcal{I}} |\lambda_i^*|. \tag{2.6.22}$$

Más aún, si se cumplen las condiciones suficientes de segundo orden del Teorema (2.6.14) y $\mu > \mu^*$, entonces x^* es un minimizador local estrícto de $\phi_1(x;\mu)$.

Dicho de otra manera, en la solución x^* del programa no lineal, cualquier movimiento dentro de la región no factible se penaliza lo suficientemente brusco para que produzca un incremento en la función de penalización a un valor mayor que $\phi_1(x^*;\mu) = f(x^*)$, y por tanto forzando al minimizador de $\phi_1(\cdot;\mu)$ a regresar a x^*.

Como los métodos de penalización trabajan minimizando directamente la función de penalización, se requiere caracterizar a los puntos estacionarios de ϕ_1. Aún cuando ϕ_1 no sea diferenciable, tiene una derivada direccional $D(\phi_1(x;\mu);p)$ a lo largo de cualquier dirección.

Definición 2.6.23. Un punto $\hat{x} \in \mathbf{R}^n$ es un punto estacionario de la función de penalización $\phi_1(x;\mu)$ si

$$D(\phi_1(\hat{x};\mu);p) \geqslant 0, \tag{2.6.24}$$

para todo $p \in \mathbf{R}^n$. Similarmente, $\hat{x}$ es un punto estacionario de la medida de inviabilidad

$$h(x) = \sum_{i\in\mathcal{E}} |c_i(x)| + \sum_{i\in\mathcal{I}} [c_i(x)]^- \tag{2.6.25}$$

si $D(h(\hat{x});p) \geqslant 0$ para todo $p \in \mathbf{R}^n$. Si un punto no es factible para (2.6.18) pero es estacionario con respecto a la medida de inviabilidad h, se dice que es un punto estacionario no factible.

El siguiente resultado complementa al Teorema (2.6.21) mostrando que los puntos estacionarios de $\phi_1(x;\mu)$ corresponden con los puntos con las condiciones KKT del problema de optimización con restricciones (2.6.18) bajo ciertas condiciones.

Teorema 2.6.26. *Suponga que $\hat{x}$ es un punto estacionario de la función de penalización $\phi_1(x;\mu)$ para toda μ mayor que un cierto límite $\hat{\mu} > 0$. Entonces si $\hat{x}$ es factible para el programa no lineal (2.6.18), satisface las condiciones KKT. Si $\hat{x}$ no es factible para (2.6.18), entonces es un punto estacionario no factible.*

Estos resultados brindan la motivación para bosquejar un algoritmo basado en la función de penalización l_1. La minimización de $\phi_1(x;\mu_k)$ se complica debido a la no diferenciabilidad de la función. Sin embargo, se sabe bien cómo calcular los pasos de minimización usando un modelo diferenciable de $\phi_1(x;\mu_k)$.

El esquema más sencillo para actualizar el parámetro de penalización μ_k es incrementarlo en un múltiplo constante, si el valor actual produce un minimizador que no es factible dentro de una tolerancia τ. Este esquema suele trabajar bien en la práctica, pero también puede ser ineficiente. Si el parámetro inicial de penalización μ_0 es demasiado pequeño, se requerirán varias iteraciones para determinar un valor apropiado. Más aún, las iteraciones podrían alejarse de la solución x^* en estos ciclos iniciales, en cuyo caso la minimización de $\phi_1(x;\mu_k)$ debe terminarse antes y quizás x_k^s debería retornarse a la iteración anterior. Si, por otro lado, μ_k es excesivamente grande, entonces la función de penalización será difícil de minimizar, y posiblemente requiera un número grande de iteraciones.

Capítulo 3

Esquemas en diferencias finitas

3.1 Diferencias Finitas Clásicas

Suponga que se tiene una ecuación diferencial (o un sistema de ecuaciones) definida sobre un dominio D junto con sus condiciones a la frontera. Este problema se puede escribir de forma simbólica como

$$Lu = f, \tag{3.1.1}$$

donde L es un operador diferencial, y f es el lado derecho de la ecuación. Por ejemplo, para escribir el problema

$$\begin{aligned} \frac{du}{dx} + \frac{x}{1+u^2} &= \cos x, \qquad 0 \leqslant x \leqslant 1, \\ u(0) &= 3, \end{aligned} \tag{3.1.2}$$

de la forma (3.1.1), se tiene

$$Lu \equiv \frac{du}{dx} + \frac{x}{1+u^2}, \qquad f \equiv \cos x,$$

donde $u(x)$ está definida en $[0,1]$, junto con la condición de frontera $u(0) = 3$.

Se asume que la solución $u(x)$ del problema (3.1.1) existe. Para calcular esta solución usando el método de diferencias finitas, en primer lugar se debe elegir un conjunto finito de puntos en el dominio D. A dicho subconjunto finito del dominio se le denomina el *mallado* D_h. La idea del método de diferencias finitas no es de calcular la solución exacta $u(x)$ de (3.1.1), sino una tabla de valores $[u]_h$ de la solución en los puntos del mallado D_h. Se asume que el mallado D_h depende de un parámetro $h > 0$, el cual se puede tomar tan pequeño como se desee. A este parámetro se le denomina el *tamaño de paso*, y la idea es que mientras más pequeño sea este parámetro, más fina será la malla.

Por ejemplo, se puede tomar $h = 1/N$, donde N es un entero positivo, y considerar el mallado $D_h = \{x_0 = 0, x_1 = h, x_2 = 2h, \ldots, x_N = 1\}$ sobre el dominio $D = [0,1]$.

Usando este mallado, la función discretizada $[u]_h$ sobre los puntos del mallado $D_h = \{x_n = nh : 1 \leqslant n \leqslant N, n \in \mathbb{N}\}$ toma los valores $u(nh)$, el cual se abrevia como u_n para los fines de la explicación. Para el cálculo de los valores de la tabla $[u]_h$, se puede usar la aproximación siguiente para el problema (3.1.2):

$$\left.\begin{aligned} \frac{u_{n+1}-u_n}{h} + \frac{x_n}{1+u_n^2} &= \cos x_n, \qquad n = 0, 1, \ldots, N-1, \\ u_0 &= 3. \end{aligned}\right\} \tag{3.1.3}$$

Nótese que la derivada del problema (3.1.2) se ha sustituido por la aproximación

$$\frac{du}{dx} \approx \frac{u(x+h)-u(x)}{h}.$$

El conjunto $u^{(h)} = \{u_0^{(h)}, u_1^{(h)}, \ldots, u_N^{(h)}\}$ se conforma de los valores aproximados de la solución $u(x)$ restringida a los puntos del mallado D_h, es decir, este conjunto se puede interpretar como una aproximación a $u(x)$ restringida a los puntos del mallado D_h. En el contexto del problema (3.1.1), a esta proyección de la aproximación de la solución $u(x)$ sobre los puntos del mallado D_h se le denomina la *forma discretizada del operador* L. Los valores $u_0^{(h)}, u_1^{(h)}, \ldots, u_N^{(h)}$ en los puntos $x_1, x_2, \ldots, x_N$ se calculan de manera consecutiva usando el sistema (3.1.3) para $n = 0, 1, \ldots, N-1$. Por simplicidad, en el sistema (3.1.3) se omitieron los superíndices (h), lo cual se hará para fines de la explicación de aquí en adelante.

En este ejemplo se considera un mallado uniforme, es decir, los puntos se encuentran separados entre sí por una distancia constante h. Esta condición no es un imperativo, ya que es posible considerar una distancia variable entre los nodos del mallado; es decir, se puede considerar un mallado $x_0 = 0$, $x_1 = x_0 + h_0$, $x_2 = x_1 + h_1$, $\ldots$, $x_N = 1$, donde los tamaños de paso h_n para $n = 0, 1, \ldots, N-1$ no son todos iguales, pero $\max h_n \to 0$, así como $h = 1/N \to 0$. Los nodos de D_h se pueden colocar sobre el dominio de manera conveniente para adaptarse al problema, por ejemplo, concentrando una mayor cantidad de puntos en los intervalos donde la solución $u(x)$ presenta mayores variaciones (ver [46] o [48]). Estas adaptaciones se realizan principalmente para acoplar la solución al problema físico que representa el sistema, o de acuerdo a las necesidades de cálculo involucradas. En el cálculo secuencial de las aproximaciones $u_1^{(h)}$, $u_2^{(h)}$, $\ldots$, $u_n^{(h)}$, se obtiene también información sobre la tasa de cambio de $u(x)$, y se sugiere que esta información se tome en cuenta para la elección del siguiente punto en el mallado x_{n+1}.

En estos ejemplos nos hemos limitado a ilustrar el concepto de mallado y la correspondiente discretización de la solución $[u]_h$. Sin embargo, el cálculo de valores que coindida de manera aproximada con la solución en los puntos del mallado de la manera que se ha discutido hasta ahora, es sólo una de las opciones posibles para calcular una solución al problema. Una opción alternativa podría ser el considerar los valores de la función $[u]_h$ en los puntos $x = h/2, 3h/2, \ldots, 1 - h/2$ dados por

$$[u]_h = \frac{1}{h}\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} u(\xi)\, d\xi.$$

Esta alternativa podría resultar conveniente en el caso de que $u(x)$ sea integrable. Por tal motivo, y a menos de que se especifique lo contrario, se asume que las funciones solución involucradas son no sólamente continuas, sino también diferenciables hasta el grado necesario según el problema, lo cual es una propiedad mucho más fuerte en una función que el solo hecho de ser integrable.

Nos interesa entonces el problema del cálculo de la tabla de valores para $[u]_h$, ya que, mientras más fino sea el mallado (es decir, conforme $h \to 0$), esta tabla proporciona una descripción cada vez más detallada de la solución u. Usando métodos para interpolación se puede construir la solución en el dominio D con una precisión creciente a medida que $h \to 0$, aunque dicha precisión depende de datos adicionales acerca de la solución (por ejemplo, cotas en la derivada), y de la distribución de los puntos del mallado.

Ejemplo 3.1.4. Considere el problema de la ecuación de advección en una dimensión espacial, definida para $x \in [-1, 1]$ y $t \geqslant 0$:

$$u_t = u_x, \qquad u = u(x,t), \text{ con la condición inicial } u_0(x) = u(0,x).$$

El mallado sobre el cual se va a calcular la solución numérica de este problema se puede definir de manera regular de la siguiente manera: si $x \in [x_{\min}, x_{\max}]$, se fija $N \in \mathbf{N}$ y con ello $\Delta x = \frac{1}{N}(x_{\max} - x_{\min})$. Dada la resolución Δx, se define $x_i = x_{\min} + i\Delta x$, que marca un conjunto de elementos del dominio uniformemente espaciados.

El dominio temporal se discretiza de manera semejante, se define $t_j = j\Delta t$, donde $j \in \mathbf{N}$ y $\Delta t \leqslant \Delta x$. De esta manera, se define un mallado $M := \{(x_i, t_j) : i, j \in \mathbf{N}, i \leqslant N\}$. Para la función u definida en el mallado, por abreviación se suele escribir $u_i^j := u(x_i, t_j)$. Al conjunto $M_n := \{(x_i, t_n) : i \in \mathbf{N}, i \leqslant N\}$, definido para una $n \in \mathbf{N}$ fija, se le suele llamar el *nivel n del mallado.*

Para la discretización del operador diferencial, se remplazan las derivadas por diferencias finitas, lo cual es una estimación del valor de la función u en los puntos "vecinos" en el mallado, lo cual es posible si u admite una expansión de Taylor y Δx es pequeño. Esto se puede hacer de varias maneras, por ejemplo:

$$\begin{aligned} u_x(x_i, t_j) &= \frac{u_{i+1}^j - u_{i-1}^j}{2\Delta x} + \mathcal{O}(\Delta x^2), \\ u_t(x_i, t_j) &= \frac{u_i^{j+1} - u_i^{j-1}}{2\Delta t} + \mathcal{O}(\Delta t^2). \end{aligned}$$

Estas aproximaciones reemplazan a las derivadas parciales en el problema original. De esta manera, con la evaluación de estas versiones aproximadas del operador diferencial en los puntos del mallado, se tiene la forma discretizada del problema:

$$\frac{u_i^{j+1} - u_i^{j-1}}{\Delta t} = \frac{u_{i+1}^j - u_{i-1}^j}{\Delta x} + \mathcal{O}(\Delta t^2, \Delta x^2).$$

Este sistema se resuelve iterativamente para calcular el valor de u_i^{j+1} para cada $j \in \mathbf{N}$ que forma parte del mallado.

Como se pudo ver en este ejemplo, un esquema para la ecuación diferencial parcial $Lu = f$ se puede escribir en general como $P_{i,j}u = R_{i,j}f$ de una manera natural, donde las expresiones $P_{i,j}u$ y $R_{i,j}f$ corresponden a las respectivas discretizaciones de los operadores, las cuales evaluadas en un punto (x_i, t_j) del mallado, involucran solo una suma finita de términos con $u_{i'}^{j'}$ o $f_{i'}^{j'}$, respectivamente. Con esta idea, se puede enunciar una primera definición para el orden de precisión de un esquema.

Definición 3.1.5. Un esquema $P_{i,j}u = R_{i,j}f$ adaptado a la ecuación diferencial $Lu = f$ tiene precisión de orden p en el tiempo y orden q en el espacio si para cada función (suficientemente) diferenciable $\phi(x,t)$,

$$P_{i,j}\phi - R_{i,j}P\phi = \mathcal{O}(\Delta t^p) + \mathcal{O}(\Delta x^q).$$

Entonces se dice que tal esquema tiene orden de precisión (p, q).

Ejemplo 3.1.6. Un modelo clásico de las ecuaciones hiperbólicas es la ecuación de onda en un sentido:

$$u_t + au_x = 0, \tag{3.1.7}$$

donde a es una constante, t y x son las variables temporal y espacial, respectivamente. Algunos esquemas en diferencias finitas utilizados para aproximar una solución a (3.1.7) son:

$$\frac{u_i^{j+1} - u_i^j}{\Delta t} + a\frac{u_{i+1}^j - u_i^j}{\Delta x} = 0, \tag{3.1.8}$$

$$\frac{u_i^{j+1} - u_i^j}{\Delta t} + a\frac{u_i^j - u_{i-1}^j}{\Delta x} = 0, \tag{3.1.9}$$

$$\frac{u_i^{j+1} - u_i^j}{\Delta t} + a\frac{u_{i+1}^j - u_{i-1}^j}{2\Delta x} = 0, \tag{3.1.10}$$

$$\frac{u_i^{j+1} - u_i^{j-1}}{2\Delta t} + a\frac{u_{i+1}^j - u_{i-1}^j}{2\Delta x} = 0, \tag{3.1.11}$$

$$\frac{u_i^{j+1} - \frac{1}{2}(u_{i+1}^j + u_{i-1}^j)}{\Delta t} + a\frac{u_{i+1}^j - u_{i-1}^j}{2\Delta x} = 0. \tag{3.1.12}$$

El esquema (3.1.8) se conoce como *esquema hacia adelante en el tiempo y hacia adelante en el espacio* por las diferencias hacia adelante que se usan en las aproximaciones de las derivadas. Similarmente, a los esquemas (3.1.9) y (3.1.10) se conoce como esquemas *hacia adelante en el tiempo y hacia atrás en el espacio* y *hacia adelante en el tiempo y centrado en el espacio*, respectivamente. Al esquema (3.1.11) se le conoce como *esquema de salto de rana*, y (3.1.12) se conoce como el *esquema de Lax-Friedrichs*.

Deducir estos esquemas es la parte sencilla del método, ya que su análisis para determinar si proporcionan buenas aproximaciones a la ecuación diferencial requiere herramientas adicionales. Es aún más complicado elaborar esquemas que sean eficientes y precisos. Sin embargo, una de las bondades del método de diferencias finitas clásicas es precisamente la variedad de esquemas que se pueden usar para aproximar una ecuación diferencial.

Cada uno de los esquemas (3.1.8)-(3.1.12) se puede escribir expresando u_i^{j+1} como una combinación lineal de los valores de u en los niveles j y $j-1$. Por ejemplo, el esquema (3.1.8) se puede escribir como

$$u_i^{j+1} = (1 + a\lambda)u_i^j - a\lambda u_{i+1}^j,$$

donde $\lambda = \Delta t/\Delta x$. Esta λ aparece seguido en el estudio de esquemas para ecuaciones hiperbólicas. Los esquemas que involucran a u en solo dos niveles, por ejemplo j y $j+1$, se llaman *esquemas de un paso*. Todos los esquemas mencionados anteriormente son esquemas de un paso, con excepción del esquema de salto de rana. Dada la información inicial u_i^0, se puede usar un esquema de un paso para evaluar u_i^j para todos los valores de j.

El esquema de salto de rana (3.1.11) es un ejemplo de un esquema en multipaso. En un esquema multipaso no es suficiente especificar los valores de u_i^0 para determinar los valores de u_i^j para todos los valores de j, para ello se debe, o bien, especificar los valores de u en suficientes niveles del dominio temporal, o bien encontrar una manera de calcular los valores de u en estos niveles iniciales de tiempo. Por ejemplo, para usar el esquema de salto de rana, se podrían especificar los valores de u_i^0 y u_i^1 para toda i en el dominio, o se podría usar el esquema (3.1.8) para calcular los valores de u_i^1 a partir de los valores de u_i^0. En estos casos, el esquema de salto de rana se puede usar para calcular u_i^j para $j > 1$.

3.1.1 Convergencia y Consistencia

La propiedad más básica que debe poseer un esquema para ser útil es que sus soluciones aproximen a la solución de la ecuación diferencial correspondiente y que la aproximación mejore a medida que la malla se hace más fina. Tal esquema se llama un esquema convergente. Para definir de manera formal este concepto, considere una ecuación diferencial parcial lineal de primer orden en la derivada temporal de la forma

$$P(\partial_t, \partial_x)u = f(x, t).$$

También se asume que para tales ecuaciones o sistemas de ecuaciones, las condiciones iniciales $u(x, 0)$ determinan una solución única.

Definición 3.1.13. Un esquema en diferencias finitas de un paso que aproxima a una ecuación diferencial parcial es un esquema convergente si para cada solución de la ecuación diferencial parcial, $u(x, t)$, y cada solución al esquema en diferencias finitas,

u_i^j, tales que u_i^0 converge a $u_0(x)$ cuando $i\Delta x$ converge a x, entonces u_i^j converge a $u(x,t)$ cuando $(i\Delta x, j\Delta t)$ converge a (x,t) mientras $(\Delta x, \Delta t) \longrightarrow (0,0)$.

En general no es sencillo demostrar que un esquema dado es convergente, si se intenta de una manera directa. Sin embargo, hay dos conceptos relacionados que son sencillos de verificar: consistencia y estabilidad.

Definición 3.1.14. Dada una ecuación diferencial parcial, $Pu = f$, y un esquema en diferencias finitas, $P_{\Delta x,\Delta t}u = f$, se dice que el esquema en diferencias finitas es consistente con la ecuación diferencial si para cada función diferenciable $\phi(x,t)$

$$P\phi - P_{\Delta x,\Delta t}\phi \longrightarrow 0, \qquad \text{si } \Delta x, \Delta t \longrightarrow 0,$$

siendo la convergencia puntual en cada punto (x,t) del dominio.

La consistencia implica que la solución a la ecuación diferencial parcial, si es que es suficientemente diferenciable, es una solución aproximada al esquema de diferencias finitas. De manera análoga, la convergencia significa que una solución al esquema de diferencias finitas aproxima a una solución de la ecuación diferencial parcial. Entonces es una pregunta natural si la consistencia es una condición suficiente para que un esquema sea convergente. Si bien es cierto que la consistencia es una condición necesaria para la convergencia, no es suficiente por si misma, como se muestra en el siguiente ejemplo:

Ejemplo 3.1.15. Considere la ecuación diferencial parcial $u_t + u_x = 0$ con el esquema hacia adelante en el tiempo y hacia adelante en el espacio:

$$\frac{u_i^{j+1} - u_i^j}{\Delta t} + \frac{u_{i+1}^j - u_i^j}{\Delta x} = 0.$$

Este esquema se puede rescribir como

$$\begin{aligned} u_i^{j+1} &= u_i^j - \frac{\Delta t}{\Delta x}(u_{i+1}^j - u_i^j), \\ &= (1+\lambda)u_i^j - \lambda u_{i+1}^j, \end{aligned} \tag{3.1.16}$$

donde $\lambda = \Delta t/\Delta x$. Se puede demostrar que este esquema es consistente (ver [78]). Como condición inicial, tome

$$u_0(x) = \begin{cases} 1 & \text{si } -1 \leqslant x \leqslant 0, \\ 0 & \text{en otro caso.} \end{cases}$$

La solución a la ecuación diferencial parcial es una traslación de u_0 a la derecha en t unidades. En particular, para $t > 0$, existen valores positivos de x para los cuales $u(t,x)$ no es cero.

Para el esquema en diferencias, se toma la condición inicial

$$u_i^0 = \begin{cases} 1 & \text{si } -1 \leqslant i\Delta x \leqslant 0, \\ 0 & \text{en otro caso.} \end{cases}$$

Como se muestra en la ecuación (3.1.16), la solución al esquema en diferencias en (t_j, x_i) depende únicamente de $x_{i'}$ para $i' \geqslant i$ en pasos anteriores. Entonces se concluye que u_i^j es siempre nulo para puntos x_i a la derecha de cero, es decir,

$$u_i^j = 0 \qquad \text{para } i > 0, j \geqslant 0.$$

Por tanto, u_i^j no puede converger a $u(t, x)$, ya que para valores positivos de t y x, la función u no se anula, pero u_i^j sí.

3.1.2 Estabilidad

En un esquema convergente se tiene que u_i^j converge a $u(x, t)$, lo cual implica que la sucesión formada por los u_i^j debe ser acotada. Esta idea es la escencia de la estabilidad.

Antes de definir la estabilidad, es conveniente hablar de la región de estabilidad. Existen restricciones para muchos esquemas sobre cómo se deben elegir Δx y Δt para que el esquema sea estable, y por tanto útil computacionalmente. Una región de estabilidad es una región acotada y no vacía en el primer cuadrante del plano $\mathbf{R}^2$, el cual tiene al origen como punto de acumulación, lo que implica la existencia de una sucesión (h_ν, k_ν) que converge al origen conforme ν crece arbitrariamente. Un ejemplo común es la región $\{(h, k) : 0 < k \leqslant ch \leqslant C\}$, donde c y C son constantes positivas.

Definición 3.1.17. Un esquema en diferencias finitas $P_{h,k}u_i^j = 0$ para una ecuación de primer orden es estable en una región Λ si existe un entero J tal que para cualquier valor positivo del tiempo T, existe una constante C_T tal que

$$h \sum_{m=-\infty}^{\infty} |u_m^j|^2 \leqslant C_T h \sum_{l=0}^{J} \sum_{m=-\infty}^{\infty} |u_m^l|^2, \tag{3.1.18}$$

para $0 \leqslant jk \leqslant T$, con $(h, k) \in \Lambda$.

El concepto de estabilidad para esquemas en diferencias finitas está muy relacionado al concepto de problema bien planteado en problemas de valores iniciales para ecuaciones diferenciales parciales. Esta discusión se restringe a ecuaciones de la forma $Pu = f$, que son de primer orden en la derivada temporal.

Definición 3.1.19. El problema de valores iniciales para la ecuación parcial de primer orden $Pu = 0$ es bien planteado si para cada tiempo $T \geqslant 0$, existe una constante C_T tal que cualquier solución $u(x, t)$ satisface

$$\int_{-\infty}^{\infty} |u(x, t)|^2 \, dx \leqslant C_T \int_{-\infty}^{\infty} |u(x, 0)|^2 \, dx \tag{3.1.20}$$

para $0 \leqslant t \leqslant T$.

Se puede demostrar que solo los problemas bien planteados de valores iniciales se pueden usar para modelar la evolución de procesos físicos.

3.1.3 El teorema de equivalencia de Lax-Richtmyer

La importancia de los conceptos de consistencia y estabilidad se ve en el teorema de equivalencia de Lax-Richtmyer, que es el teorema fundamental en la teoría de esquemas en diferencias finitas para problemas de valores iniciales.

Teorema 3.1.21. ***(Teorema de equivalencia de Lax-Richtmyer)*** *Un esquema consistente en diferencias finitas para una ecuación diferencial parcial para el cual el problema de valores iniciales es bien planteado es convergente sí, y sólamente si, el esquema es estable.*

La demostración de este teorema se puede consultar en [78]. La utilidad de este teorema radica en la caracterización simple de los esquemas convergentes. Determinar si un esquema dado es convergente puede ser muy complicado si se trata de verificar usando la definición de manera directa. En lugar de ello, verificar la consistencia o la estabilidad de un esquema es bastante más sencillo. Entonces el resultado más complejo - la convergencia - es reemplazado por una prueba equivalente de consistencia y estabilidad, que es la principal utilidad de este resultado.

3.1.4 La condición de Courant-Friedrichs-Lewy

La ecuación de onda (3.1.7) se emplea también para modelar fenómenos de advección escalar en una dimensión espacial. En los esquemas (3.1.8)-(3.1.12) se puede escribir el término u_i^{j+1} como combinación lineal de los valores de u en los niveles j y $j-1$, para lo cual usualmente se denomina $\lambda = \Delta t/\Delta x$. En el caso de que esta λ sea constante para un esquema en diferencias finitas, la condición $|a\lambda| \leqslant 1$ es necesaria para la estabilidad de varios esquemas en diferencias finitas para la ecuación (3.1.7).

En un esquema explícito en diferencias finitas se puede escribir el término u_i^{j+1} como una suma finita de términos $u_{i'}^{j'}$ con $j' \leqslant j$. Todos los esquemas mencionados hasta ahora son explícitos. El siguiente resultado se aplica a los esquemas de un paso que se han mencionado en este capítulo:

Teorema 3.1.22. *Para un esquema explícito de la ecuación hiperbólica (3.1.7) de la forma* $u_i^{j+1} = \alpha u_{i-1}^j + \beta u_i^j + \gamma u_{i+1}^j$ *con* $\lambda = \Delta t/\Delta x$ *constante, una condición necesaria para la estabilidad es la* condición de Courant-Friedricks-Lewy (CFL)*,*

$$|a\lambda| \leqslant 1.$$

Para sistemas de ecuaciones en los cuales u *es un vector y* α*,* β *y* γ *son matrices, se debe tener* $|a_i\lambda| \leqslant 1$ *para todos los valores propios* a_i *de la matriz* A*.*

La demostración del teorema (3.1.22) se puede consultar en [78]. Un análisis más detallado de la estabilidad de la ecuación (3.1.7) se puede consultar en [58].

Se puede usar un argumento similar para mostrar que no hay esquemas explícitos consistentes para ecuaciones diferenciales hiperbólicas que sean estables para todos los valores de λ, con λ constante mientras $\Delta x, \Delta t \to 0$. El siguiente resultado se atribuye a Courant, Friedricks y Lewy.

Teorema 3.1.23. *No existen esquemas en diferencias finitas para sistemas hiperbólicos de ecuaciones diferenciales parciales que sean explícitos, incondicionalmente estables y consistentes.*

La *velocidad numérica* de propagación para un esquema de la forma considerada en el teorema (3.1.22) es $\Delta x/\Delta t = \lambda^{-1}$, dado que la información se puede propagar en un paso en el espacio por cada paso en el tiempo. La condición CFL se puede escribir como

$$|a| \leqslant \lambda^{-1},$$

lo cual se puede interpretar afirmado que la velocidad numérica de propagación debe ser al menos igual a la velocidad de propagación de la ecuación diferencial. Esta es la idea básica de estos teoremas. Si el esquema numérico no puede propagar la solución al menos a la misma velocidad de la solución de la ecuación diferencial, entonces la solución del esquema no puede converger a la solución de la ecuación diferencial parcial.

3.2 El problema de Poisson en diferencias finitas clásicas

Un ejemplo clásico de aplicación de los esquemas en diferencias finitas clásicas es la ecuación de difusión, también conocida como la ecuación de conducción de calor, la cual en su forma bidimensional en el espacio toma la forma:

$$u_t = (\kappa u_x)_x + (\kappa u_y)_y + \psi, \tag{3.2.1}$$

donde $\kappa(x, y) > 0$ es un coeficiente de difusión o conducción de calor que puede variar con la posición espacial, y $\psi(x, y, t)$ es un término fuente. La solución $u(x, y, t)$ generalmente depende de la posición y el tiempo. También se requieren condiciones iniciales $u(x, y, 0)$ en el dominio espacial Ω de u, así como condiciones de frontera en cada punto de la frontera de Ω para cada instante t. Si las condiciones de frontera y los términos fuente son independientes del tiempo, entonces se espera un estado estacionario, el cual se puede encontrar resolviendo la ecuación elíptica

$$(\kappa u_x)_x + (\kappa u_y)_y = f, \tag{3.2.2}$$

donde se fija $f(x, y) = -\psi(x, y)$, junto con condiciones de frontera.

Si se considera el caso más simple, donde $\kappa \equiv 1$, se tiene el *problema de Poisson*

$$u_{xx} + u_{yy} = f. \tag{3.2.3}$$

Este problema se completa con las condiciones de frontera especificadas en $\partial\Omega$.

3.2.1 El esténcil de cinco puntos para el Laplaciano

Para fijar ideas, considere el problema de Poisson (3.2.3) definido en el cuadrado unitario $0 \leqslant x \leqslant 1$, $0 \leqslant y \leqslant 1$ y suponga condiciones de Dirichlet a la frontera. Se usará un mallado uniforme de puntos (x_i, y_j), donde $x_i = i\Delta x$ y $y_j = j\Delta y$.

Sea u_{ij} la aproximación a $u(x_i, y_j)$. Para discretizar la ecuación (3.2.3), se remplazan las derivadas parciales en x y y con diferencias finitas centradas, es decir:

$$\frac{1}{(\Delta x)^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}) + \frac{1}{(\Delta y)^2}(u_{i,j-1} - 2u_{i,j} + u_{i,j+1}) = f_{i,j}. \tag{3.2.4}$$

Por simplicidad, se considera el caso de un mallado regular en el que $\Delta x = \Delta y = h$. De esta manera, se puede escribir (3.2.4) como:

$$\frac{1}{h^2}(u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} - 4u_{i,j}) = f_{i,j}. \tag{3.2.5}$$

Esta discretización sobre un mallado regular se muestra en la figura (3.1), donde se muestra la posición del esténcil junto con los pesos de cada nodo, según la ecuación (3.2.5).

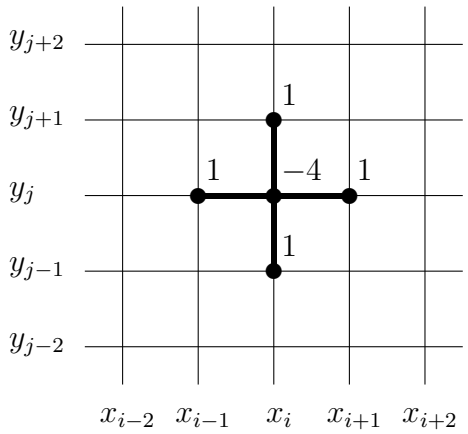


Figura 3.1: Esténcil de cinco puntos.

3.2.2 Precisión y estabilidad

El error local de truncamiento τ_{ij} en el punto (i, j) del mallado se define como

$$\tau_{ij} = \frac{1}{h^2}(u(x_{i-1}, y_j) + u(x_{i+1}, y_j) + u(x_i, y_{j-1}) + u(x_i, y_{j+1}) - 4u(x_i, y_j)) - f(x_i, y_j).$$

Separando esta expresión hasta segundo orden en las direcciones x y y, se tiene que

$$\tau_{ij} = \frac{1}{12}h^2(u_{xxxx} + u_{yyyy}) + \mathcal{O}(h^4).$$

El error global $E_{ij} = u_{ij} - u(x_i, y_j)$ es solución del sistema lineal

$$A^h E^h = -\tau^h,$$

donde A^h es la matriz de discretización con tamaño de paso h. Este método es globalmente preciso hasta segundo orden en alguna norma si es que es estable, o sea, si $||(A^h)^{-1}||$ es uniformemente acotada conforme $h \to 0$.

Esto es fácil de verificar en la norma 2, ya que se pueden calcular explícitamente el radio espectral de la matriz. Los valores y vectores propios de A se pueden indizar con los parámetros p y k correspondientes a los números de onda en las direcciones x y y para $p, k = 1, 2, \ldots, m$. El vector propio $u^{p,q}$ tiene los m^2 elementos

$$u_{ij}^{p,q} = \sin(p\pi ih)\sin(q\pi ih).$$

El valor propio correspondiente es

$$\lambda_{p,q} = \frac{2}{h^2}((\cos(p\pi h) - 1) + (\cos(q\pi h) - 1)).$$

Los valores propios son estríctamente negativos (A es negativa definida) y el más cercano al origen es

$$\lambda_{1,1} = -2\pi^2 + \mathcal{O}(h^2).$$

El radio espectral de $(A^h)^{-1}$, que es también su 2-norma, es entonces

$$\rho((A^h)^{-1}) = \frac{1}{\lambda_{1,1}} \approx -\frac{1}{2\pi^2}.$$

Por tanto, el método es estable en la norma 2.

Por otro lado, el número de condición en la norma 2 se define como

$$\kappa_2(A) = ||A||_2||A^{-1}||_2.$$

Dado que $||(A^h)^{-1}||_2 \approx -1/2\pi^2$ para h pequeña, y la norma de A está dada por su radio espectral. El valor propio más grande de A (en magnitud) es

$$\lambda_{m,m} \approx -\frac{8}{h^2},$$

y así

$$\kappa_2(A) \approx \frac{4}{\pi^2 h^2} = \mathcal{O}\left(\frac{1}{h^2}\right), \qquad \text{cuando } h \to 0.$$

El hecho de que la matriz sea muy mal condicionada mientras se redefine la malla es responsable de que los métodos iterativos se vuelvan lentos.

3.3 La ecuación de difusión en diferencias finitas clásicas

Considere la ecuación de difusión en una dimensión espacial

$$u_t = \kappa u_{xx}. \tag{3.3.1}$$

A esta ecuación también se le conoce como *ecuación de calor*, y es un ejemplo clásico de ecuación parabólica. Para fines de esta expliacación, se asume $\kappa = 1$. Considere también las condiciones iniciales

$$u(x, 0) = \eta(x),$$

así como condiciones de Dirichlet a la frontera

$$\begin{aligned} u(0, t) &= g_0(t) \qquad \text{para } t > 0, \\ u(1, t) &= g_1(t) \qquad \text{para } t > 0, \end{aligned}$$

para $0 \leqslant x \leqslant 1$.

La discretización del dominio espacial y temporal está dada por el mallado (x_i, t_n), donde

$$x_i = ih, \qquad t_n = nk.$$

Se toma $h = \Delta x$ como el tamaño de paso en el dominio temporal y $k = \Delta t$ es el tamaño de paso en el dominio temporal. Sea $U_i^n \approx u(x_i, t_n)$ la aproximación numérica en el punto (x_i, t_n).

Para calcular una solución numérica a la ecuación de calor, se sugiere un esquema con un paso hacia adelante en el tiempo, el cual permita aproximar U_i^{n+1} para cualquier i a partir de los valores de U_i^n del paso anterior, o tal vez usando también algunos pasos anteriores en un esquema de paso múltiple.

Un ejemplo de discretización podría ser

$$\frac{U_i^{n+1} - U_i^n}{k} = \frac{1}{h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n).$$

Este esquema usa una diferencia centrada en el espacio y hacia adelante en el tiempo. Este es un método explícito que permite calcular U_i^{n+1} explícitamente usando la información del paso anterior:

$$U_i^{n+1} = U_i^n + \frac{k}{h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n). \tag{3.3.2}$$

Este esténcil se muestra en la figura (3.2).

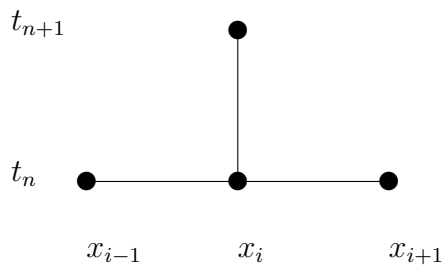


Figura 3.2: Esténcil para el esquema (3.3.2).

Otro método muy usado en la práctica es el *método de Crank-Nicholson*,

$$\begin{aligned}\frac{U_i^{n+1}-U_i^n}{k} &= \frac{1}{2}(D^2U_i^n + D^2U_i^{n+1}) \\ &= \frac{1}{2h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n + U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}),\end{aligned}$$

el cual se puede escribir como

$$U_i^{n+1} = U_i^n + \frac{k}{2h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n + U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}), \tag{3.3.3}$$

o también como

$$-rU_{i-1}^{n+1} + (1+2r)U_i^{n+1} - rU_{i+1}^{n+1} = rU_{i-1}^n + (1-2r)U_i^n + rU_{i+1}^n, \tag{3.3.4}$$

donde $r = k/2h^2$. El esténcil para el método de Crank-Nicholson se muestra en la figura (3.3).

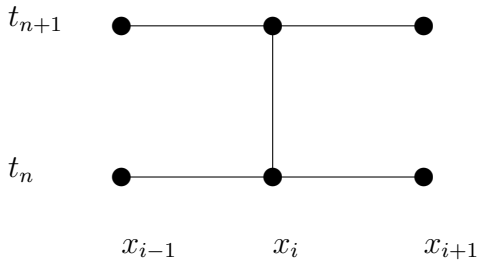


Figura 3.3: Esténcil para el esquema (3.3.3).

Crank-Nicholson (3.3.4) es un método implícito que genera un sistema tridiagonal de ecuaciones lineales para resolver para todos los valores de U_i^{n+1} de manera simultánea. En forma matricial, este esquema genera el sistema

$$\begin{bmatrix} (1+2r) & -r & & & & & \\ -r & (1+2r) & -r & & & & \\ & -r & (1+2r) & -r & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -r & (1+2r) & -r & \\ & & & & -r & (1+2r) \end{bmatrix} \begin{bmatrix} U_1^{n+1} \\ U_2^{n+1} \\ U_3^{n+1} \\ \vdots \\ U_{m-1}^{n+1} \\ U_m^{n+1} \end{bmatrix}$$

$$= \begin{bmatrix} r(g_0(t_n) + g_0(t_{n+1})) + (1-2r)U_1^n + rU_2^n \\ rU_1^n + (1-2r)U_2^n + rU_3^n \\ rU_2^n + (1-2r)U_3^n + rU_4^n \\ \vdots \\ rU_{m-2}^n + (1-2r)U_{m-1}^n + rU_m^n \\ rU_{m-1}^n + (1-2r)U_m^n + r(g_1(t_n) + g_1(t_{n+1})) \end{bmatrix}.$$

Nótense las condiciones de frontera $u(0,t) = g_0(t)$ y $u(1,t) = g_1(t)$ en el sistema. Dada la estructura tridiagonal de este sistema, existen métodos eficientes para resolver este método implícito.

3.4 Diferencias finitas generalizadas

Los esquemas en diferencias finitas clásicas suelen dar buenos resultados bajo ciertas condiciones, sin mencionar su practicidad y simplicidad. Una cualidad deseable en un esquema en diferencias finitas clásicas es preservar ciertas simetrías, las cuales pueden estar presentes en el dominio del problema, el esténcil o la discretización de los operadores. Preservas estas simetrías en un esquema de diferencias finitas tiene sus ventajas, por ejemplo, la estructura matricial del sistema (3.3.4), la cual permite calcular una solución de manera eficiente.

Sin embargo, no siempre es posible preservar toda la estructura y simetrías deseables en un esquema en diferencias finitas. Muchas veces los problemas físicos que se desea modelar no presentan una estructura tan definida, por ejemplo, al calcular un esquema para resolver una ecuación diferencial parcial sobre un dominio irregular. En estas situaciones, los problemas físicos demandan que los esquemas se adapten a la situación particular, perdiendo parte de la estrucura y simetría a cambio de una mejor adaptación al problema, lo cual puede llevar a mejores resultados. Es esta idea la que lleva a generalizar los esquemas en diferencias finitas: brindar una mejor adaptación a un problema.

Si bien es cierto que no es posible preservar de manera intrínseca la simetría y estructura de los esquemas en diferencias finitas clásicas, esta adaptación trae consigo nuevas estructuras en la discretización de los operadores, lo que abre las puertas a explorar nuevas propiedades relacionadas a ciertos esquemas.

La discretización del dominio de la ecuación diferencial es un proceso que se debe llevar a cabo de manera complementaria a la discretización del operador. Hoy en día existen una variedad de métodos para generación de mallados, de entre los cuales mencionaremos en primera instancia a los métodos de tipo algebraico, que se basan en una interpolación a lo largo del dominio de la ecuación diferencial. En un dominio rectangular, una interpolación es simplemente definir un mallado de manera regular, dividiendo al dominio en rectángulos pequeños, por ejemplo. En dominios más sofisticados o irregulares, se puede recurrir a la interpolación transfinita en dos o tres dimensiones, la cual es básicamente una interpolación de tipo Lagrange: se inicia con una superficie, y se transforma de manera continua a través de un mapeo entre el espacio físico y el espacio computacional. Al definir este tipo de transformaciones, se busca que el mapeo preserve ciertas propiedades: se define un difeomorfísmo en la frontera del dominio, y lo que se busca es extender este mapeo de manera continua y diferenciable al interior del dominio.

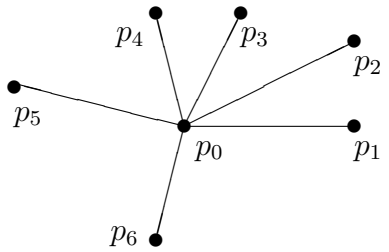
Uno de los principales grupos de generadores de malla se obtiene con ecuaciones diferenciales, donde se busca que el mapeo del espacio físico al espacio computacional sea armónico, es decir, que ambas componentes de la transformación satisfagan la ecuación de Laplace. En regiones con una geometría relativamente simple, estos generadores proporcionan mallados bastante buenos. En este caso, el problema elíptico correspondiente al mapeo inverso lleva a las ecuaciones de Winslow.

Como ya se mencionó, la discretización del dominio y la discretización del operador son procesos que se retroalimentan entre sí. Aún en el caso de un dominio rectangular y regular, una vez que se obtiene un mallado, se puede observar la solución para determinar si existen regiones en el dominio donde la norma del gradiente es demasiado grande, para resolver de nueva cuenta y refinar el mallado en esa zona en particular. De esta manera, la discretización del operador diferencial sugiere una discretización para el dominio.

En esta sección se presentan los esquemas en diferencias finitas generalizadas utilizados en el desarrollo del presente trabajo, así como algunas de sus propiedades básicas.

3.4.1 Los esquemas propuestos

De manera general, se puede pensar que en la formulación de un esquema en diferencias finitas generalizadas se pretende adaptar, tanto la discretización del dominio como del operador involucrados, al problema físico en cuestión, de tal manera que la solución numérica obtenida cumpla con las expectativas deseadas. En esta sección se presentan las propuestas de esquemas en diferencias generalizadas empleados en el desarrollo de este proyecto.



En una primera propuesta, se piensa en un esténcil de siete nodos, que comprenden seis nodos colocados alrededor de un nodo central. Los nodos se han denotado como $p_i = (x_i, y_i)$ para $i = 0, 1, \ldots, 6$, donde p_0 es el nodo central, y el resto de los nodos se enumeran en orden opuesto a las manecillas del reloj, comenzando por el nodo a la derecha del nodo central.

Al usar estos esténciles en esquemas en diferencias finitas generalizadas para estudiar problemas de tipo Poisson, surgen matrices que siguen una estructura particular, las cuales tiene la forma:

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ y_1 & y_2 & y_3 & y_4 & y_5 & y_6 \\ x_1^2 & x_2^2 & x_3^2 & x_4^2 & x_5^2 & x_6^2 \\ x_1y_1 & x_2y_2 & x_3y_3 & x_4y_4 & x_5y_5 & x_6y_6 \\ y_1^2 & y_2^2 & y_3^2 & y_4^2 & y_5^2 & y_6^2 \end{pmatrix}. \tag{3.4.1}$$

Al trabajar con matrices con esta estructura, se busca establecer condiciones necesarias sobre el mallado para garantizar que la matriz (3.4.1) sea no singular, así como establecer condiciones para que el sistema no homogéneo de ecuaciones lineales $\mathbf{M\Gamma} = \mathbf{B}$ tenga una solución no trivial $\mathbf{\Gamma}$, donde este vector $\mathbf{\Gamma}$ está formado por los pesos en los nodos vecinos del nodo central, los cuales se buscan de tal manera que se satisfaga la siguiente condición de consistencia. Sea φ una función de clase $\mathcal{C}^2$ definida sobre un dominio en el plano. Considere el operador lineal de segundo orden:

$$L(\varphi, K_1(x,y), K_2(x,y), K_3(x,y), K_4(x,y), K_5(x,y)) =$$
$$K_1(x,y)\frac{\partial^2\varphi}{\partial x^2} + K_2(x,y)\frac{\partial^2\varphi}{\partial y^2} + K_3(x,y)\frac{\partial\varphi}{\partial x} + K_4(x,y)\frac{\partial\varphi}{\partial y} + K_5(x,y)\varphi.$$

Nótese la dependencia de los coeficientes K_i, $i = 1, \ldots, 5$ en la posición (x, y). A partir de la evaluación de este operador en el nodo central del esténcil, se define la combinación lineal

$$L_0 := \sum_{l=1}^{6} \Gamma_l \left(p_0, p_l, K_1(x_0,y_0), K_2(x_0,y_0), K_3(x_0,y_0), K_4(x_0,y_0), K_5(x_0,y_0)\right) \varphi(p_l),$$

Nótese que la dependencia de cada coeficiente Γ_l con la posición de cada nodo del esténcil. Los coeficientes Γ_l se buscan de tal manera que se cumpla la condición de consistencia:

$$\tau(p_0) := \left[L(\varphi, K_1(x,y), K_2(x,y), K_3(x,y), K_4(x,y), K_5(x,y))\right]_{(x_0,y_0)} - L_0 \to 0,$$

conforme $p_1, \ldots p_6 \to p_0$, de acuerdo con [16]. Para fines de esta explicación, se abrevia:

$$\Gamma_l := \Gamma_l\left(p_0, p_l, K_1(x_l, y_l), K_2(x_l, y_l), K_3(x_l, y_l), K_4(x_l, y_l), K_5(x_l, y_l)\right).$$

Sean Δx_l y Δy_l las componentes x y y de $p_l - p_0$, respectivamente. Entonces el error local de truncamiento $\tau(p_0)$ produce

$$\begin{aligned}
\tau(p_0) &= \\
\left(K_5(x_0,y_0) - \sum_{i=0}^{6}\Gamma_i\right)\varphi(p_0) + \left(K_3(x_0,y_0) - \sum_{i=1}^{6}\Gamma_i\Delta x_i\right)\frac{\partial\varphi}{\partial x}(p_0) &+ \\
\left(K_4(x_0,y_0) - \sum_{i=1}^{6}\Gamma_i\Delta y_i\right)\frac{\partial\varphi}{\partial y}(p_0) + \left(K_1(x_0,y_0) - \sum_{i=1}^{6}\frac{\Gamma_i(\Delta x_i)^2}{2}\right)\frac{\partial^2\varphi}{\partial x^2}(p_0) &+ \\
\left(-\sum_{i=1}^{6}\Gamma_i\Delta x_i\Delta y_i\right)\frac{\partial^2\varphi}{\partial x\partial y}(p_0) + \left(K_2(x_0,y_0) - \sum_{i=1}^{6}\frac{\Gamma_i(\Delta y_i)^2}{2}\right)\frac{\partial^2\varphi}{\partial y^2}(p_0) &+ \\
\mathcal{O}\left(\text{máx}\{\Delta x_i, \Delta y_i\}\right)^3.
\end{aligned}$$

De esta manera, dado que la función φ es de clase $\mathcal{C}^2$, la ecuación anterior se puede escribir como un sistema de ecuaciones lineales de la forma $\mathbf{M\Gamma} = \mathbf{B}$:

$$\begin{pmatrix}
1 & 1 & \dots & 1 \\
0 & \Delta x_1 & \dots & \Delta x_6 \\
0 & \Delta y_1 & \dots & \Delta y_6 \\
0 & (\Delta x_1)^2 & \dots & (\Delta x_6)^2 \\
0 & \Delta x_1\Delta y_1 & \dots & \Delta x_6\Delta y_6 \\
0 & (\Delta y_1)^2 & \dots & (\Delta y_6)^2
\end{pmatrix}
\begin{pmatrix}
\Gamma_0 \\ \Gamma_1 \\ \Gamma_2 \\ . \\ . \\ . \\ \Gamma_6
\end{pmatrix} =
\begin{pmatrix}
K_5(x_0,y_0) \\ K_3(x_0,y_0) \\ K_4(x_0,y_0) \\ 2K_1(x_0,y_0) \\ 0 \\ 2K_2(x_0,y_0)
\end{pmatrix}. \tag{3.4.2}$$

Cabe destacar que, en general, este sistema de ecuaciones lineales no está bien determinado. Un truco práctico para calcular la solución de este sistema consiste en remover la el primer renglón y la primera columna de la matriz $\mathbf{M}$, junto con Γ_0 y el respectivo lado derecho $K_5(x_0, y_0)$, ya que Γ_0 se determina mediante la primera ecuación del sistema

$$\sum_{i=0}^{6}\Gamma_i = K_5(x_0, y_0).$$

Un ejemplo con un problema estacionario

Para poner un ejemplo del uso de los esquemas que se proponen, considere el problema estacionario de difusión-advección en 2D:

$$u\frac{\partial C}{\partial x} + v\frac{\partial C}{\partial y} = \frac{\partial}{\partial x}\left(A\frac{\partial C}{\partial x}\right) + \frac{\partial}{\partial y}\left(A\frac{\partial C}{\partial y}\right) \tag{3.4.3}$$

definido sobre el dominio $\Omega = [0,1]\times[0,1]$, con los parámetros $u = v = 0.1$, $A(y) = 1 + e^{-10y}$ para $0 \leqslant y \leqslant 1$, y las condiciones de frontera $C(0,y) = C(1,y) = 0.05$ para $0 \leqslant y \leqslant 1$, $\frac{\partial C}{\partial n} = 0$ en $(0,1)\times\{1\}$ y $\frac{\partial C}{\partial n} = g(x)$ en $(0,1)\times\{0\}$, donde

$$g(x) = \begin{cases} 0.5 & \frac{3}{8} \leqslant x \leqslant \frac{5}{8} \\ 0 & \text{de otro modo} \end{cases}$$

El dominio Ω se discretiza de manera regular, en cuadrados de la misma longitud. Para los fines de este ejemplo, se usaron 201 nodos a lo largo de cada dirección x y y. Los nodos de esta discretización se enumeran desde 1 hasta 201 en ambas direcciones, por lo que los nodos tienen la forma $p_{i,j} = (x_i, y_j)$ con $i, j = 1, 2, \ldots, 201$. En cada cuadrado del dominio discretizado, se añade la diagonal que va de la esquina superior derecha a la esquina inferior izquierda. De esta manera, el dominio se discretiza en elementos triángulares de la siguiente manera:

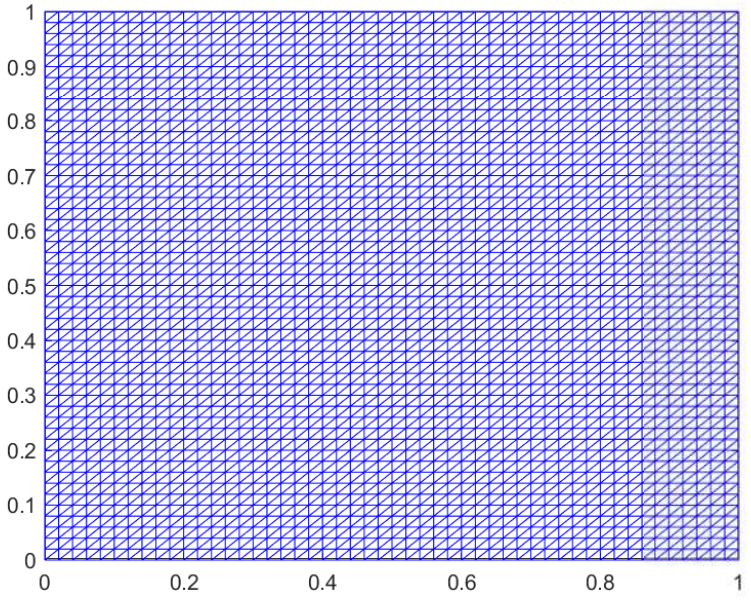


Figura 3.4: Mallado usado para el cálculo de la solución numérica de (3.4.3).

La diagonal se añade en cada cuadrado del mallado para construir el esténcil de siete puntos del que se habla en la sección anterior. Este esténcil es una modificación de (3.1) en la que el nodo central se conecta con seis nodos vecinos, formado el esténcil de la figura (3.5).

Ahora se plantea el sistema (3.4.2) de acuerdo a la discretización del dominio (3.4) y de acuerdo con el problema (3.4.3). De acuerdo a la distribución de los nodos, es claro que para cualquier $i = 1, 2, \ldots 6$, se tiene que $\Delta x_i = \Delta y_i$; sea Δ esta cantidad en común. Removiendo el primer renglón y la primer columna de la matriz $\mathbf{M}$ de (3.4.2), se plantea el sistema $\mathbf{M}^{-}\mathbf{\Gamma}^{-} = \mathbf{B}^{-}$, donde

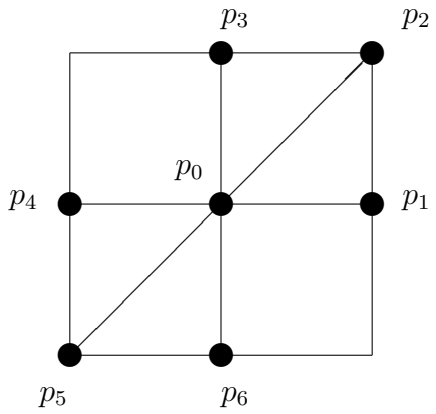


Figura 3.5: Esténcil de siete puntos.

$$\mathbf{M}^- := \begin{pmatrix} \Delta & \Delta & 0 & -\Delta & -\Delta & 0 \\ 0 & \Delta & \Delta & 0 & -\Delta & -\Delta \\ \Delta^2 & \Delta^2 & 0 & \Delta^2 & \Delta^2 & 0 \\ 0 & \Delta^2 & 0 & 0 & \Delta^2 & 0 \\ 0 & \Delta^2 & \Delta^2 & 0 & \Delta^2 & \Delta^2 \end{pmatrix}, \tag{3.4.4}$$

$$\mathbf{\Gamma}^- = \begin{pmatrix} \Gamma_1 \\ \Gamma_2 \\ . \\ . \\ . \\ \Gamma_6 \end{pmatrix}, \quad \mathbf{B}^- = \begin{pmatrix} -u(x_0, y_0) \\ \frac{\partial A}{\partial y}(x_0, y_0) - v(x_0, y_0) \\ 2A(x_0, y_0) \\ 0 \\ 2A(x_0, y_0) \end{pmatrix}.$$

En el planteamiento de este sistema, cabe mencionar que los nodos de la frontera requieren un tratamiento adecuado de la información proporcionada en el problema. Los segmentos de la frontera con condiciones de Dirichlet pasan al lado derecho en sus respectivas posiciones para formar parte del vector $\mathbf{B}^-$. En cuanto a los segmentos de la frontera con condiciones de Neumann, la discretización del operador requiere una condición adicional.

En el caso del segmento de frontera $\{(x, 1) : 0 < x < 1\}$, la condición de frontera es

$$\frac{\partial C}{\partial n} = \left(\frac{\partial C}{\partial x}, \frac{\partial C}{\partial y}\right) \bullet (0, 1) = \frac{\partial C}{\partial y}$$

Para la discretización de este operador en la frontera, se añade un nodo adicional en cada fila, como se muestra en la figura (3.6). La parte azul de la figura corresponde a la parte del mallado (3.4), mientras que la parte negra es el agregado al esténcil.

Para cada nodo p_0 de la frontera superior, se añade el nodo p_3 del esténcil (3.5), pero se descarta el nodo p_2 del mismo para no agregar grados de libertad adicionales.

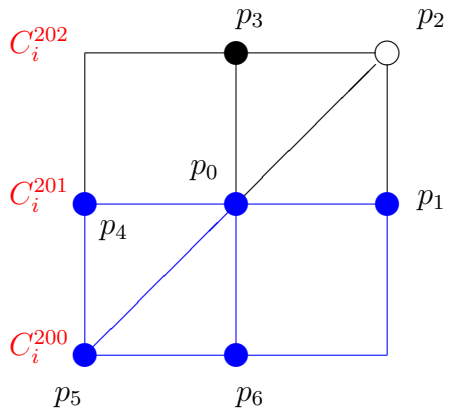


Figura 3.6: Nodo adicional en la frontera superior.

De acuerdo a esta distribución de nodos, se usa una aproximación de segundo orden para la derivada normal en esta frontera:

$$\frac{\partial C}{\partial y} \approx \frac{C_i^{202} - C_i^{200}}{2\Delta y} = 0, \qquad i = 2, 3, \ldots 200.$$

De donde se tiene que el peso en el nodo adicional debe coincidir con el peso del nodo en el rengón 200.

El tratamiento del problema en la frontera inferior es muy similar. En este caso, la condición de frontera es:

$$\frac{\partial C}{\partial n} = \left(\frac{\partial C}{\partial x}, \frac{\partial C}{\partial y}\right) \bullet (0, -1) = -\frac{\partial C}{\partial y}$$

De manera análoga, se añade un nodo al esténcil en cada nodo de la frontera inferior, como se muestra en la figura (3.7). La parte azul es la que corresponde al mallado (3.4), y la parte negra es el agregado al esténcil.

Para cada nodo p_0 de la frontera inferior, se agrega el nodo p_6 del esténcil (3.5), pero no se añade el nodo p_5 para no agregar más grados de libertad. De acuerdo a esta discretización, y usando una aproximación de segundo orden para la derivada normal, se tiene

$$\frac{\partial C}{\partial y} \approx \frac{C_i^2 - C_i^{-1}}{2\Delta y}, \qquad i = 2, 3, \ldots 200.$$

Entonces la condición de frontera en el segmento $\{(x, 0) : 0 < x < 1\}$ se puede escrbir como

$$\frac{C_i^{-1} - C_i^2}{2\Delta y} = g(x_i), \qquad i = 2, 3, \ldots 200.$$

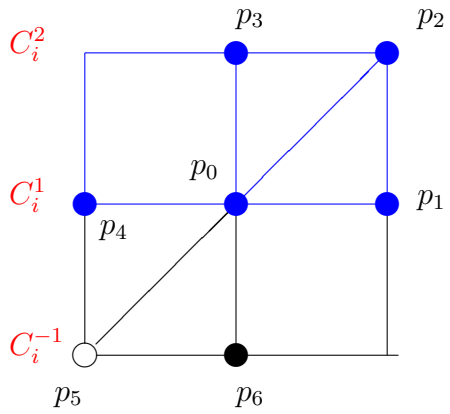


Figura 3.7: Nodo adicional en la frontera inferior.

De esta expresión se obtiene el peso para el nodo adicional en el renglón -1 como $C_i^{-1} = C_i^2 + 2\Delta y \cdot g(x_i)$ para $i = 2, 3, \ldots 200$.

Realizando estas adaptaciones con los nodos adicionales en las fronteras con condiciones de Neumann, se planteó un sistema de ecuaciones lineales de acuerdo al ensamblaje de la matriz mostrado en [45], donde se incluyen todos los nodos del mallado con sus respectivos pesos, así como el vector que contiene la información sobre las condiciones de frontera del problema y las partes no homogéneas del sistema, con lo cual se obtuvo la solución numérica mostrada en (3.8).

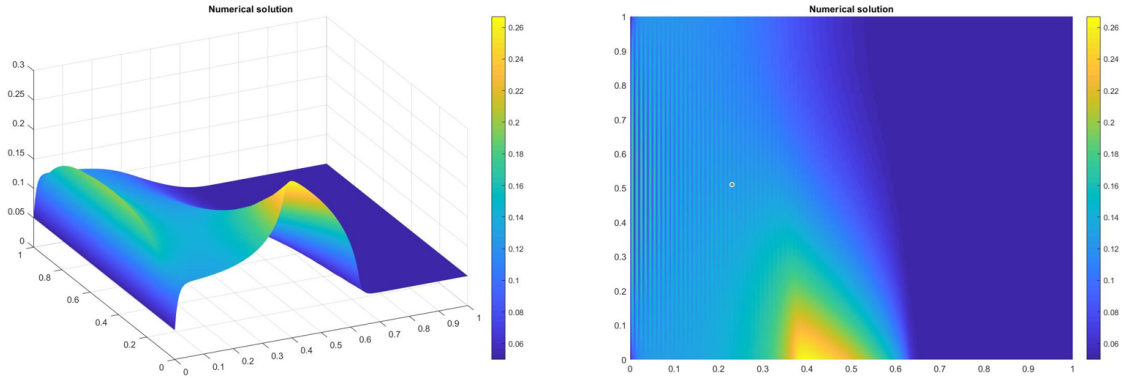


Figura 3.8: Solución numérica obtenida para el problema (3.4.3).

El esténcil (3.5) usado en este esquema tiene una simetría muy particular, esta estructura determina algunas de las propiedades en el sistema de ecuaciones lineales (3.4.2) al plantear la discretización del problema (3.4.3). En primer lugar, es importante destacar que la matriz de diferenciación resultante del ensamble del sistema según [45] es una matriz tridiagonal por bloques, la cual tiene rango completo por filas, sin embargo, dicha matriz carece de normalidad, lo que impide realizar cierto análisis ma-

tricial, en particular, no es posible calcular cotas específicas para sus valores propios, ni tampoco es inmediato concluir que los errores locales de truncamiento carecen de un factor de crecimiento. Sin embargo, cabe resaltar que las disretizaciones empleadas tanto para el operador como para el dominio, arrojan resultados aceptables; en particular, esta discretización permite modelar la condición de frontera de Neumann en la frontera inferior del dominio, donde el gradiente crece de manera abrupta. A diferencia de otros modelos u otras discretizaciones que requieren hacer un refinamiento cerca de esta pendiente abrupta, o que en su defecto utilizan funciones de base radial como alternativa, el modelo propuesto arroja buenos resultados con la simpleza de su planteamiento y sin necesidad de hacer mayores adecuaciones a la estructura propuesta. Este modelo está basado en el esquema propuesto por F.J. Domínguez-Mota *et al.* en [27].

3.4.2 Estructura matricial

El ejemplo anterior emplea una discretización regular del dominio $\Omega = [0,1] \times [0,1]$ en cuadrados, con la cual se calcula una discretización para el operador en cuestión. Para los fines de este proyecto, es de interés especial analizar la estructura y propiedades básicas de las matrices involucradas en la discretización de operadores de tipo Poisson en el dominio regular Ω.

Considere el problema de Poisson $\nabla^2 u = f$ con $u \in \Omega$ y condiciones de Dirichlet homogéneas a la frontera. Así como en el ejemplo anterior, suponga que el dominio Ω se discretiza de manera regular en un mallado de m cuadrados por lado, fijando $\Delta = 1/(m+1)$. Sea $u_{k,l}$ la aproximación a la solución $u(k \cdot \Delta, l \cdot \Delta)$. De esta manera, la función $u_{k,l}$ se define sobre el mallado bidimensional correspondiente a la discretización regular del dominio Ω, en el cual se pueden tomar las diferencias en cualquiera de las direcciones del mallado, lo cual no crea ambigüedades si se especifica de manera clara la dirección en la que actúa el operador, lo cual se puede indicar con un subíndice, e.g. $\Delta_{0,x}$.

Sea $v = v(x,y)$ con $(x,y) \in \operatorname{cl}\Omega$, una función arbitraria y suficientemente diferenciable. Para cada punto interno del mallado, se tiene que

$$\begin{aligned} \left.\frac{\partial^2 v}{\partial x^2}\right]_{\substack{x=x_0+k\Delta,\\ y=y_0+l\Delta}} &= \frac{1}{\Delta^2}\Delta_{0,x}^2 v_{k,l} + \mathcal{O}(\Delta^2), \\ \left.\frac{\partial^2 v}{\partial y^2}\right]_{\substack{x=x_0+k\Delta,\\ y=y_0+l\Delta}} &= \frac{1}{\Delta^2}\Delta_{0,y}^2 v_{k,l} + \mathcal{O}(\Delta^2), \end{aligned}$$

donde $v_{k,l}$ es el valor de v en el (k,l)-ésimo punto del mallado. Por tanto,

$$\frac{1}{\Delta^2}(\Delta_{0,x}^2 + \Delta_{0,y}^2),$$

aproxima ∇^2 hasta orden $\mathcal{O}(\Delta^2)$, lo que motiva el reemplazo del operador de Laplace por su versión discretizada

$$\frac{1}{\Delta^2}(\Delta_{0,x}^2 + \Delta_{0,y}^2)u_{k,l} = f_{k,l}, \tag{3.4.5}$$

en cada par (k, l) que corresponde a la parte interna del mallado. Por supuesto, $f_{k,l} = f(x_0 + k\Delta, y_0 + l\Delta)$.

El ensamblaje del sistema, de acuerdo a [45], requiere de un acomodo de los nodos $u_{k,l}$ en un vector $\mathbf{u} \in \mathbb{R}^s$ con $s = m^2$, es decir, para cualquier permitación $\{(k_i, l_i) : i = 1, 2, \ldots, s\}$ del conjunto $\{(k, l)\}_{k,l=1,2,\ldots,s}$, sea

$$\mathbf{u} = \begin{bmatrix} u_{k_1,l_1} \\ u_{k_2,l_2} \\ . \\ . \\ . \\ u_{k_s,l_s} \end{bmatrix}.$$

Según la discretización del operador, se plantea un sistema de ecuaciones lineales de la forma

$$\mathbf{Au} = \mathbf{b}, \tag{3.4.6}$$

donde $\mathbf{A}$ es una matriz $s \times s$, y $\mathbf{b} \in \mathbb{R}^s$ es un vector que incluye las partes no homogéneas del problema $\nabla^2 u = f$, y las contribuciones de las condiciones de frontera, si es que las hay. Dado que cualquier permutación de los s nodos de la malla genera un arreglo diferente en el vector $\mathbf{u}$, se tienen $s! = (m^2)!$ maneras distintas de plantear el sistema (3.4.6). Por fortuna, ninguna de las propiedades importantes que se abordan en este análsis depende del arreglo del vector $\mathbf{u}$.

Con el propósito de facilitar el cálculo de la solución del sistema de ecuaciones lineales (3.4.6), conviene estudiar las propiedades de la matriz $\mathbf{A}$ de dicho sistema. Para tal propósito, cabe destacar el trabajo de Benvenuti y Farina [8] sobre regiones de eigenvalores para matrices positivas, cuyos resultados principales fueron implementados y puestos a prueba como parte de los propósitos del presente proyecto.

Definición 3.4.7. Dada una matriz cuadrada M, se denota por $\sigma(M)$ el *espectro* de M, es decir, el conjunto de todos los valores propios de M. Dado $\lambda \in \sigma(M)$, se define deg λ como el tamaño del mayor bloque diagonal que contiene a λ en la forma canónica de Jordan de M.
Para una matriz cuadrada y no negativa M, cualquier eigenvalor λ_0 de M tal que $|\lambda_0| = \rho(M) = \text{máx}\{|\lambda| : \lambda \in \sigma(M)\}$ se llama *eigenvalor de radio dominante* de M, y $\rho(M)$ se denomina el *radio espectral* de M.
Una matriz M es *positiva* si todas sus entradas son no negativas y al menos una de sus entradas es positiva. Una matriz M es *Metzler* si todas sus entradas fuera de la diagonal son no negativas y al menos una de estas entradas es positiva. Una matriz con

al menos una entrada negativa en la diagonal principal se llama una matriz *propiamente Metzler*.

El siguiente resultado, debido a Benvenuti y Farina [8], caracteriza completamente al eigenvalor de radio dominante de una matriz cuadrada positiva.

Teorema 3.4.8. ***(Benvenuti-Farina)*** *El eigenvalor de radio dominante de una matriz cuadrada positiva M de dimensión n son todas las raíces de $\lambda^k - \rho(M)^k = 0$ para algunos (posiblemente más de uno) valores de $k = 1, \ldots, n$. En particular, si M no es nilpotende, entonces uno de sus eigenvalores de radio dominante es real positivo (el eigenvalor de Frobenius $\lambda_F = \rho(M)$) y deg $\lambda_F \geqslant$ deg λ para cualquier otro eigenvalor de radio dominante λ.*

El siguiente resultado, debido también a Benvenuti y Farina [8], permite establecer cotas para el radio espectral de una matriz cuadrada positiva.

Teorema 3.4.9. ***(Benvenuti-Farina)*** *Sea M una matriz cuadrada positiva de dimensión n. Sean $v_r^{(i)}$ y $v_c^{(i)}$ los conjuntos de índices de los renglones y columnas que no tienen todas sus entradas nulas de una matriz $M^{(i)}$. Sea $M^{(0)} = M$ y de manera recursiva defina las matrices $M^{(i)}$ para $i \geqslant 1$ como*

$$M^{(i+1)} = [m_{kj}^{(i)}], \qquad k, j \in v_r^{(i)} \cap v_c^{(i)}$$

para $i = 0, \ldots, \overline{n}$ con $\overline{n}$ tal que $M^{(\overline{n}+1)} = M^{(\overline{n})}$ ó $v_r^{(\overline{n})} \cap v_c^{(\overline{n})} = \emptyset$. Entonces, si $v_r^{(\overline{n})} \cap v_c^{(\overline{n})} = \emptyset$, se tiene $\rho(M) = 0$; de otra manera, sean

$$r_i = \sum_{j \in v_r^{(\overline{n})} \cap v_c^{(\overline{n})}} m_{ij}^{(\overline{n})}, \qquad c_j = \sum_{i \in v_r^{(\overline{n})} \cap v_c^{(\overline{n})}} m_{ij}^{(\overline{n})}.$$

Entonces

$$\text{máx} \left\{ \min_{i \in v_r^{(\overline{n})} \cap v_c^{(\overline{n})}} \left[\frac{1}{r_i} \sum_{j \in v_r^{(\overline{n})} \cap v_c^{(\overline{n})}} m_{ij}^{(\overline{n})} r_j \right], \min_{j \in v_r^{(\overline{n})} \cap v_c^{(\overline{n})}} \left[\frac{1}{c_j} \sum_{i \in v_r^{(\overline{n})} \cap v_c^{(\overline{n})}} m_{ij}^{(\overline{n})} c_i \right] \right\} \leqslant \rho(M),$$

$$\rho(M) \leqslant \text{mín} \left\{ \max_{i \in v_r^{(\overline{n})} \cap v_c^{(\overline{n})}} \left[\frac{1}{r_i} \sum_{j \in v_r^{(\overline{n})} \cap v_c^{(\overline{n})}} m_{ij}^{(\overline{n})} r_j \right], \max_{j \in v_r^{(\overline{n})} \cap v_c^{(\overline{n})}} \left[\frac{1}{c_j} \sum_{i \in v_r^{(\overline{n})} \cap v_c^{(\overline{n})}} m_{ij}^{(\overline{n})} c_i \right] \right\}.$$

Con el propósito de analizar regiones de eigenvalores correspondientes a matrices de la forma (3.4.1), el teorema (3.4.9) se implementó como un software usando código en Matlab. A continuación se muestran algunos ejemplos.

Ejemplo 3.4.10. Considere la matriz de prueba

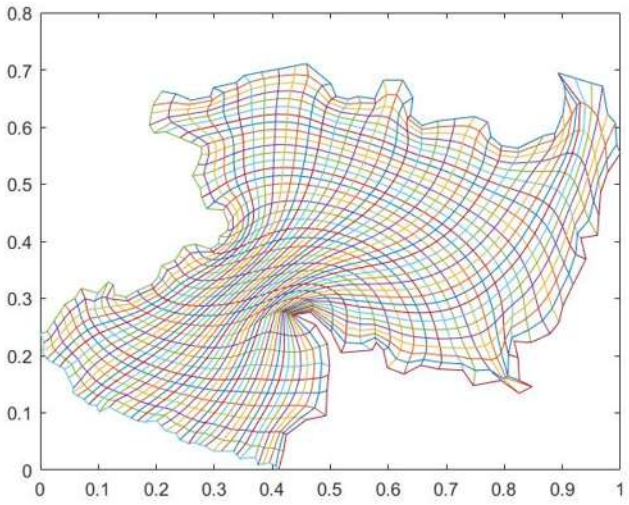
$$M = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 2 & 3 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 1 & 2 & 1 \\ 1 & 0 & 2 & 0 & 0 \end{pmatrix}$$

Nuestro software arrojó las siguientes cotas para el radio espectral de M:

```
Command Window
New to MATLAB? See resources for Getting Started.
>> [min_A,max_A]=eig_bounds(A)

min_A =

    3.2500


max_A =

    3.3333

fx >>
```

Para esta matriz M, el valor real del radio espectral es $\rho(M) = 3.3$, por lo que la aproximación dada por el software es bastante aceptable.

Ejemplo 3.4.11. En el siguiente ejemplo, se utilizó un esténcil de nueve puntos (ver [24]) para discretizar el Laplaciano y resolver la ecuación de Laplace $\nabla^2\varphi = 0$ en el dominio Ω cuya discretización se muestra en el siguiente mallado:



Se agregaron condiciones de Dirichlet a la frontera dadas por $\varphi\restriction_{\partial\Omega} = 2e^{2x+y}$. Usando el esténcil de nueve puntos, se planteó un esquema en diferencias finitas generalizadas para establecer el sistema de ecuaciones lineales (3.4.6). Se utilizó el software para obtener las cotas para el radio espectral de la matriz M asociada a este sistema, obteniendo los siguientes resultados:

```
Command Window
New to MATLAB? See resources for Getting Started.
>> [rho_min,rho_max]=eig_bounds(K1)

rho_min =

   9.4852e+03


rho_max =

   1.3358e+05

fx >>
```

Estas cotas varían por más de dos órdenes de magnitud, por lo que estas aproximaciones no aportan estimaciones finas y aplicables sobre el eigenvalor dominante.

Así pues, parece que el software desarrollado parece funcionar bien para matrices pequeñas, pero las cotas calculadas para matrices asociados a problemas de interés para este proyecto no aportan mucha información sobre el radio espectral de las matrices asociadas a estos sistemas.

Otro resultado importante de Benvenuti y Farina [8] es la caracterización de una región en el plano complejo que contiene a todos los eigenvalores de una matriz positiva. Sea Θ_n^ρ el conjunto de puntos del plano complejo que son eigenvalores de matrices positivas de tamaño $n \times n$ con radio espectral ρ. Dado que $\Theta_n^\rho = \rho\Theta_n^1$, se tiene el siguiente resultado:

Teorema 3.4.12. ***(Benvenuti-Farina)*** *La región Θ_n^1 es simétrica con respecto al eje real, está incluída en el disco $|z| \leqslant 1$ e intersecta a la circunferencia $|z| = 1$ en los puntos de la forma $e^{\frac{2\pi a}{b}}$, donde a y b son primos relativos que satisfacen $0 \leqslant a \leqslant b \leqslant n$. La frontera de Θ_n^1 consiste de estos puntos y de arcos circulares que los conectan en orden circular. Sean $e^{\frac{2\pi a_1}{b_1}}$ y $e^{\frac{2\pi a_2}{b_2}}$ $(b_1 \leqslant b_2)$ extremos de un arcos. Cada uno de estos arcos está dado por la ecuación paramétrica siguiente, donde $s \in [0, 1]$.*

$$\lambda^{b_2}(\lambda^{b_1} - s)^{[n/b_1]} = (1 - s)^{[n/b_1]}\lambda^{b_1[n/b_1]}.$$

En la sección anterior se presentó el ejemplo del esquema en diferencias finitas generalizadas para calcular una solución numérica al problema estacionario (3.4.3) usando el esténcil de siete puntos (3.5). Al final de dicho ejemplo se comentaron algunas de las complicaciones relacionadas al análisis matricial de la matriz cuadrada $\mathbf{A}$ de (3.4.6) asociada al ensamble de dicho esquema. Una manera de mejorar este enfoque consiste en añadir los dos nodos faltantes al esténcil de siete puntos y completar un esténcil de nueve puntos, como el de la figura (3.9)

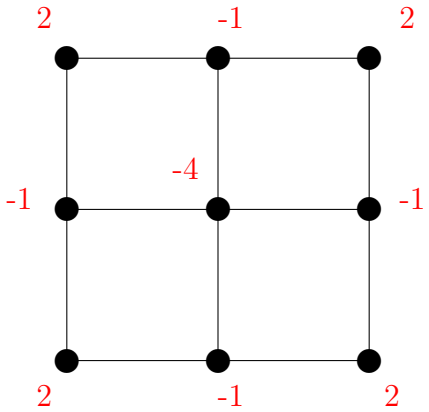


Figura 3.9: Esténcil de nueve puntos y los pesos de cada nodo.

Usando este esténcil para el problema de Poisson $\nabla^2 u = f$ en $\Omega = [0, 1]^2$, el operador de Laplace se puede discretizar de manera local como:

$$\begin{aligned} &2\left(u_{k+1,l+1}+u_{k-1,l+1}+u_{k-1,l-1}+u_{k+1,l-1}\right) \\ &-\left(u_{k+1,l}+u_{k-1,l}+u_{k,l+1}+u_{k,l-1}+4u_{k,l}\right) \quad = \quad \Delta^2 f_{k,l}, \end{aligned} \tag{3.4.13}$$

donde $f_{k,l} = f(k\Delta, l\Delta)$. Al emplear el esténcil de nueve puntos para la discretización del Laplaciano, se plantea el sistema de ecuaciones lineales (3.4.6) correspondiente para el esquema en diferencias finitas generalizadas. En este caso, se tienen los siguientes resultados relacionados a la matriz de diferenciación $\mathbf{A}$:

Lema 3.4.14. *La matriz* $\mathbf{A}$ *de (3.4.6) obtenida al usar el esténcil (3.9) para la discretización del operador de Laplace, es una matriz simétrica y el conjunto de sus valores propios es*

$$\sigma(\mathbf{A}) = \{\lambda_{\alpha,\beta} : \alpha, \beta = 1, 2, \ldots, m\},$$

donde

$$\lambda_{\alpha,\beta} = -4\left\{\cos^2\left[\frac{\alpha\pi}{2(m+1)}\right] + \cos^2\left[\frac{\beta\pi}{2(m+1)}\right] - 2\cos\left[\frac{\alpha\pi}{(m+1)}\right]\cdot\cos\left[\frac{\beta\pi}{(m+1)}\right]\right\}, \tag{3.4.15}$$

para $\alpha, \beta = 1, 2, \ldots, m$.

Demostración. La simetría en los pesos del esténcil (3.9) hace que la matriz $\mathbf{A}$ herede una estructura simétrica.
Para calcular los valores propios de la matriz $\mathbf{A}$, se pasa por alto el arreglo de los nodos de la malla, después de todo, las permutaciones simétricas conservan los valores propios. Suponga que se puede demostrar la existencia de una función no trivial $(v_{k,l})_{k,l=0,1,\ldots,m+1}$ tal que $v_{k,0} = v_{k,m+1} = v_{0,l} = v_{m+1,l} = 0$ para $k, l = 1, 2, \ldots, m$, y de tal manera que el sistema de ecuaciones lineales

$$\begin{aligned} &2\left(v_{k+1,l+1}+v_{k-1,l+1}+v_{k-1,l-1}+v_{k+1,l-1}\right) \\ &-\left(v_{k+1,l}+v_{k-1,l}+v_{k,l+1}+v_{k,l-1}+4v_{k,l}\right) \quad = \quad \lambda v_{k,l}, \end{aligned} \tag{3.4.16}$$

con $k, l = 1, 2, \ldots, m$, se cumple para algún λ; en cuyo caso, salvo reacomodo, se tendría que $v_{k,l}$ es un vector propio asociado al valor propio λ de la matriz $\mathbf{A}$.
Dados $\alpha, \beta \in \{1, 2, \ldots, m\}$ sea

$$v_{k,l} = \sin\left(\frac{k\alpha\pi}{m+1}\right)\sin\left(\frac{l\beta\pi}{m+1}\right), \qquad k, l = 0, 1, \ldots, m+1.$$

Observe que se cumple la condición de frontera de Dirichlet homogénea $v_{k,0} = v_{k,m+1} =$

$v_{0,l} = v_{m+1,l} = 0$ para $k, l = 1, 2, \ldots, m$. Sustituyendo en (3.4.16),

$$\begin{aligned}
2\left(v_{k+1,l+1} + v_{k-1,l+1} + v_{k-1,l-1} + v_{k+1,l-1}\right) - \left(v_{k+1,l} + v_{k-1,l} + v_{k,l+1} + v_{k,l-1} + 4v_{k,l}\right) \;=\; & \\
2\left\{\sin\left(\frac{(k+1)\alpha\pi}{m+1}\right)\left[\sin\left(\frac{(l+1)\beta\pi}{m+1}\right) + \sin\left(\frac{(l-1)\beta\pi}{m+1}\right)\right]\right. & \\
\left. + \sin\left(\frac{(k-1)\alpha\pi}{m+1}\right)\left[\sin\left(\frac{(l+1)\beta\pi}{m+1}\right) + \sin\left(\frac{(l-1)\beta\pi}{m+1}\right)\right]\right\} & \\
- \left\{\left[\sin\left(\frac{(k+1)\alpha\pi}{m+1}\right) + \sin\left(\frac{(k-1)\alpha\pi}{m+1}\right)\right]\sin\left(\frac{l\beta\pi}{m+1}\right)\right. & \\
\left. + \sin\left(\frac{k\alpha\pi}{m+1}\right)\left[\sin\left(\frac{(l+1)\beta\pi}{m+1}\right) + \sin\left(\frac{(l-1)\beta\pi}{m+1}\right)\right] + 4\sin\left(\frac{k\alpha\pi}{m+1}\right)\sin\left(\frac{l\beta\pi}{m+1}\right)\right\} &
\end{aligned}$$

Para simplificar, se usa la identidad trigonométrica

$$\sin(\alpha - \beta) + \sin(\alpha + \beta) = 2\sin\alpha\cos\beta$$

entonces se tiene, para el lado derecho

$$\begin{aligned}
= 2\left\{2\sin\left(\frac{(k+1)\alpha\pi}{m+1}\right)\sin\left(\frac{l\beta\pi}{m+1}\right)\cos\left(\frac{\beta\pi}{m+1}\right)\right. & \\
\left. + 2\sin\left(\frac{(k-1)\alpha\pi}{m+1}\right)\sin\left(\frac{l\beta\pi}{m+1}\right)\cos\left(\frac{\beta\pi}{m+1}\right)\right\} & \\
- \left\{2\sin\left(\frac{k\alpha\pi}{m+1}\right)\cos\left(\frac{\alpha\pi}{m+1}\right)\sin\left(\frac{l\beta\pi}{m+1}\right)\right. & \\
\left. + 2\sin\left(\frac{k\alpha\pi}{m+1}\right)\sin\left(\frac{l\beta\pi}{m+1}\right)\cos\left(\frac{\beta\pi}{m+1}\right) + 4\sin\left(\frac{k\alpha\pi}{m+1}\right)\sin\left(\frac{l\beta\pi}{m+1}\right)\right\} & \\
= -4\left\{\cos^2\left[\frac{\alpha\pi}{2(m+1)}\right] + \cos^2\left[\frac{\beta\pi}{2(m+1)}\right] - 2\cos\left[\frac{\alpha\pi}{(m+1)}\right]\cdot\cos\left[\frac{\beta\pi}{(m+1)}\right]\right\} v_{k,l} &
\end{aligned}$$

Para $k, l = 1, 2, \ldots m,$. En el último paso se usó la identidad trigonométrica

$$1 + \cos\theta = 2\cos^2\left(\frac{\theta}{2}\right).$$

Se tiene entonces que (3.4.16) se cumple para $\lambda = \lambda_{\alpha,\beta}$, lo que concluye la demostración del lema. ■

Corolario 3.4.17. *La matriz* $\mathbf{A}$ *es negativa definida y, a fortiori, no singular.*

Demostración. Por el lema anterior, la matriz $\mathbf{A}$ es simétrica. De la expresión para $\lambda_{\alpha,\beta}$, no es claro que el valor del eigenvalor es negativo para cualesquiera $\alpha, \beta = 1, 2, \ldots, m$. En la figura (3.10) se muestra la gráfica correspondiente a $-\frac{1}{4}\lambda_{\alpha,\beta}$, en donde se añade el coefiente $-\frac{1}{4}$ para simplificar el coeficiente -4 de (3.4.15). Como se puede ver en la gráfica (3.10), todos los valores de $-\frac{1}{4}\lambda_{\alpha,\beta}$ son positivos, por lo que al mutiplicar con el coeficiente -4 de (3.4.15), se tiene que todos los valores propios de $\mathbf{A}$ son negativos, y por lo tanto $\mathbf{A}$ es negativa definida y no singular. ■

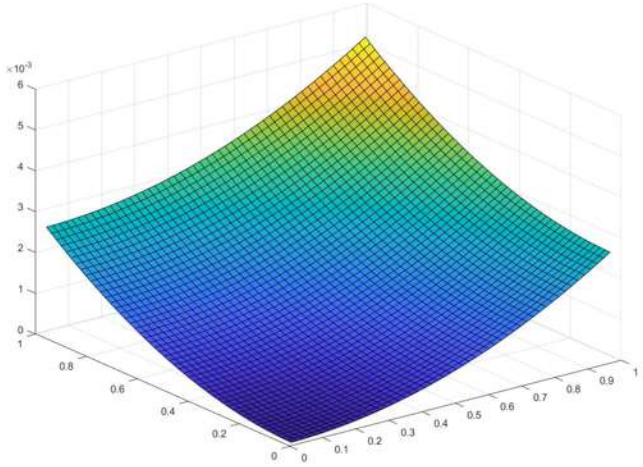


Figura 3.10: Gráfica de $\lambda_{\alpha,\beta}$ en el dominio Ω.

Valores propios del operador de Laplace

Los valores y vetores propios del operador de Laplace se pueden evaluar de manera directa en el dominio $\Omega = [0,1]^2$. Dados dos enteros positivos α, β, sea $v(x,y) = \sin(\alpha\pi x)\sin(\beta\pi x)$, con $x, y \in [0,1]$. Nótese que v cumple las condiciones de frontera de Dirichlet homogéneas, más aún, $\nabla^2 v = -(\alpha^2+\beta^2)\pi^2 v$, y por tanto v es una función propia asociada al valor propio $-(\alpha^2+\beta^2)\pi^2$. Se puede demostrar que *todas* las funciones propias del operador de Laplace en $\Omega = [0,1]^2$ tienen esta forma. El vector $v_{k,l}$ del Lema (3.4.14) se puede obtener al probar la función propia v en los puntos del mallado $\left\{\left(\frac{k}{m+1}, \frac{l}{k+1}\right)\right\}_{k,l=0,1,\ldots,m+1}$, para $\alpha, \beta = 1, 2, \ldots, m$, mientras que $\Delta^{-2}\lambda_{\alpha,\beta}$ es una buena aproximación a $-(\alpha^2+\beta^2)\pi^2$ si α, β son pequeños en comparación con m. Expandiendo $\cos^2\theta$ en series de potencias y teniendo en cuenta que $(m+1)\Delta = 1$, se obtiene

$$\begin{aligned}
\frac{\lambda_{\alpha,\beta}}{\Delta^2} &= -4\left\{\left(1 - \left[\frac{\alpha\pi}{2(m+1)}\right]^2 + \frac{1}{3}\left[\frac{\alpha\pi}{2(m+1)}\right]^4 + \ldots\right)\right. \\
&\quad + \left(1 - \left[\frac{\beta\pi}{2(m+1)}\right]^2 + \frac{1}{3}\left[\frac{\beta\pi}{2(m+1)}\right]^4 + \ldots\right) \\
&\quad -2\left(1 - \left[\frac{\alpha\pi}{(m+1)}\right]^2 + \frac{1}{24}\left[\frac{\alpha\pi}{(m+1)}\right]^4 + \ldots\right) \times \\
&\quad \left.\times\left(1 - \left[\frac{\beta\pi}{(m+1)}\right]^2 + \frac{1}{24}\left[\frac{\beta\pi}{(m+1)}\right]^4 + \ldots\right)\right\} \\
&= -\frac{7}{2}(\alpha^2+\beta^2)\pi^2 + \frac{1}{4}(\alpha^4+\beta^4)\pi^4(\Delta^2) + \mathcal{O}(\Delta^4).
\end{aligned}$$

Entonces, salvo un múltiplo escalar $-7/2$, el valor propuesto $\lambda_{\alpha,\beta}$ aproxima a los

valores propios exactos del operador de Laplace.

Sean u la solución exacta al problema de Poisson $\nabla^2 u = f$ en el dominio $\Omega = [0,1]^2$, $\tilde{u}_{k,l} = u(k\Delta, l\Delta)$ y denotemos por $e_{k,l}$ el error de la discretización (3.4.5) con el esténcil de nueve puntos (3.9) en el (k,l)-ésimo punto del mallado, $e_{k,l} = u_{k,l} - \tilde{u}_{k,l}$, para $k,l = 0,1,\ldots,m+1$. Estas ecuaciones lineales se representan de la forma (3.4.6) y $\mathbf{e}$ denota un arreglo de $\{e_{k,l}\}$ en un vector en $\mathbb{R}^s$, $s = m^2$, cuyo orden coincide con el de $\mathbf{u}$. Se medirá la magnitud de $\mathbf{e}$ con la norma Euclideana $||\cdot||$.

Teorema 3.4.18. *Sujeto a que la función f sea lo suficientemente diferenciable, además de las condiciones de frontera, existe un número $c > 0$, independiente de Δ, tal que*

$$||\mathbf{e}|| \leqslant c \cdot \Delta^2, \qquad \Delta \to 0. \tag{3.4.19}$$

Demostración. Dado que $\Delta^{-2}(\Delta^2_{0,x} + \Delta^2_{0,y})$ aproxima al operador de Laplace hasta orden $\mathcal{O}(\Delta^2)$, se cumple que

$$\begin{aligned} &2\left(\tilde{u}_{k+1,l+1} + \tilde{u}_{k-1,l+1} + \tilde{u}_{k-1,l-1} + \tilde{u}_{k+1,l-1}\right) \\ &-\left(\tilde{u}_{k+1,l} + \tilde{u}_{k-1,l} + \tilde{u}_{k,l+1} + \tilde{u}_{k,l-1} + 4\tilde{u}_{k,l}\right) \;=\; \Delta^2 f_{k,l} + \mathcal{O}(\Delta^4), \end{aligned} \tag{3.4.20}$$

para $\Delta \to 0$. Restando (3.4.20) de (3.4.13), se tiene

$$\begin{aligned} &2\left(e_{k+1,l+1} + e_{k-1,l+1} + e_{k-1,l-1} + e_{k+1,l-1}\right) \\ &-\left(e_{k+1,l} + e_{k-1,l} + e_{k,l+1} + e_{k,l-1} + 4e_{k,l}\right) \;=\; \mathcal{O}(\Delta^4), \qquad \Delta \to 0, \end{aligned}$$

o, en notación vectorial, y poniendo la debida atención al hecho que $u_{k,l}$ y $\tilde{u}_{k,l}$ coinciden a lo largo de la frontera

$$\mathbf{A}\mathbf{e} = \delta_\Delta, \tag{3.4.21}$$

donde $\delta_\Delta \in \mathbb{R}^{m^2}$ es tal que $||\delta_\Delta|| = \mathcal{O}(\Delta^4)$. De (3.4.21) se sigue que

$$\mathbf{e} = \mathbf{A}^{-1}\delta_\Delta. \tag{3.4.22}$$

Por el Lema (3.4.14), $\mathbf{A}$ es simétrica, y por tanto también $\mathbf{A}^{-1}$ es simétrica, y su norma Euclideana $||\mathbf{A}^{-1}||$ es igual a su radio espectral $\rho(\mathbf{A}^{-1})$. Éste último se puede calcular con (3.4.15), dado que $\lambda \in \sigma(B)$ es lo mismo que $\lambda^{-1} \in \sigma(B^{-1})$ para cualquier matriz no singular B. Entonces, tomando en cuenta que $(m+1)\Delta = 1$,

$$\begin{aligned} \rho(\mathbf{A}^{-1}) \;&=\; \max_{\alpha,\beta=1,2,\ldots,m} \frac{1}{4}\left\{\cos^2\left[\frac{\alpha\pi}{2(m+1)}\right] + \cos^2\left[\frac{\beta\pi}{2(m+1)}\right]\right. \\ &\qquad \left. -2\cos\left[\frac{\alpha\pi}{(m+1)}\right]\cdot\cos\left[\frac{\beta\pi}{(m+1)}\right]\right\}^{-1} \\ &=\; \frac{1}{8\cos^2\left(\frac{1}{2}\pi\Delta\right) - 8\cos^2(\pi\Delta)}. \end{aligned}$$

Dado que

$$\lim_{\Delta\to 0}\left[\frac{\Delta^2}{8\cos^2\left(\frac{1}{2}\pi\Delta\right) - 8\cos^2(\pi\Delta)}\right] = \frac{1}{6\pi^2},$$

se sigue que para cualquier constante $c_1 > (6\pi^2)^{-1}$ se cumple que

$$||\mathbf{A}^{-1}|| = \rho(\mathbf{A}^{-1}) \leqslant c_1 \Delta^{-2}, \qquad \Delta \to 0. \tag{3.4.23}$$

Si tanto f como las condiciones de frontera son lo suficientemente diferenciables, u es por sí misma suficiente diferenciable y existe una constante $c_2 > 0$ tal que

$$||\delta \cdot \Delta|| \leqslant c_2 \cdot \Delta^4$$

(recuerde que δ depende únicamente de la solución exacta). Sustituyendo esta expresión junto con (3.4.23) en la desigualdad (3.4.22), se cumple (3.4.19) con $c = c_1 c_2$. ■

Para concluir este capítulo, conviene resaltar la propiedad de adaptabilidad de las discretizaciones que se han empleado en este proyecto, pues así como en el ejemplo (3.4.1) o en [27], el uso de estas mallas se puede extender a mallas no estructuradas o nubes de puntos. Estas discretizaciones se probaron con mallados de 3×3 nodos, los cuales fueros dispuestos al azar. En comparación con la estructura del esténcil (3.9), se generó una estructura aleatoria donde la numeración de los nodos se alternó; en particular, la posición del nodo central se fue alternando entre los puntos del esténcil aleatorio, y se registró el error local en cada uno:

- En la figura (3.11) se muestra el esténcil generado aleatoriamente, donde la posición del nodo central se alterna entre todos los nodos del esténcil. Se señalan los esnténciles donde se obtuvieron los errores locales máximo y mínimo.
- La figura (3.12) muestra las gráficas correspondientes a los errores locales de cada esténcil de (3.11). Todos los errores calculados son del órden de 10^{-14}.
- En la figura (3.13) se muestra el esténcil generado aleatoriamente, indicando las magnitudes máxima y mínima de los errores locales, y la posición del nodo central correspondiente a cada uno.

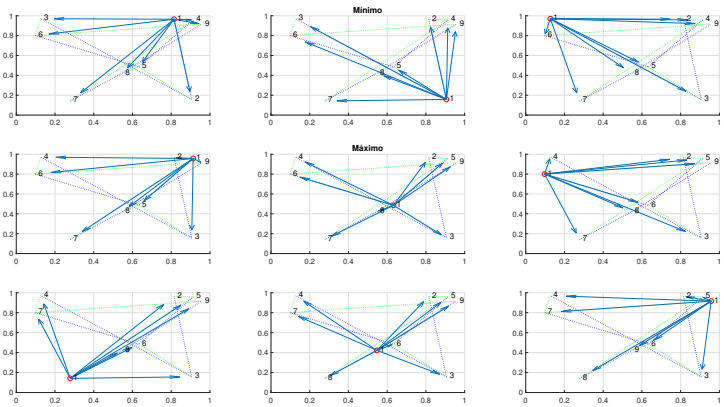


Figura 3.11: Esténcil aleatorio de nueve puntos; posición del nodo central.

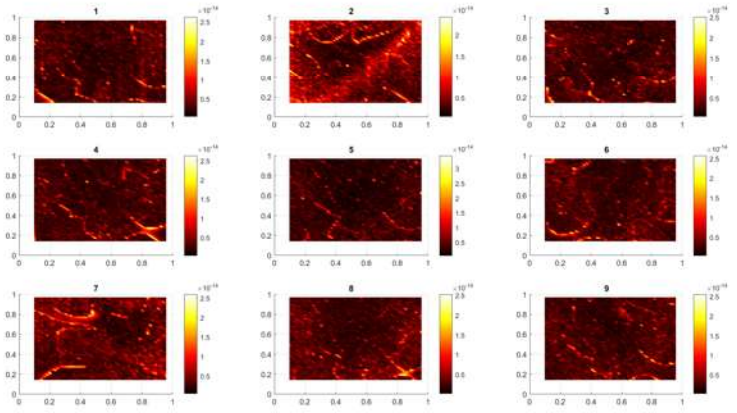


Figura 3.12: Esténcil aleatorio de nueve puntos; errores locales en cada esténcil.

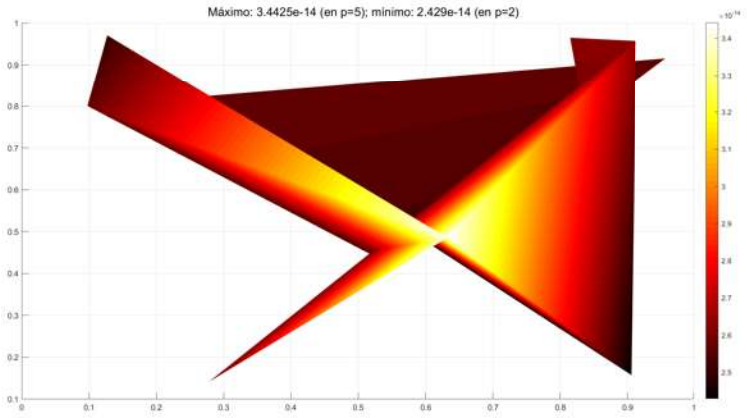


Figura 3.13: Esténcil aleatorio de nueve puntos; errores locales máximos y mínimos.

Capítulo 4

Modelado en problemas inversos

Las discretizaciones que se presentaron en el capítulo pasado se usaron también en el modelado de problemas inversos asociados al ejemplo estacionario de difusión-advección en 2D que se discutió anteriormente:

$$u\frac{\partial C}{\partial x} + v\frac{\partial C}{\partial y} = \frac{\partial}{\partial x}\left(A\frac{\partial C}{\partial x}\right) + \frac{\partial}{\partial y}\left(A\frac{\partial C}{\partial y}\right), \tag{4.0.1}$$

para C definida sobre el dominio $\Omega = [0,1] \times [0,1]$, con los parámetros constantes $u = v = 0.1$, $A(y) = 1 + e^{-10y}$ para $0 \leqslant y \leqslant 1$, y las condiciones de frontera $C(0,y) = C(1,y) = 0.05$ para $0 \leqslant y \leqslant 1$, $\frac{\partial C}{\partial n} = 0$ en $(0,1) \times \{1\}$ y $\frac{\partial C}{\partial n} = g(x)$ en $(0,1) \times \{0\}$, donde

$$g(x) = \begin{cases} 0.5 & \frac{3}{8} \leqslant x \leqslant \frac{5}{8} \\ 0 & \text{de otro modo} \end{cases}$$

En el contexto de los problemas inversos, el problema que se abordó en este proyecto se refiere de manera específica a la identificación de parámetros: considere la solución al problema (4.0.1), suponga que dicha solución representa una medición real de la concentración de una cierta sustancia suspendida en la atmósfera. El problema inverso que se discute en este capítulo consiste en la determinación de parámetros físicos involucrados en este fenómeno de disfusión-advección; específicamente, este problema trata de determinar los parámetros u y v de (4.0.1), los cuales corresponden a las velocidades de transporte.

Cabe mencionar que esta clase de problemas en particular requieren de especial atención en el tratamiento de las condiciones a la frontera. En particular, la solución numérica que se obtuvo para (4.0.1) en el capítulo anterior presenta un crecimiento abrupto en el gradiente en la frontera inferior. Sin embago, la discretización empleada permitió obtener buenos resultados. Este tipo de crecimientos del gradiente en la frontera fueron modelados con éxito, como se muestra en [27].

Para el modelado de estos parámetros, Dilley y Yen [23] propusieron las siguientes funciones, dadas como leyes de potencias:

$$u(x,y) = (U_1 - ax)\left(\frac{y}{y_1}\right)^m, \qquad v(y) = \frac{ay}{m+1}\left(\frac{y}{y_1}\right)^m, \tag{4.0.2}$$

donde U_1, a, y_1, m son parámetros a determinar, los cuales están estrechamente relacionados a la física del problema.

Uno de los propósitos de este proyecto es proponer modelos para los parámetros u y v de (4.0.1), que arrojen buenos resultados y que a la vez, resulten más sencillos que los propuestos en [23].

4.1 El modelo exponencial

El primer modelo consiste en funciones de tipo exponencial, dadas como series de potencias de la siguiente manera:

$$u(x) = A_1 x^{\alpha_1} + B_1, \qquad v(y) = A_2 y^{\alpha_2} + B_2, \tag{4.1.1}$$

donde los parámetros $A_1, A_2, B_1, B_2, \alpha_1$ y α_2 son constantes a determinar. En una primera comparación, nótese que, a diferencia de las leyes de potencias (4.0.2), estas funciones dependen cada una de una única componente espacial.

El cálculo de estos parámetros se realizó tomando los datos de la solución numérica obtenida para (4.0.1) en el capítulo anterior, realizando un ajuste de mínimos cuadrados no lineales regularizado, usando el algoritmo de región de confianza, como los que se discutieron en el Capítulo 2. Los valores encontrados para estos parámetros se muestran a continuación:

A_1	413.86	A_2	41.38
B_1	0.03	B_2	0.03
α_1	413.86	α_2	0.03

Cuadro 4.1: Valores encontrados para los parámetros del modelo (4.1.1).

Una vez obtenidos estos parámetros, sus valores se usaron para graficar la función de concentración que se deseaba recuperar, la cual corresponde a la solución numérica de (4.0.2) que se obtuvo en el capítulo anterior. Ambas funciones se muestran en la figura (4.1).

Como se puede ver en la figura (4.1), las soluciones son bastante similares, no solo en apariencia, sino también en norma. Se calculó la diferencia entre estas soluciones con relación a las normas $||\cdot||_\infty$ y $||\cdot||_2$; las diferencias obtenidas se muestran en la tabla (4.2).

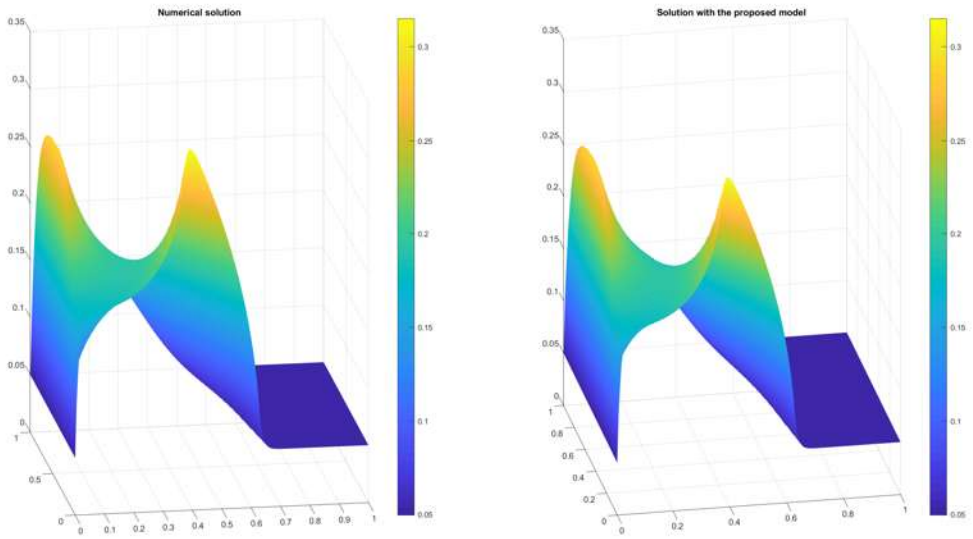


Figura 4.1: *Izquierda:* solución numérica de (4.0.2). *Derecha:* modelo obtenido usando las leyes de potencias (4.1.1).

Norma $\|\| \cdot \|\|_\infty$:	1.608×10^{-10}
Norma $\|\| \cdot \|\|_2$:	4.6×10^{-9}

Cuadro 4.2: Norma de la diferencia entre las funciones de la figura (4.1).

4.2 El modelo racional

El segundo modelo propuesto para los parámetros u y v de (4.0.1) consiste en funciones de tipo racional, dadas como cocientes de polinomios lineales. Los modelos propuestos son los siguientes:

$$u(x) = \frac{A_1 x + A_2}{A_3 x + A_4}, \qquad v(y) = \frac{B_1 x + B_2}{B_3 x + B_4}, \tag{4.2.1}$$

donde los coeficientes A_i, B_i para $i = 1, 2, 3, 4$ son constantes a determinar. De manera similar al modelo (4.1.1), estas funciones dependen cada una de una sola componente espacial.

Los parámetros constantes A_i, B_i para $i = 1, 2, 3, 4$ se calcularon tomando los datos de la solución numérica obtenida para (4.0.1) en el capítulo anterior, realizando un ajuste de mínimos cuadrados no lineales regularizado, usando el algoritmo de región de confianza. Los valores encontrados para estos parámetros se muestran en la tabla (4.3)

Los valores obtenidos para estos parámetros se usaron para graficar la función de concentración que se deseaba restaurar, y que corresponde a la solución numérica de (4.0.1) que se obtuvo con el esquema en diferencias finitas generalizadas que se describió en el capítulo anterior. Estas funciones se muestran en la figura (4.2).

A_1	0.0499	B_1	0.0501
A_2	0.0499	B_2	0.0501
A_3	0.0501	B_3	0.0499
A_4	0.0501	B_4	0.0499

Cuadro 4.3: Valores encontrados para los parámetros del modelo (4.2.1).

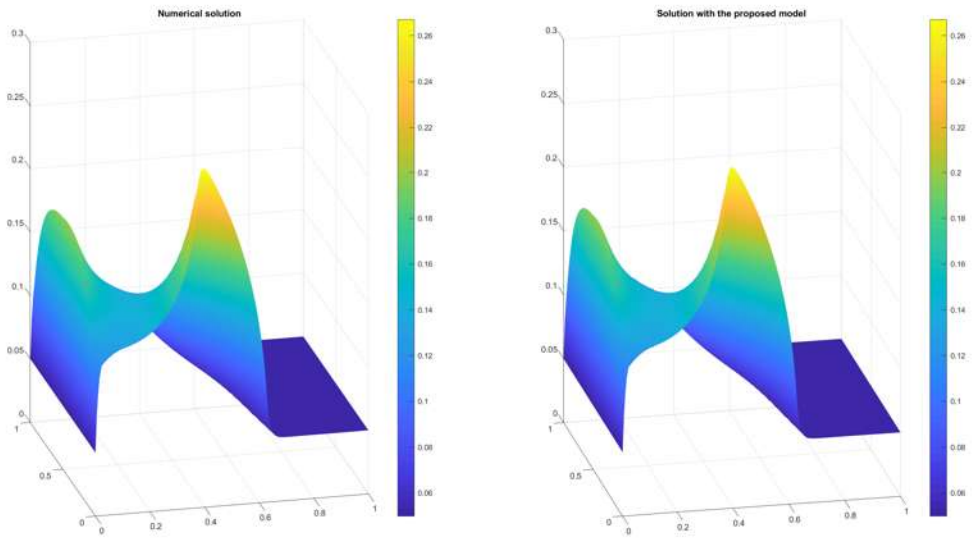


Figura 4.2: *Izquierda:* solución numérica de (4.0.2). *Derecha:* modelo obtenido usando las funciones racionales (4.2.1).

De manera similar al modelo exponencial, estas funciones resultan también bastante similares. La diferencia entre estas dos funciones, con respecto a las normas $||\cdot||_\infty$ y $||\cdot||_2$ se muestran a continuación:

Norma $||\cdot||_\infty$:	6.57×10^{-11}
Norma $||\cdot||_2$:	3.21×10^{-9}

Cuadro 4.4: Norma de la diferencia entre las funciones de la figura (4.2).

4.3 Conclusiones

El esquema en diferencias finitas generalizadas que se presentó en el capítulo anterior permitió obtener una solución numérica aceptable al problema (4.0.1), esto a pesar de las condiciones a la frontera del problema. Si bien es cierto que las discretizaciones son muy similares a las que se emplean en esquemas en diferencias finitas clásicas, los

nodos adicionales que se añadieron en las fronteras con condiciones de Neumann permitieron obtener una solución numérica bastante precisa. En las figuras (4.1) y (4.2) se puede observar que los efectos del mal condicionamiento del problema siguen estando presentes, pero tales efectos son parte de la formulación del problema (4.0.1) y son independientes del esquema que se use para calcular una solución numérica.

También es importante resaltar que los modelos propuestos (4.1.1) y (4.2.1) permitieron replicar con un buen grado de precisión la solución numérica de (4.0.1) que se obtuvo en el Capítulo anterior. Si bien es muy cierto que los modelos propuestos por Dilley y Yen en [23] están mucho más relacionados con la física del problema, nuestra propuesta permitió replicar la solución numérica de (4.0.1) con un buen grado de precisión y empleando funciones con un menor grado de complejidad, lo cual es una característica bastante deseable en el cálculo de soluciones numéricas.

4.4 Futuros proyectos

En el futuro se espera continuar trabajando con los modelos en diferencias finitas generalizadas presentados en este proyecto. El siguiente proyecto consiste en usar los esquemas propuestos para estudiar fenónemos de difusión-advección no estacionarios, así como adaptar la propuesta de este esquema para dominios irregulares, en los cuales se pueda adaptar la simetría del esténcil a geometrías irregulares.

Los resultados obtenidos del problema inverso presentado en este proyecto no representa más que un *benchmark* de los esquemas que se proponen. Un proyecto que se tiene pensado a corto plazo consiste en emplear los esquemas propuestos usando datos reales obtenidos de una estación meteorológica en la ciudad de Morelia, Michoacán, y obtener modelos numéricos para condiciones climáticas reales en esta entidad.

Los esquemas propuestos en este proyecto destacan por su sencillez y los buenos resultados obtenidos. La inclusión de los nodos adicionales en las fronteras con condiciones de Neumann permitieron obtener una solución numérica aceptable a pesar del crecimiento abrupto del gradiente en una de las fronteras. Otro proyecto que podría resultar interesante es adaptar los esquemas propuestos para estudiar fenómenos de dispersión entre dos interfaces distintas. Estos problemas suelen presentar condiciones de frontera en las que se tienen discontinuidades o crecimientos bastante abruptos del gradiente en la frontera entre las interfaces. Creemos que sería interesante poner a prueba nuestra propuesta en uno de estos problemas.

Otra línea que es de interés para este proyecto es la adaptación de los esquemas propuestos en mallados adaptativos. Creemos que la simetría heredada por el esténcil en la estructura de la matriz de diferenciación tiene propiedades que pueden ser preservadas para mallas adaptativas.

Bibliografía

[1] M. Abouali and J.E. Castillo. Solving adjective equations using Castillo-Grone's Mimetic Operators.

[2] M. Abouali and J.E. Castillo. Stability and performance analysis of the Castillo-Grone mimetic operators in conjunction with RK3 time discretization in solving advective equations. *Procedia Computer Science*, 18:465–472, 2013.

[3] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, volume 55 of *Applied Mathematics*. National Bureau of Standards, first edition, 1964.

[4] M. Agarwal and A. Tandon. Modeling of the urban heat island in the form of mesoscale wind and of its effect on air pollution dispersal. *Applied Mathematical Modelling*, (34):2520–2530, 2010. 3

[5] S. Bartels. *Numerical Approxiamtion of Partial Differential Equations*, volume 64 of *Texts in Applied Mathematics*. Springer, first edition, 2016.

[6] S. Bartels. *Numerical Methods for Nonlinear Partial Differential Equations*. Number 47 in Springer Series in Computational Mathematics. Springer, first edition, 2017.

[7] L. Beilina, E. Karchevskii, and M. Karchevskii. *Numerical Linear Algebra: Theory and Applications*. Springer, first edition, 2017.

[8] L. Benvenuti and L. Farina. Eigenvalue regions for positive systems. *Systems & Control Letters*, (51):325–330, 2004. 27, 68, 69, 71

[9] A. Berman and R.J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Number 9 in Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1994. 27

[10] J.F. Bonnans, C.L. Gilbert, and C.A. Sagastizábal. *Numerical Optimization. Theoretical and Practical Aspects*. Springer, second edition, 2006.

[11] G.R. Buchanan. *Finite Element Analysis*. Schaum's Outlines Series. McGraw-Hill, first edition, 1995.

[12] D. Buske, M.T. Vilhena, T. Tirabasi, and B. Bodmann. Air Pollution Steady-State Advection-Diffusoin Equation: The General Three-Dimensional Solution. *Journal of Environmental Protcetion*, 3:1124–1134, 2012. 2

[13] V.E. Cardoso-Nungaray. *Discrete Volume Method. A variational approach for brittle fracture*. PhD thesis, Centro de Invetigación en Matemáticas, 2017.

[14] J.E. Castillo, J.M. Hyman, M.J. Shashkov, and S. Steinberg. The sensitivity and accuracy of fourth order finite-difference schemes on nonuniform grids in one dimension. *Computers Math. Applic.*, 30(8):41–55, 1995.

[15] J.E. Castillo, J.M. Hyman, M.J. Shashkov, and S. Steinberg. Fourth- and sixth-order onservative finite difference approximations of the divergence and gradient. *Applied Numerical Mathematics*, 37:171–187, 2001.

[16] M. Celia and W. Gray. *Numerical Methods for Differential Equations*. Prentice-Hall, 1992. 61

[17] HF Chan, CM Fan, and CW Kuo. Generalized finite difference method for solving two-dimensional non-linear obstacle problems. *Engineering Analysis with Bouondary Elements*, 37:1189–1196, 2013.

[18] G. Chavent. *Nonlinear Least Squares for Inverse Problems*. Springer, 2009. 22

[19] C. Chávez-Negrete, F.J. Domínguez-Mota, and D. Santana-Quinteros. Numerical solution of Richards' equation of water flow by generalized finite differences. *Computers and Geotechnics*, (101):168–175, 2018. 3

[20] J. Crank. *The Mathematics of Diffusion*. Claredon Press - Oxford, second edition, 1975. 13

[21] L.B. da Veiga, A. Chernov, L. Mascotto, and A. Russo. Basic principles of hp virtual elements on quasiuniform meshes. *Mathematical Models and Methods in Applied Sciences*, 26(8):1567–1598, 2016.

[22] L. Dieci and J. Lorenz. Block M-matrices and Computation of Invariant Tori. *SIAM J. Sci. Stat. Comput.*, 13(4):885–903, 1992.

[23] J.F. Dilley and K.T. Yen. Effect of a mesoscale type wind on the pollutant distribution from a line source. *Atmospheric Environment Pergamon Press*, 5:843–851, 1971. 78, 79, 82

[24] F.J. Domínguez-Mota, S. Mendoza-Armenta, G. Tinoco-Guerrero, and J.G. Tinoco-Ruiz. Finite difference schemes satisfying an optimality condition for the unsteady heat equation. *Mathematics and computers in simulation*, (106):76–83, 2014. 3, 70

[25] F.J. Domínguez-Mota, G. Tinoco-Guerrero, A. Gaona-Arias, M.L. Ruiz-Zavala, and J.G. Tinoco-Ruiz. A stability analysis for a generalized finite-difference scheme applied to the pure advection equation. *Mathematics and Computers in Simulation*, 147:293–300, 2018. 3

[26] F.J. Domínguez-Mota, G. Tinoco-Guerrero, and J.G. Tinoco-Ruiz. A study of the stability for a generalized finite-difference scheme applied to the advection-diffusion equation. *Mathematics and Computers in Simulation*, 176:301–311, 2020. 3

[27] F.J. Domínguez-Mota, J.G. Tinoco-Ruiz, C. Chávez, J.S. Lucas-Martínez, and D. Sanatana-Quinteros. Generalized finite difference solution for the Motz Problem. *Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería*, 37(1), 2021. 3, 67, 76, 78

[28] D.R. Durran. *Numerical Methods for Fluid Dynamics*. Number 32 in Texts in Applied Mathematics. Springer, second edition, 2010. 13, 15, 17

[29] H.W. Engl. *Inverse Problems*. Number 8. Sociedad Matemática Mexicana, 1995.

[30] M. Esmaeilzadeh, R.M. Barron, and R. Balachandar. Numerical solution of partial differential equations in arbitrary shaped domains using cartesian cut-stencil finite difference method. Part I: Concepts and Fundamentals. *Numer. Math. Theor. Meth. Appl.*, 13(4):881–907, 2020.

[31] F.R. Gantmacher. *The Theory of Matrices*, volume 1. Chelsea Publishing Company, first edition, 1959.

[32] M.S. Gockenbach. *Understanding and Implementing the Finite Element Method*. Society for Industrial and Applied Mathematics, first edition, 2006.

[33] S.K. Godunov and V.S. Ryabenkii. *Difference Schemes*, volume 19 of *Studies in Mathematics and its Applications*. Elsevier Science Publishers B.V., first edition, 1987.

[34] J. Gottlieb and P. DuChateau. *Parameter identification and inverse problems in hydrology, geology and ecology*, volume 23 of *Water Science and Technology Library*. Kluwer Academic Publishers, first edition, 1996.

[35] D.S. Grebenkov and B.T. Nguyen. Geometrical Structure of Laplacian Eigenfunctions. *Society for Industrial and Applied Mathematics*, 55(4):601–667, 2013.

[36] C. Grossmann, H.G. Roos, and M. Stynes. *Numerical Treatment of Partial Differential Equations*. Springer, first edition, 2007.

[37] Y Gu, J Lei, CM Fan, and XQ He. The generalized finite difference method for an inverse time-dependent source problem associated with three-dimensional heat equation. *Engineering Analysis with Bouondary Elements*, 91:73–81, 2018. 3

[38] A.H. Hasanoğlu and V.G. Romanov. *Introduction to Inverse Problems for Differential Equations*. Springer, first edition, 2010.

[39] C. Hirsch. *Numerical Computation of Internal and External Flows*, volume 1 of *Fundamentals of Numerical Discretization*. John Wiley & Sons Ltd., first edition, 1988.

[40] C. Hirsch. *Numerical Computation of Internal and External Flows*, volume 2 of *Computational Methods for Inviscid and Viscous Flows*. John Wiley & Sons Ltd., first edition, 1990.

[41] E. Holzbecher. *Environmental Modeling using Matlab*. Springer, second edition, 2012. 13, 15, 17

[42] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, second edition, 2013.

[43] C.H. Huang. A theory of dispersion in turbulent shear flow. *Atmospheric Environment*, 13:453–463, 1979.

[44] TZ Huang and Y Zhu. Estimation of $||A^{-1}||_\infty$ for weakly chained diagonally dominant M-matrices. *Linear Algebra and Its Applications*, 432:670–677, 2010.

[45] A. Iserles. *A first course in the numerical analysis of differential equations*. Cambridge texts in Applied Mathematics, first edition, 1996. 66, 68

[46] S.A. Ivanenko. *Selected Chapters on Grid Generation and Applications*. Russian Academy of Sciences, second edition, 2010. 47

[47] P.S. Jensen. Finite difference techniques for variable grids. *Computers & Structures*, 2:17–29, 1972. 2

[48] H.B. Keller. *Numerical Methods for Two-Point Boundary-Value Problems*. Dover Books on Mathematics, 2018. 47

[49] C.T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*. Society for Industrial and Applied Mathematics, 1995.

[50] C.T. Kelley. *Iterative Methods for Optimization*. Society for Industrial and Applied Mathematics, 1999.

[51] P. Knupp and S. Steinberg. *Fundamentals of Grid Generation*. CRC Press, 2000.

[52] D.E. Knuth. *The art of computer programming: fundamental algorithms*, volume 1. Addison Wesley Longman, third edition, 1997.

[53] LY Kolotilina. Bounds for the infinity norm of the inverse of certain M- and H-matrices. *Linear Algebra and Its Applications*, 430:692–702, 2009.

[54] J. Kuhnert and S. Tiwari. Finite Pointset Method Based on the Projection Method for Simulations of the Incompressible Navier-Stokes Equations. *Meshfree Methods for Partial Differential Equations*, 2003. 2

[55] P. Kumar and M. Sharan. An analytical model for crosswind integrated concentrations released from a continuous source in a finite atmospheric boundary level. *Atmospheric Environment*, 43:2268–2277, 2009.

[56] P. Kumar and M. Sharan. An analytical model for dispersion of pollutants from a continuous source in the atmospheric boudary layer. *Proceedings of The Royal Society*, 466:383–406, 2010.

[57] Y.W. Kwon and H. Bang. *The Finite Element Method using MATLAB*. CRC Mechanical Engineering Series. CRC Press, first edition, 1997.

[58] R.J. LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations*. Society for Industrial and Applied Mathematics, 2007. 53

[59] P. Li, W Chen, ZJ Fu, and CM Fan. Generalized finite difference method for solving the double-diffusive natural convection in fluid-saturated porous media. *Engineering Analysis with Boundary Elements*, 95:175–186, 2018.

[60] PW Li and CM Fan. Generalized finite difference method for solving two-dimensional Burgers' equation. *Procedia Engineering*, 79:55–60, 2014.

[61] PW Li and CM Fan. Generalized finite difference method for two-dimensional shallow water equations. *Engineering Analysis with Boundary Elements*, 80:58–71, 2017.

[62] Z.C. Li and T.T. Lu. Singularities and Treatments of Elliptic Boundary Value Problems. *Mathematical and Computer Modelling*, 31:97–145, 2000. 2

[63] JS Lin and L.M. Hildemann. Analytical solutions of the atmospheric diffusion equation with multiple sources and height-dependent wind speed and eddy diffusivities. *Atmospheric Environment*, 30(2):239–254, 1996.

[64] T. Liszka and J. Orkisz. The Finite Difference Method at arbitrary irregular grids and its applications in Applied Mathematics. *Computers & Structures*, 11:83–95, 1980.

[65] D. Medková. *The Laplace Equation. Boundary value problems on bounded and unbounded Lipschitz domains.* Springer Nature. Springer International Publishing, first edition, 2018. 2

[66] W. Mitkowski. Dynamical properties of Metzler systems. *Bulletin of the Polish Academy of Sciences*, 56(4), 2008.

[67] D.M. Moreira, M.T. Vilhena, and T. Tirabassi. The state-of-art of the GILTT method to simulate pollutant dispersion in the atmosphere. *Atmospheric Research*, 92:1–17, 2009. 2

[68] H. Motz. The treatment of singularities of partial differential equations by relaxation methods. *Relaxation methods applied to engineering problems*, 4:371–377, 1947. 2

[69] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer, second edition, 2006. 20, 22, 31, 32, 35, 36, 39, 41, 43

[70] E. Oñate. Derivation of stabilized equations for numerical solution of advective-diffusive transport and fluid flow problems. *Computer Methods in Applied Mechanics and Engineering*, 151:233–265, 1998.

[71] V. Pereyra and E.G. Sewell. Mesh selection for discrete solution of boundary problems in ordinary differential equations. *Numer. Math.*, 23:261–268, 1975.

[72] R.J. Plemmons. M-Matrix Characterizations. Nonsingular M-Matrices. *Linear Algebra and Its Applications*, 18:175–188, 1977.

[73] E.O. Reséndiz-Flores and I.D. García-Calvillo. Application of the finite pointset method to non-stationary heat conduction problems. *International Journal of Heat and Mass Transfer*, 71:720–723, 2014. 2

[74] N. Robidoux. *Numerical solution of the steady diffusion equation with discontinuous coefficients*. PhD thesis, Université de Montréal, 1984.

[75] R. Schaback. *A Practical Guide to Radial Basis Functions*, 2007.

[76] Y. Seity, P. Brousseau, S. Malardel, G. Hello, P. Bénard, F. Bouttier, C. Lac, and V. Masson. The AROME-France Convective-Scale Operational Model. *Monthly Weather Review*, 139, 2011.

[77] E. Sousa. The controversial stability analysis. *Applied Mathematics and Computation*, 145:777–794, 2003.

[78] J.C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. Society for Industrial and Applied Mathematics, second edition, 2004. 51, 53

[79] M. Stynes. Steady-state convection-diffusion problems. *Acta Numerica*, pages 445–508, 2005. 2

[80] O. Temel and J. Beeck. Two-equation eddy viscosity models based on the Monin-Obukhov similarity theory. *Applied Mathematical Modelling*, 000:1–16, 2016.

[81] J.F. Thompson, B.K. Soni, and N.P. Weatherill. *Handbook of Grid Generation*. CRC Press, first edition, 1999.

[82] F. Ureña Prieto, J.J. Benito-Muñoz, and L. Gavete-Corvinos. Influence of several factors in the generalized finite difference method. *Applied Mathematical Modeling*, 25:1039–1053, 2001. 2

[83] F. Ureña Prieto, J.J. Benito-Muñoz, and L. Gavete-Corvinos. Solving parabolic and hyperbolic equations by the generalized finite difference method. *Journal of Computational and Applied Mathematics*, 209:208–233, 2007. 2

[84] F. Ureña Prieto, J.J. Benito-Muñoz, and L. Gavete-Corvinos. Application of the generalized finite difference method to solve the advection-diffusion equation. *Journal of Computational and Applied Mathematics*, 235:1849–1855, 2011.

[85] F. Ureña Prieto, J.J. Benito-Muñoz, L. Gavete-Corvinos, and E. Salete. A note on the application of the generalized finite difference method to seismic wave propagation in 2D. *Journal of Computational and Applied Mathematics*, 236:3016–3025, 2012. 2

[86] F. Ureña Prieto, J.J. Benito-Muñoz, L. Gavete-Corvinos, E. Salete, A. García, and M. Ureña. Solving second order non-linear ellipctic partial differential equations using generalized finite difference method. *Journal of Computational and Applied Mathematics*, 318:378–387, 2017. 2

[87] C.R. Vogel. *Computational Methods for Inverse Problems*. Society for Industrial and Applied Mathematics, 2002. 20, 22

[88] S. Vukovic and L. Sopta. Upwind schemes with exact conservation property for one-dimensional open channel flow equations. *Society for Industrial and Applied Mathematics*, 24(5):1630–1649, 2003.

[89] O.C. Zienkiewicz and R.L. Taylor. *The Finite Element Method*, volume 1. Butterworth-Heinemann, fifth edition, 2000.

[90] O.C. Zienkiewicz and R.L. Taylor. *The Finite Element Method*, volume 2. Butterworth-Heinemann, fifth edition, 2000.