UNIVERSIDAD DE EL SALVADOR.

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA.

ESCUELA DE MATEMÁTICA.

Universidad de El Salvador

Hacia la libertad por la cultura

TRABAJO DE GRADUACIÓN:

MÉTODOS DE OPTIMIZACIÓN PARA ESTIMACIÓN DE PARÁMETROS EN MODELOS EPIDEMIOLÓGICOS.

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

LICENCIADAS EN MATEMÁTICA.

Estudiantes: María Elizabeth Moreno Ramos. *Carné:* MR09064
Karen Brizeida Campos Martínez. *Carné:* CM09055

Asesor: MSc. Carlos Ernesto Gámez Rodríguez.

Ciudad Universitaria, 18 de septiembre de 2015.

AUTORIDADES UNIVERSITARIAS
PERIODO 2011-2015

UNIVERSIDAD DE EL SALVADOR

Rector: Ing. Mario Roberto Nieto Lovo
Vicerrectora Académica: Maestra Ana María Glower De Alvarado
Vicerrector Administrativo: Maestro Óscar Noé Navarrete
Secretaria General: Dra. Ana Leticia De Amaya
Defensora De Los Derechos Universitarios: Licda. Claudia María Melgar De Zambrana
Fiscal:Lic. Francisco Cruz Letona

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

Decano: M.Sc. Martín Enrique Guerra Cáceres
Vicedacano: Lic. Ramón Arístides Paz Sánchez
Secretario: Lic. Carlos Antonio Quintanilla Aparicio

ESCUELA DE MATEMATICA

Director: Dr. José Nerys Funes Torres
Secretaria: Licda. Alba Idalia Córdova Cuellar

Índice general

Agradecimientos.

A Dios todo poderoso por habernos regalado vida, sabiduría y paciencia y no permitir que nos rindiéramos en ningún momento y permitirnos culminar nuestra carrera.

A nuestros padres por todo su cariño, apoyo, dedicación, sacrificio y por ayudarnos a ser mejores personas cada día.

A nuestros amigos y amigas con quienes compartimos alegrías y tristezas y que siempre nos dieron ánimos para salir adelante y nunca darnos por vencidas.

A nuestro asesor, MSc. Carlos Gámez, por su disposición y colaboración en cada aspecto de este trabajo.

Al MSc. Porfirio Rodríguez y MSc. René Palacios por revisar el perfil de este proyecto y hacer las correcciones pertinentes.

Introducción.

La relación entre la Matemática y la Biología no ha sido tan estrecha como debería. Con trabajos de muchos investigadores se ha ido vislumbrando la necesidad de relacionar estas dos importantes ramas del conocimiento, y en especial durante el último siglo se han hecho grandes avances en cuanto a este propósito. Para mantener esta tendencia es necesario conocer cada vez más la utilidad y las posibilidades que se generan al construir estos vínculos.

Los trabajos de Vito Volterra en Biología Matemática, han servido para enriquecer tanto las Matemáticas como la Biología y fueron un gran avance en la aplicación de las Matemáticas al campo de la Biología. De forma similar en el paso del tiempo se han seguido otras investigaciones de gran importancia para contribuir en el avance de la aplicación matemática al campo de la biología dentro de las cuales se construyeron modelos epidemiológicos que consisten en sistemas de ecuaciones diferenciales (ED) con parámetros asociados.

A diferencia de los sistemas físicos, para los cuales se dispone de un modelo totalmente establecido por leyes bien conocidas, en los sistemas biológicos como la modelación epidemiológica exigen una gran capacidad de intuición a la hora de establecer las relaciones de causalidad entre las variables. De ahí que, en el estudio de enfermedades epidémicas, a través de los modelos que las describen, es importante el problema de estimar los parámetros que rigen su desarrollo, para su posterior interpretación; así como validar la efectividad de cada modelo, o sea, la certeza de las intuiciones utilizadas en su diseño.

En general, el problema de estimación de parámetros en modelos descritos por ecuaciones diferenciales ordinarias (**EDOs**), con datos o mediciones reales disponibles en el tiempo, puede enfocarse como un problema de optimización. Por ejemplo la modelación de enfermedades infecciosas estudia la dinámica y los factores que causan la proliferación de una enfermedad cuando ésta invade a una población en particular. En el presente trabajo se estudia un grupo de modelos, representados por un sistema de ecuaciones diferenciales, que describen adecuadamente, por ejemplo el proceso de infección y de transmisión de la enfermedad mediante la clasificación de la población en diferentes clases epidemiológicas. Nuestro principal objetivo es probar la validez de la exactitud de uno de los modelos estudiados.

La optimización es un área de la matemática aplicada que permite modelar y resolver problemas de la vida real; sus principios y métodos se usan para resolver problemas cuantitativos en disciplinas como Física, Biología, Ingeniería y Economía. Su objetivo principal es hallar los valores de una función donde se obtiene el máximo o mínimo dependiendo de una serie de restricciones.

La metodología empleada consiste en formular un problema de optimización donde la función objetivo será la acumulación de errores entre los datos predichos por el modelo y los datos reales. Dicha función objetivo da una medida entre la discrepancia de los valores reales y los obtenidos teóricamente tomando como parámetros los parámetros del sistema de ecuaciones diferenciales ED.

Objetivos.

- Buscar e identificar un método óptimo para estimar parámetros en modelos epidemiológicos.

- Mostrar que la matemática es una herramienta sumamente importante para entender distintos fenómenos biológicos.

- Estudiar y entender los modelos epidemiológicos introductorios.

- Entender la teoría básica de optimización.

- Estudiar el método de Levenberg-Marquardt.

- Estudiar métodos de búsqueda de linea para la estimación de parámetros.

Antecedentes y Justificación.

Antecedentes.

Algunas de las figuras relevantes en la modelación de poblaciones.

Alfred James Lotka (1880-1949).

Figura 1: Alfred J. Lotka

Químico, demógrafo y matemático norteamericano de origen ucraniano escribió un libro de biología teórica y varios artículos sobre procesos oscilantes en Química, en donde de manera independiente a Volterra trabajó con la misma ecuación logística de Verhulst pero con el fin de describir una reacción química en la cual las concentraciones oscilan y estableció el modelo que hoy se conoce con el nombre de ambos Lotka-Volterra y que representa aún la base de los estudios teóricos acerca de la dinámica de poblaciones y otros modelos matemáticos en campos tan diversos como la economía, interdependencia compleja, sostenibilidad, tratamiento de plagas, etc.

Vito Volterra (1860-1940).

Figura 2: Vito Volterra

Físico y matemático italiano fue catedrático de la universidad de Roma y senador. Su oposición al fascismo y su origen judío significaron la expulsión de su cátedra y de las sociedades científicas italianas. Exiliado en Francia hasta 1939, impartió cursos en distintos países, entre ellos España. Volterra desarrolló la solución a las ecuaciones integrales de límites variables que lleva su nombre y tras la primera guerra mundial, en la que se

alistó en el cuerpo de ingenieros, se interesó por la aplicación matemática en la biología, extendiendo y desarrollando la obra del matemático belga Pierre François Verhulst, uno de los "padres" de la ecuación logística.

Thomas Robert Malthus (1766).

Figura 3: Thomas Robert Malthus

Nacido en Surrey (condado no metropolitano en el sudeste de Inglaterra, Reino Unido) el 14 de febrero de 1766, su principal estudio fue el Ensayo sobre el Principio de la Población (1798), en el que afirmaba que la población tiende a crecer en progresión geométrica, mientras que los alimentos sólo aumentan en progresión aritmética, por lo que la población se encuentra siempre limitada por los medios de subsistencia. Malthus fue educado según los principios pedagógicos de Jean-Jacques Rousseau, de quien su padre era íntimo amigo. Completó sus estudios en el Jesus College de Cambridge. Después de graduarse en filosofía y teología, fue ordenado pastor anglicano y estuvo durante un tiempo al frente de la parroquia de Albury. En 1793 fue designado miembro del equipo de dirección del Jesus College, puesto al que tuvo que renunciar en 1804 al contraer matrimonio.

Reseña histórica

Los problemas de optimización de funciones son una de las aplicaciones más inmediatas e interesantes del cálculo de derivadas. El problema es determinar los extremos relativos (máximos o mínimos) de una función; se aplican en diferentes contextos, permitiendo resolver problemas de optimización geométricos, biológicos, físicos, económicos entre otros. Sabemos que en la vida cotidiana con frecuencia estamos afrontando muchos problemas de optimización; por ejemplo, buscamos el mejor camino para ir de un lugar a otro, (no necesariamente el más corto), tratamos de hacer la mejor elección al hacer una compra, buscamos la mejor ubicación cuando vamos a un cine o a un teatro, tratamos de enseñar lo mejor posible. Evidentemente, en ninguno de estos casos usamos matemática formalizada y rigurosa para encontrar lo que nos proponemos, pues afrontamos los problemas con los criterios que nos dan la experiencia y la intuición, aunque no necesariamente encontremos la solución óptima.

En una perspectiva más amplia, observamos que los problemas de optimización son parte fundamental de la matemática que ya estaban presentes en los tratados de los griegos de la antigüedad. Una muestra de ello es el libro V de la obra sobre cónicas escrita en ocho tomos por Apolonio, considerado uno de los griegos más importantes de la antigüedad, que vivió entre los años 262 y 190 A.C. en el cual se dedica a estudiar segmentos de longitud máxima y longitud mínima trazados respecto a una cónica. Ciertamente, un hito histórico está marcado por el desarrollo del cálculo diferencial en el siglo XVII y el uso de derivadas para resolver problemas de máximos y mínimos, con lo cual se amplió aún más las aplica-

ciones de las matemáticas en diversos campos de la ciencia y la tecnología y gracias, sobre todo, a Euler se creó el cálculo de variaciones, considerando la obtención de funciones que optimizan funcionales, lo cual proporcionó valiosas herramientas matemáticas para afrontar problemas más avanzados. Otro hito importante en la historia de la optimización se marca en la primera mitad del siglo XX al desarrollarse la programación lineal. Kantorovich y Koopmans recibieron el premio Nobel de economía en 1975, como reconocimiento a sus aportes a la teoría de la asignación óptima de recursos, con la teoría matemática de la programación lineal.

En esta breve mirada histórica, es importante mencionar que Fermat (1601-1665), antes que Newton y Leibinitz publicaran sus trabajos sobre el cálculo diferencial, inventó métodos ingeniosos para obtener valores máximos y mínimos; que Jean Baptiste-Joseph Fourier (1768-1830) mostró aproximaciones intuitivas a métodos de optimización actualmente considerados en la programación lineal; y que el tratamiento riguroso de las ideas de Newton y Leibinitz y de muchos otros anteriores a ellos, que aportaron ideas relevantes al análisis matemático fue desarrollado reciéntemente en el siglo XIX, con Cauchy, Weierstrass y Dedekind.

En matemática y en otras ciencias la optimización matemática es la selección de un mejor elemento de un conjunto de alternativas disponibles.

Por otro lado; a partir del año 1950 la modelación matemática ha sido utilizada en diferentes areas de trabajo como salud entre otras. Los orígenes de la modelación de epidemias, sin embargo, cuenta con una historia más larga. Hace 80 años Kermack y McKendrick propusieron el más importante de los modelos matemáticos para epidemias. Aunque no se trate de un olvido total de ese trabajo durante sus primeros 50 años , todo indica que su amplia difusión comenzó a partir de 1979 con el desarrollo de nuevas herramientas computacionales. Este modelo, conocido en la literatura como “SIR” (por sus siglas; Susceptible, Infectado y Removido) ha impactado positivamente en el área de modelación y control de epidemias.

El problema de la estimación de parámetros o de calibración de modelos consiste en determinar de forma indirecta parámetros desconocidos a partir de las observaciones, que se formula como un problema de optimización no lineal. El proceso de validación de modelos matemáticos que describen aplicaciones prácticas implica, generalmente, la estimación de los parámetros desconocidos que en ellos intervienen.

A diferencia de los sistemas físicos, para los cuales se dispone de un modelo totalmente establecido por leyes bien conocidas, la modelación de procesos epidemiológicos exige una gran dosis de intuición a la hora de establecer las relaciones de causalidad entre las variables.

En general, el problema de estimación de parámetros en modelos descritos por ecuaciones diferenciales ordinarias (EDOs), con datos o mediciones reales disponibles en el tiempo, puede enfocarse como un problema de optimización. En nuestro caso, se minimiza una suma de funciones residuales entre los datos y los valores correspondientes obtenidos de la solución del sistema.

Justificación.

En el estudio de los métodos de optimización se determinan diferentes parámetros para analizar diferentes modelos epidemiológicos. Consideramos que la relevancia del modelo SIR se manifiesta desde diferentes ángulos, entre los que se destacan su simplicidad, su valor didáctico, su aplicabilidad a datos reales, su extensibilidad para el estudio de epidemias con mecanismos más complejos.

1. Su simplicidad. La cual se hace palpable en la sencillez de su estructura pues solo cuenta con tres eslabones en su cadena. Esto permite una fácil obtención de las ecuaciones que definen el sistema.

2. Su valor didáctico. Diferentes aspectos comunes a toda epidemia, como su carácter explosivo que puede ilustrarse a partir del tratamiento de este modelo.

3. Su aplicabilidad a datos reales. En la medida que un modelo refleja la realidad, su utilidad es mayor y mayor es la certeza de que lo que se obtiene en el terreno de la modelación debe manifestarse en el curso de una epidemia.

4. Su extensibilidad para el estudio de epidemias con mecanismos más complejos.

5. La posibilidad de estimar sus parámetros.

6. Nuevas aplicaciones prácticas basadas en el modelo.

Nuestro trabajo se enfatiza en el número 5 y aborda nuevas interrogantes que impone la realidad de las epidemias de nuestro tiempo al ejercicio de su modelación. En particular, proponemos un nuevo enfoque para la estimación de los parámetros del modelo a partir de datos reales. Asimismo, se obtienen soluciones numéricas para un modelo de epidemias asociadas a vectores.

En nuestra opinión, es de mucha importancia conocer mas de cerca el análisis de métodos de optimización para estimar parámetros en modelos conocidos en nuestro ámbito para que así pudiera revertirse en toma de decisiones más acertadas .

Planteamiento del problema.

El modelamiento de enfermedades infecciosas estudia la dinámica y los factores que causan la proliferación de una enfermedad cuando ésta invade una población. El presente trabajo estudia un grupo de modelos (SI, SIS, SIR), representados por un sistema de ecuaciones diferenciales, que describen adecuadamente el proceso de infección y de transmisión de enfermedades mediante la clasificación de la población en diferentes clases epidemiológicas. Donde la correspondiente relación entre susceptibles, infecciosos y recuperados esta dada por:

$$\frac{dS}{dt} = -\beta SI,\ \frac{dI}{dt} = \beta SI - \gamma I,\ \frac{dR}{dt} = \gamma I.$$

En donde podemos identificar cada una de las tasas de infeccción o contagio. El objetivo principal de este trabajo es buscar métodos de optimización para la estimación de parámetros en modelos epidemiológicos dentro de los cuales daremos relevancia al modelo SIR. El estudio sistemático de la propagación de enfermedades infecciosas es una poderosa herramienta para reproducir, analizar, controlar y prevenir eventuales brotes de enfermedades.

Uno de los parámetros más importantes usados en epidemiología es la tasa de reproducción básica $R_0 = \beta N/\gamma$, definida como el número promedio de nuevas infecciones causadas por un individuo infectado cuando es introducido en una población completamente susceptible. Para que una enfermedad persista R_0 deberá ser mayor a uno, de lo contrario la enfermedad no se propaga en una población. Asumiendo una población totalmente homogénea (todos sus habitantes con las mismas características epidemiológicas y hábitos sociales), el número de infecciones es proporcional a la probabilidad de tener un contacto efectivo con un individuo susceptible.

Para un modelo SIR endémico, los nacimientos son el único flujo de entrada al compartimiento de susceptibles (S). Si se asume que la tasa de natalidad y la tasa de mortalidad son iguales para que la población N permanezca constante, entonces se puede utilizar un mismo parámetro para esta tasa.Teniendo en cuenta que para resolver el sistema del modelo SIR haremos uso de la adimensionalización para reducir el número de variables en un sistema mas simple.

Como todos los modelos que tratan de reproducir o representar fenómenos naturales, estos modelos asumen una población homogénea. También asumen interacción homogénea entre los individuos de una población, lo que es bastante irreal. Por ejemplo, es sabido que los niños tienen una tasa de contacto mayor entre ellos que con los adultos, y que éstos últimos a su vez interactúan menos entre sí. La gran ventaja que ofrecen estos modelos, es que nos permiten obtener y estudiar algunas propiedades epidemiológicas importantes de algún brote conservando cierto nivel de simplicidad.

Metodología.

A continuación se describen los aspectos importantes de la metodología de trabajo:

1. **Tipo de investigación.**
   Este proyecto de investigación es de carácter bibliográfico-descriptivo.

   1.1. **Bibliográfico**:
   Se ha hecho una extensa recopilación de libros impresos y de libros obtenidos por Internet para contar con el suficiente material que cubra las necesidades del estudio y de las que puedan surgir más adelante. El objetivo es compilar coherentemente la información más útil y destacada del tema.

   1.2. **Descriptivo**:
   Ya que se describe cada uno de los aspectos importantes para la resolución del problema a resolver.

2. **Forma de Trabajo**
   Se tendrán reuniones periódicas con el asesor del trabajo para tratar los diferentes aspectos de la investigación como estudiar y discutir la teoría, y tratar los diferentes aspectos del trabajo escrito.

3. **Exposiciones**
   Se tendrán dos exposiciones:

   Primera Exposición (Pública) : Presentación del Perfil del Proyecto de Investigación.

   Segunda Exposición (Pública) : Presentación Final del Trabajo de Investigación: resumen de resultados y aplicaciones.

Capítulo 1

Preliminares

1.1. Estabilidad de Ecuaciones Diferenciales

Si se tiene un sistema de ED de primer orden de la forma

$$\frac{dx}{dt} = F(x,y) \qquad \frac{dy}{dt} = G(x,y) \tag{1.1}$$

que nos permite modelar una variedad de fenómenos naturales, donde x e y son variables dependientes y t una variable independiente y además se asume que $F(x,y)$ y $G(x,y)$ son funciones continuas y derivables en una región $\mathbb{R}^2$, se tiene que por el Teorema de Existencia y Unicidad dado t_0 y cualquier punto (x_0, y_0) de $\mathbb{R}^2$ existe solo una solución $x = x(t)$ y $y = y(t)$ para el sistema de ED definida en algún intervalo que contiene a t_0 y que satisface las condiciones iniciales $x(t_0) = x_0$, $y(t_0) = y_0$. Donde las ecuaciones $x(t) = x$ y $y(t) = y$ describen una curva solución denominada **trayectoria**.

Definición 1 *: Un punto crítico del sistema (1.1) es un punto (x_*, y_*) tal que $F(x_*, y_*) = G(x_*, y_*) = 0$.*

Si (x_*, y_*) es un punto crítico del sistema (1.1), entonces las funciones constantes $x = x_*$, $y = y_*$ tienen derivadas $x^{'}(t) = 0$ y $y^{'}(t) = 0$, donde a este tipo de solución se le llama **solución de equilibrio**.

El sistema de ecuaciones que estamos considerando, donde los valores de la derivada son independientes del tiempo t se conoce como sistema **autónomo**.

Definición 2 *: Un punto crítico (x_*, y_*) se dice que es estable siempre que el punto inicial (x_0, y_0) esté suficientemente cercano a (x_*, y_*), entonces $(x(t), y(t))$ permanece cercano a (x_*, y_*) $\forall t > 0$. En notación vectorial, con $\boldsymbol{X}(t) = (x(t), y(t))$, la distancia entre el punto inicial $\boldsymbol{X}_0 = (x_0, y_0)$ y el punto crítico $\boldsymbol{X}_* = (x_*, y_*)$*

$$|\boldsymbol{X}_0 - \boldsymbol{X}_*| = \sqrt{(x_0 - x_*)^2 + (y_0 - y_*)^2}.$$

Así, el punto crítico $\boldsymbol{X}_*$ es estable siempre que, para cada $\varepsilon > 0$, exista un $\delta > 0$ tal que

$$|\boldsymbol{X}_0 - \boldsymbol{X}_*| < \delta \quad implica \quad que \quad |\boldsymbol{X}(t) - \boldsymbol{X}_*| < \varepsilon$$

para todo $t > 0$. Y un punto crítico (x_*, y_*) se dice que es inestable cuando no es estable.

Teorema 1 *: **Estabilidad de Sistemas Lineales***
*Sean λ_1 y λ_2 los eigenvalores de la matriz de coeficientes **A** del sistema lineal de dos dimensiones*

$$\begin{aligned} \frac{dx}{dt} &= ax + by \\ \frac{dy}{dt} &= cx + dy \end{aligned}$$

con $ad - bc \neq 0$. Entonces el punto crítico (0,0) es:

***1-** Asintoticamente estable si las partes reales de λ_1 y λ_2 son ambas negativas;*

***2-** Estable pero no asintoticamente estable si las partes reales de λ_1 y λ_2 son ambas cero (de tal manera que $\lambda_1, \lambda_2 = \pm qi$);*

***3-** Inestable si λ_1 ó λ_2 tienen una parte real positiva.*

Teorema 2 *: **Estabilidad de Sistemas Casi Lineales (Hartman Grobman)***
*Sean λ_1 y λ_2 los eigenvalores de la matriz de coeficientes **A** del sistema lineal de dos dimensiones*

$$\begin{aligned} \frac{dx}{dt} &= ax + by \\ \frac{dy}{dt} &= cx + dy \end{aligned}$$

asociado con el sistema casi lineal

$$\begin{aligned} \frac{dx}{dt} &= ax + by + r(x, y) \\ \frac{dy}{dt} &= cx + dy + s(x, y) \end{aligned}$$

entonces:

***1-** Si $\lambda_1 = \lambda_2$ eigenvalores reales e iguales, por consiguiente el punto crítico (0,0) del sistema asociado es un nodo o un punto espiral, y asintoticamente estable si $\lambda_1 = \lambda_2 < 0$ e inestable si $\lambda_1 = \lambda_2 > 0$.*

***2-** Si λ_1 y λ_2 son imaginarios puros, entonces (0,0) es un centro o un punto espiral, y puede ser asintoticamente estable, ó inestable.*

***3-** En caso contrario, esto es, que λ_1 y λ_2 no sean reales iguales o imaginarios puros, el punto crítico (0,0) del sistema asociado es del mismo tipo y con la misma estabilidad que el punto crítico (0,0) del sistema lineal dado en el teorema anterior.*

1.2. Crecimiento y Decrecimiento Natural

La ecuación diferencial $\frac{dx}{dt} = Kx$ (K constante) sirve como modelo matemático para un notable y amplio número de fenómenos naturales que involucran una variable cuya razón de cambio en el tiempo es proporcional a su tamaño actual. Resolviendo esta ecuación por el método de separación de variables se obtiene la ecuación exponencial o ecuación de crecimiento natural dada por:

$$x(t) = x_0 \exp(Kt);\ x_0 \text{ valor inicial de la población.}$$

Algunos ejemplos de dicha ecuación diferencial son:

Ejemplo 1 . ***Crecimiento de una Población.*** *Suponga que $N(t)$ es el número de individuos en una población (de humanos, insectos o bacterias) con tasas de nacimiento y mortalidad constantes b y d. Entonces, durante un intervalo corto de tiempo Δt, ocurren aproximadamente $bN(t)\Delta t$ nacimientos y $dN(t)\Delta t$ muertes, de tal manera que el cambio en $N(t)$ esta dado aproximadamente por*

$$\Delta N \approx (b-d)N(t)\Delta t.$$

Y por tanto

$$\frac{dN}{dt} = \lim_{\Delta t \to 0} \frac{\Delta N}{\Delta t} = rN,$$

donde $r = b - d$.

Ejemplo 2 . ***Interés Compuesto.*** *Sea $A(t)$ el número de dólares en una cuenta de ahorros en el tiempo t (en años), y supongase que el interés es compuesto y continuo a una tasa de interés anual r. Compuesto y continuo significa que durante un intervalo corto de tiempo Δt, la cantidad de interés sumado a la cuenta es aproximadamente de $\Delta A = rA(t)\Delta t$, tal que*

$$\frac{dA}{dt} = \lim_{\Delta t \to 0} \frac{\Delta A}{\Delta t} = rA.$$

Definición 3 *: La serie de Taylor de una función f real o compleja, $f(x)$ infinitamente diferenciable en el entorno de un número real o complejo a es la siguiente serie de potencia:*

$$f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + ...$$

que puede ser representada de una manera más compacta como:

$$\sum_{n=0}^{\infty} \frac{f^n(a)}{n!}(x-a)^n.$$

El siguiente teorema proporciona un método para el cálculo de conjuntos de parámetros adimensionales de un conjunto de variables dadas, incluso cuando la ecuación es desconocida, pues la elección de parámetros adimensionales no es única, por lo que este teorema solo proporciona una forma de generar conjuntos de parámetros adimensionales.

Teorema 3 *: **Buckingham Pi***
El número de términos adimensionales que se pueden formar, ρ es igual a la nulidad de la matriz de dimensión M y K es el rango.
Tenemos una ecuación física significativa dada por $f(q_1, q_2, ..., q_n) = 0$ donde los q_i son las n variables físicas que están dadas en términos de K unidades físicas independientes, entonces la ecuación anterior puede ser reformulada como $F(\pi_1, \pi_2, ..., \pi_\rho) = 0$ donde los π_i son parámetros sin dimensión construidos a partir de los q_i por $\rho = n - K$ ecuaciones adimensionales, el llamado grupo π de la forma $\pi_i = q_1^{a_1} q_2^{a_2} ... q_n^{a_n}$ donde los exponentes a_i son números racionales (que siempre pueden tomarse como enteros).

Prueba:
Dado un sistema de n variables dimensionales (variables físicas) en K dimensiones, escribimos la matriz M dimensional, cuyas filas son las dimensiones y cuyas columnas son las variables: la (i, j) entrada es la potencia de la $i - esima$ unidad en la $j - esima$ variable. La matriz puede interpretarse como tomar una combinación de las cantidades dimensionales y dar las dimensiones de este producto. Así

$$M = \begin{pmatrix} a_1 \\ . \\ . \\ . \\ a_n \end{pmatrix}$$

son las unidades de $q_1^{a_1} q_2^{a_2} ... q_n^{a_n}$.
Una variable adimensional es una combinación cuyas unidades son todos ceros (por lo tanto, adimensional), que es equivalente a la del núcleo de esta matriz; una variable adimensional es una relación lineal entre las unidades de variables adimensionales.

Por el teorema de rango-nulidad, un sistema de n vectores en K dimensiones (donde son necesarias todas la dimensiones) satisface un $(\rho = n - K)$ espacio de dimension de las relaciones. Cualquier elección de la base tendrá ρ elementos, que son las variables adimensionales.

□

Las variables adimensionales siempre se pueden tomar para ser combinaciones de números enteros de las variables dimensionales. No es matemáticamente elección natural de variables adimensionales; algunas de las variables adimensionales son físicamente mas significativas, y estas son las que se utilizan de forma óptima.

Para ver como funciona este teorema veamos un ejemplo.

Ejemplo 3 *. ¿Supongamos que un carro es conducido a 100km/h; en cuanto tiempo habra recorrido 200km?.*

Esta ecuación tiene dos unidades fisicas fundamentales: tiempo t y longitud l y tres variables dimensionales: distancia D, tiempo tomado T y velocidad V. Por lo que habran $3 - 2 = 1$ cantidad adimensional. Las cantidades de las unidades dimensionales son:

$$D \sim l, T \sim t, V \sim l/t.$$

La matriz dimensional es la siguiente:

$$M = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix}$$

Las filas corresponden a las dimensiones l y t y las columnas a las variables dimensionales D, T, V. Por ejemplo la 3ra columna (1 -1) que contiene la variable V (velocidad) que tiene unidades $l^1 t^{-1} = l/t$.

Para una constante sin dimensiones $\Pi = D^{a_1} T^{a_2} V^{a_3}$ estamos tratando de hallar un vector $a = [a_1, a_2, a_3]$ tal que su producto con la matriz nos de el vector [0 0]. En algebra lineal este vector es llamado el kernel de la matriz dimensional y su combinación es el conjunto nulo, el cual es un caso particular de una dimensión.

Por lo que el kernel de la matriz es $a = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$.Y las variables sin dimensión son $\Pi = D^{-1} T^1 V^1 = TV/D$ ó en terminos dimensionales $\Pi \sim (l)^{-1}(t)^1(l/t)^1 \sim 1$.

Teorema 4 *: **Principio de Superposición Lineal***
Sean $y_1, y_2, ..., y_k$ soluciones de la ED de orden n $\left(a_n(x)\dfrac{d^n y}{dx^n} + ...a_0(x)y = 0\right)$ donde x esta en un intervalo I. La combinación lineal $y = c_1 y_1(x) + c_2 y_2(x) + ... + c_n y_n(x)$ con c_i constantes arbitrarias, también es una solución de la ED de orden n.

Prueba:
Probaremos para $k = 2$. Sea $L = a_n(x)D^n + a_{n-1}(x)D^{n-1} + ... + a_1(x)D + a_0(x)$ ($D_y = dy/dx$) el operador diferencial y sean $y_1(x)$ y $y_2(x)$ soluciones de la ecuación homogénea $L(y) = 0$. Si definimos $y = c_1 y_1(x) + c_2 y_2(x)$ entonces por la linealidad de L, $L(y) = L(c_1 y_1(x) + c_2 y_2(x)) = c_1 L(y_1) + c_2 L(y_2) = c_1 0 + c_2 0 = 0$ por tanto y es solución de la ED.

□

Teorema 5 *: Sean c_1 y $c_2 \in \mathbb{R}$. Supongamos que $x^2 = c_1 x + c_2$ tiene dos raíces distintas r_1 y r_2. Entonces a_n es una solución a la relación de recurrencia $a_n = c_1 a_{n-1} + c_2 a_{n-2} \Longleftrightarrow a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$ es solución, para $n \in \mathbb{N}$ y α_1, α_2 constantes.*

Prueba:
($\Longleftarrow$)
Supongamos que $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$ y verifiquemos que $a_n = c_1 a_{n-1} + c_2 a_{n-2}$ es solución.

$$\begin{aligned} c_1 a_{n-1} &= c_1 \alpha_1 r_1^{n-1} + c_1 \alpha_2 r_2^{n-1} \\ c_2 a_{n-2} &= c_2 \alpha_1 r_1^{n-2} + c_2 \alpha_2 r_2^{n-2} \end{aligned}$$

Ahora sumando ambas ecuaciones y haciendo uso de la hipótesis se tiene:

$$\begin{aligned} c_1 a_{n-1} + c_2 a_{n-2} &= \alpha_1 r_1^{n-2}(c_1 r_1 + c_2) + \alpha_2 r_2^{n-2}(c_1 r_2 + c_2) \\ &= \alpha_1 r_1^{n-2} r_1^2 + \alpha_2 r_2^{n-2} r_2^2 \\ &= \alpha_1 r_1^n + \alpha_2 r_2^n \\ &= a_n \end{aligned}$$

($\Longrightarrow$)
De la primera parte del problema sabemos que $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$ satisface la relación de recurrencia $a_n = c_1 a_{n-1} + c_2 a_{n-2}$. Falta demostrar que $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$ satisface las condiciones iniciales para α_1, α_2. Del teorema se deriva la unicidad de la solución lineal homogénea de la relación de recurrencia.
Para ver si $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$ satisface las condiciones iniciales para α_1, α_2 consideremos $a_1 = \alpha_1 r_1 + \alpha_2 r_2$ y $a_0 = \alpha_1 + \alpha_2$ de donde tenemos un sistema lineal de 2 variables y cuyas soluciones son:

$$\alpha_1 = \frac{a_1 - a_0 r_2}{r_1 - r_2}, \alpha_2 = \frac{a_0 r_1 - a_1}{r_1 - r_2}.$$

□

1.3. Cálculo

Definición 4 *: Dada una función real f de n variables reales:*

$$\begin{array}{rccc} f: & \mathbb{R}^n & \rightarrow & \mathbb{R} \\ & x & \rightarrow & f(x) \end{array}$$

Si todas las segundas derivadas parciales de f existen, se define la matriz hessiana de f como: $H_f(x)$ donde $H_f(x)_{i,j} = \frac{\partial^2 f(x)}{\partial x_i \partial y_j}$ tomando la siguiente forma.

$$H_f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ . & . & & . \\ . & . & \cdots & . \\ . & . & & . \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Definición 5 *: El vector gradiente de f evaluado en un punto genérico x del dominio de f, $\nabla f(x)$, indica la dirección en la cual el campo f varía más rápidamente y su módulo representa el ritmo de variación de f en la dirección de dicho vector gradiente.*

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, ..., \frac{\partial f(x)}{x_n} \right)$$

Definición 6 *: Una matriz definida positiva es una matriz hermitaña (matriz cuadrada de elementos complejos que tiene la característica de ser igual a su propia traspuesta conjugada) que en muchos aspectos es similar a un número real positivo.*

Sea M una matriz hermitaña cuadrada $n \times n$. Esta matriz M se dice definida positiva si cumple con que para todo vector no nulo $z \in \mathbb{C}^n$ $\boldsymbol{z}^t M \boldsymbol{z} > 0$.

1.4. Método de Newton Raphson y Runge Kutta

En esta sección introduciremos dos métodos necesarios para el estudio de los métodos que se emplearan para aproximar la solución de un problema. Dichos métodos son el método de Newton, tanto para aproximar una ecuación o un sistema de ecuaciones no lineales, y el método de Runge Kutta, para resolver problemas de valor inicial de ecuaciones diferenciales ordinarias.

1.4.1. Método de Newton.

El método de Newton (o método de Newton-Raphson), es una de las técnicas numéricas para resolver un problema de búsqueda de raíces $f(x) = 0$, mas poderosas y conocidas. El estudio de este método lo haremos por los polinomios de Taylor.

Supongamos que $f \in C^2[a, b]$. Sea $\overline{x} \in [a, b]$ una aproximación de p talque $f'(\overline{x}) \neq 0$ y $|p - \overline{x}|$ es pequeño. Consideremos el primer polinomio de Taylor para $f(x)$ expandido alrededor de $\overline{x}$.

$$f(x) = f(\overline{x}) + (x - \overline{x})f'(\overline{x}) + \frac{(x - \overline{x})^2}{2}f''(\xi(x)),$$

donde $\xi(x)$ esta entre x y $\overline{x}$. Dado que $f(p) = 0$ esta ecuación, con $x = p$, da

$$0 = f(\overline{x}) + (p - \overline{x})f'(\overline{x}) + \frac{(p - \overline{x})^2}{2}f''(\xi(p)).$$

El método de Newton se obtiene suponiendo que, como $|p - \overline{x}|$ es tan pequeño, el termino que contiene $(p - \overline{x})^2$ es mucho menor y así

$$0 \approx f(\overline{x}) + (p - \overline{x})f'(\overline{x}),$$

despejando p de esta ecuación obtenemos

$$p \approx \overline{x} - \frac{f(\overline{x})}{f'(\overline{x})}.$$

Esto nos prepara para introducir el método de Newton, el cual comienza con una aproximación p_0 y genera la sucesión $\{p_n\}_{n=0}^{\infty}$ definida por

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \quad para\ n \geqslant 1.$$

1.4.2. Método de Runge-Kutta

Es un método de resolución numérica de ecuaciones diferenciales ordinarias. Para introducirlo definamos el problema de valor inicial (PVI):

$$y' = f(t, y), \ \ y(t_0) = y_0$$

Entonces el método de Runge-Kutta de cuarto orden para este problema esta dado por la siguiente ecuación:

$$y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

donde:

$$k_1 = hf(t_n, y_n),$$
$$k_2 = hf(t_n + \frac{h}{2}, y_n + \frac{k_1}{2}),$$
$$k_3 = hf(t_n + \frac{h}{2}, y_n + \frac{k_2}{2}),$$
$$k_4 = hf(t_n + h, y_n + k_3),$$

$\forall n \geqslant 0$. Por ser de cuarto orden, se tiene que el error de paso por teorema de Taylor es de $\mathcal{O}(h^5)$, mientras que el error acumulado tiene orden $\mathcal{O}(h^4)$.

1.5. Dinámica de Población

El modelo de crecimiento de Malthus es el abuelo de todos los modelos de población, y comenzamos con una simple derivación de la famosa ley de crecimiento exponencial.

El crecimiento exponencial sin comprobarlo obviamente no se produce en la naturaleza y las tasas de crecimiento de la población pueden ser reguladas por la comida limitada u otros recursos del medio ambiente y por la competencia de los individuos dentro de una especie o entre especies.

Vamos a desarrollar modelos para tres tipos de regulación. El primer modelo es la ecuación logística, el segundo modelo es una extensión del modelo logístico a la competencia de especies y el tercer modelo son las famosas ecuaciones depredador-presa de Lotka-Volterra. Debido a que todos estos modelos matemáticos son ecuaciones diferenciales no lineales, se consideraran métodos numéricos para obtener y analizar soluciones aproximadas de las ecuaciones.

1.5.1. El Modelo de Crecimiento de Malthus

Sea $N(t)$ el número de individuos de una población en el tiempo t y sean b y d el promedio de tasa de natalidad y de mortalidad por habitante respectivamente. En un corto tiempo Δt el número de nacimientos en la población es $b\Delta tN$, y el número de muertes es $d\Delta tN$. Una ecuación para N en el tiempo $t + \Delta t$ esta determinada por:

$$N(t + \Delta t) = N(t) + b\Delta tN - d\Delta tN$$

$$\frac{N(t + \Delta t) - N(t)}{\Delta t} = (b - d)N(t)$$

y cuando $\Delta t \to 0$ se tiene

$$\frac{dN}{dt} = (b - d)N(t)$$

con un tamaño de la población inicial N_0 y con $r = b - d$ positivo, la solución para $N = N(t)$ crece exponencialmente y esta dada por $N(t) = N_0 \exp(rt)$.

1.5.2. La Ecuación Logística

La ley de crecimiento exponencial para el tamaño de población para tiempos largos es no realista ya que con el tiempo el crecimiento sera revisado por el consumo excesivo de recursos. Asumamos que el medio ambiente tiene una capacidad de carga K.

Para modelar el crecimiento demográfico con capacidad de carga K ambiental, buscamos una ecuación no lineal de la forma $\frac{dN}{dt} = rNF(N)$ donde $F(N)$ proporciona un modelo para la regulación ambiental. Esta función debe satisfacer que $F(0) = 1$ (la población crece exponencialmente con tasa de crecimiento r cuando N es pequeño), $F(K) = 0$ (la población deja de crecer en la capacidad de carga) y $F(N) < 0$ cuando $N > K$ (la población decae cuando es mas grande que la capacidad de carga). La función mas simple $F(N)$ que satisface estas condiciones esta dada por $F(N) = 1 - \frac{N}{K}$. El modelo resultante es la ecuación logística dada por

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right), \tag{1.2}$$

que es un modelo muy importante para muchos procesos, además del crecimiento de la población.

Aunque (1.2) es una ecuación no lineal, una solución analítica se puede encontrar mediante la separación de variables. Antes de abordar esta álgebra, lo primero que ilustraremos serán algunos conceptos básicos que se utilizarán en el análisis de ecuaciones diferenciales no lineales.

Los puntos fijos también llamados puntos de equilibrio, de una ecuación diferencial como (1.2) se definen como los valores de N donde $\frac{dN}{dt} = 0$. Aquí, vemos que los puntos fijos de (1.2) son $N = 0$ y $N = k$. Si el valor inicial de N esta en alguno de estos puntos fijos, entonces N permanece fija en todo tiempo. Los puntos fijos, sin embargo, pueden ser estables o inestables. Un punto fijo es estable si una pequeña perturbación del punto fijo decae a cero, de modo que la solución regresa al punto fijo. Del mismo modo, un punto fijo es inestable si una pequeña perturbación crece exponencialmente de manera que la solución se aleja del punto fijo. Al cálculo de la estabilidad por medio de pequeñas perturbaciones se denomina análisis de estabilidad lineal.

Por ejemplo considerar en general la ecuación diferencial unidimensional (usando la notación $\dot{x} = \frac{dN}{dt}$)

$$\dot{x} = f(x) \tag{1.3}$$

Con x_* un punto fijo de la ecuación, es decir $f(x_*) = 0$. Para determinar analíticamente si x_* es un punto fijo estable o inestable, perturbamos la solución $x = x(t)$ en la forma

$$x(t) = x_* + \varepsilon(t) \tag{1.4}$$

donde $\varepsilon(0)$ es inicialmente pequeño pero distinto de cero. Sustituyendo (1.4) en (1.3) se obtiene

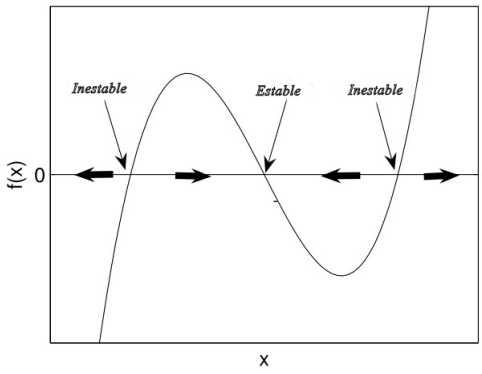


Figura 1.1: *Estabilidad unidimensional mediante el método gráfico.*

$$\begin{aligned} \dot{\varepsilon} &= f(x_* + \varepsilon) &= f(x_*) + \varepsilon f'(x_*) + ... \\ &= & \varepsilon f'(x_*) + ... \end{aligned}$$

donde la segunda igualdad utiliza un desarrollo en serie de Taylor de $f(x)$ sobre x_* y la tercera igualdad utiliza $f(x_*) = 0$, si $f'(x_*) \neq 0$, podemos prescindir de términos de orden superior para ε pequeño e integrando obtenemos que $\varepsilon(t) = \varepsilon(0)\exp(f'(x_*)t)$.

Si la perturbación $\varepsilon(t)$ en el punto fijo x_* tiende a cero cuando $t \to \infty$ entonces $f'(x_*) < 0$, de otra forma $f'(x_*) > 0$. Por lo tanto la condición de estabilidad en x_* es

$$x_* \quad \text{es} \quad \begin{cases} \text{un punto fijo estable si } f'(x_*) < 0, \\ \\ \text{un punto fijo inestable si } f'(x_*) > 0. \end{cases}$$

Otro enfoque equivalente pero más simple de analizar la estabilidad de los puntos fijos de una ecuación no lineal unidimensional, tales como (1.3) es trazar $f(x)$ versus x. Mostramos un ejemplo genérico en la Figura (1.1). Los puntos fijos son los interceptos con el eje X de la gráfica. Las flechas de dirección en el eje X se pueden dibujar basándose en el signo de $f(x)$.

Si $f(x) < 0$, entonces las flechas apuntan hacia la izquierda; si $f(x) > 0$ apuntan hacia la derecha. Las flechas indican la dirección del movimiento de una partícula en la posición x satisfaciendo $\dot{x} = f(x)$. Como se ilustra en la Figura (1.1), puntos fijos con flechas en ambos lados apuntando hacia dentro son estables, y puntos fijos con flechas en ambos lados apuntando hacia fuera son inestables.

En la ecuación logística (1.2), los puntos fijos son $N_* = 0, K$. Un bosquejo de $F(N) = rN\left(1 - \frac{N}{K}\right)$ versus N, con $r, K > 0$ en la Figura (1.2) muestra inmediatamente que $N_* = 0$ es un punto fijo inestable y $N_* = K$ es un punto fijo estable.

El enfoque analítico calcula $F'(N) = r\left(1 - \frac{2N}{K}\right)$, de modo que $F'(0) = r > 0$ y $F'(K) = -r < 0$. Una vez más llegamos a la conclusión que $N_* = 0$ es inestable y $N_* = K$ es estable.

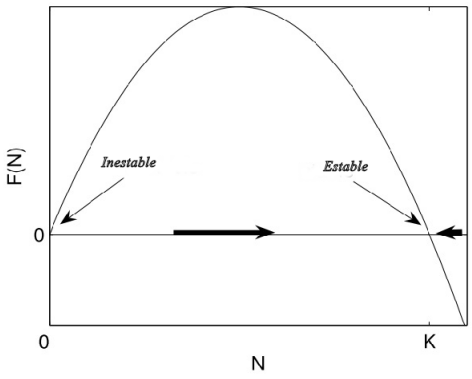


Figura 1.2: *Estabilidad de los puntos fijos de la ecuación logística.*

Ahora resolvamos analíticamente la ecuación logística. Aunque esta ecuación es relativamente simple se puede resolver en primer lugar mediante la no-dimensionalización, para ilustrar esta importante técnica que más tarde resultará ser más útil. Tal vez aquí se puede adivinar la unidad apropiada de tiempo para ser $\frac{1}{r}$ y la unidad apropiada del tamaño de la población a ser K. Sin embargo, preferimos mostrar una técnica mas general que puede aplicarse de manera útil a las ecuaciones para que no sea necesario adivinar las variables sin dimensión.

Tenemos nuestra ecuación logistica en función tamaño de la población N, tiempo t, tasa de crecimiento r, capacidad de carga K. Entonces tenemos $n = 4$ variables dimensionales independientes ya que estan en términos de tiempo y número así:

$$\begin{array}{rcl} [N] & = & p \\ [t] & = & T \\ [r] & = & T^{-1} \\ [K] & = & p. \end{array}$$

Donde p: número. Teniendo $k = 2$ variables dimensionales independientes, entonces $\rho = n - k = 4 - 2 = 2$ combinaciones adimensionales dadas por η y τ que representan los grupos **Pi**.

Luego resolvemos nuestro sistema matricial para el vector columna a, b, c, d, donde las columnas son las variables relevantes y cuyas filas las variables fundamentales.

$$M = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \qquad \text{y} \qquad V = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}$$

Entonces buscamos el kernel de M así:

$$\begin{bmatrix} 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Obteniendo las combinaciones

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \end{pmatrix} \right\}.$$

Entonces $\eta = N^1 t^0 r^0 K^{-1} = \dfrac{N}{K}$ y $\tau = N^0 t^1 r^1 K^0 = tr$. Haciendo un trabajo algebraico la ecuación original

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right),$$

se convierte en

$$\frac{d\eta}{d\tau} = \eta(1-\eta) \tag{1.5}$$

con condiciones iniciales $\eta(0) = \eta_0 = \frac{N_0}{K}$, donde N_0 es el tamaño inicial de la población. Teniendo en cuenta que la ecuación logística adimensional (1.5) no tiene parámetros libres, mientras que la forma dimensional de la ecuación (1.2) contiene r y K.

Reducir el número de parámetros libres (r y K) por el número de unidades independientes (tiempo y tamaño de la población) es una característica general de adimensionalizar. El resultado teórico es conocido como el teorema de Buckingham. Reducir el número de parámetros libres en un problema para el mínimo absoluto es especialmente importante antes de proceder a una solución numérica. El espacio de los parámetros que deben ser expresados puede reducirse sustancialmente.

Resolviendo la ecuación logística adimensional (1.5) mediante separación de variables e integrando de $\tau = 0$ hasta τ y de η_0 a η

$$\int_{\eta_0}^{\eta} \frac{d\eta}{\eta(1-\eta)} = \int_0^{\tau} d\tau$$

La integral en el lado izquierdo puede resolverse utilizando el método de fracciones parciales:

$$\frac{1}{\eta(1-\eta)} = \frac{A}{\eta} + \frac{B}{1-\eta}$$

$$= \frac{A + (B-A)\eta}{\eta(1-\eta)}$$

igualando los coeficientes de los numeradores proporcionales a η^0 y η^1, se obtiene que $A = 1$ y $B = 1$. Por lo tanto,

$$\begin{aligned}\int_{\eta_0}^{\eta}\frac{d\eta}{\eta(1-\eta)} &= \int_{\eta_0}^{\eta}\frac{d\eta}{\eta} + \int_{\eta_0}^{\eta}\frac{d\eta}{(1-\eta)}\\ &= \ln\left(\frac{\eta}{\eta_0}\right) - \ln\left(\frac{1-\eta}{1-\eta_0}\right)\\ &= \ln\left(\frac{\eta(1-\eta_0)}{\eta_0(1-\eta)}\right)\\ &= \tau\end{aligned}$$

despejando η de

$$\frac{\eta(1-\eta_0)}{\eta_0(1-\eta)} = \exp(\tau)$$

se tiene

$$\eta = \frac{\eta_0}{\eta_0 + (1-\eta_0)\exp(-\tau)}.$$

Volviendo a las variables dimensionales, obtenemos

$$N(t) = \frac{N_0}{\frac{N_0}{K} + \left(1 - \frac{N_0}{K}\right)\exp(-rt)}. \tag{1.6}$$

Hay varias maneras de escribir el resultado final dado por (1.6). Decidimos escribir (1.6) en esa forma para observar fácilmente los siguientes resultados:

1. $N(0) = N_0$;
2. $\lim_{t\to\infty} N(t) = K$;
3. $\lim_{K\to\infty} N(t) = N_0\exp(rt)$.

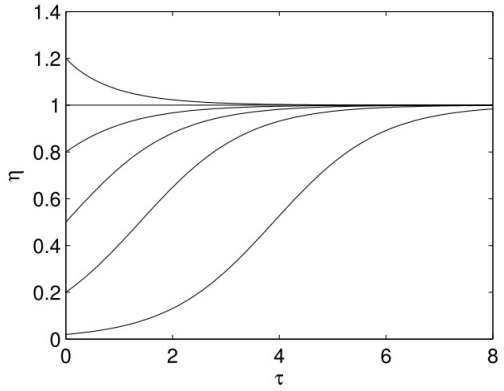


Figura 1.3: *Las soluciones de la ecuación logística adimensional.*

En la Figura (1.3) representamos la solución a la ecuación logística adimensional para las condiciones iniciales $\eta_0 = 0.02, 0.2, 0.5, 0.8, 1.0, 1.2$. La curva mas baja es la característica en

forma de S (ese) usualmente asociada con la solución de la ecuación logística. Esta curva signoidal aparece en otros tipos de modelos.
La secuencia de Matlab que muestra el gráfico es:

```
eta0=[0.02 .2 .5 .8 1 1.2];
tau=linspace(0,8);
for i=1:length(eta0)
    eta=eta0(i)./(eta0(i)+(1-eta0(i)).*exp(-tau));
    plot(tau,eta);hold on
end
axis([0 8 0 1.25]);
xlabel('\tau'); ylabel('\eta'); title('Logistic Equation');
```

1.5.3. Modelo de Competición de Especies

Supongamos que dos especies compiten por los mismos recursos. Para construir un modelo, podemos comenzar con las ecuaciones logísticas para ambas especies. Las diferentes especies tendrían diferente tasa de crecimiento y diferentes capacidades de carga.

Si dejamos que N_1 y N_2 sean el número de individuos de la especie uno y dos respectivamente, entonces

$$\begin{aligned} \frac{dN_1}{dt} &= r_1 N_1 \left(1 - \frac{N_1}{K_1}\right), \\ \frac{dN_2}{dt} &= r_2 N_2 \left(1 - \frac{N_2}{K_2}\right). \end{aligned}$$

Estas son ecuaciones acopladas (separadas que no tienen relación) de modo que asintoticamente $N_1 \to K_1$ y $N_2 \to K_2$. ¿Cómo vamos a modelar la competencia entre las especies?. Si N_1 es mucho menor que K_1 y N_2 mucho más pequeña que K_2, entonces los recursos son abundantes y las poblaciones crecen de forma exponencial con tasas de crecimiento r_1 y r_2. Si las especies uno y dos compiten, entonces el crecimiento de la especie uno reduce los recursos disponibles para la especie dos y viceversa. Dado que no se conoce el impacto que causa una especie sobre la otra introducimos dos parámetros adicionales para modelar la competencia. Una modificación razonable que relaciona las dos ecuaciones logísticas es:

$$\frac{dN_1}{dt} = r_1 N_1 \left(1 - \frac{N_1 + \alpha_{12} N_2}{K_2}\right), \quad \frac{dN_2}{dt} = r_2 N_2 \left(1 - \frac{N_2 + \alpha_{21} N_1}{K_2}\right) \tag{1.7}$$

donde α_{12} y α_{21} son parámetros adimensionales que modelan el consumo de los recursos propios de una especie sobre la otra. Por ejemplo, supongamos que ambas especies comen la misma comida, pero la especie dos consume dos veces de lo de la uno. Dado que un individuo de la especie dos consume el equivalente de dos individuos de la especie uno, el modelo correcto es $\alpha_{12} = 2$ y $\alpha_{21} = \frac{1}{2}$.

Otro ejemplo, supone que las especies uno y dos ocupan el mismo lugar, consumen recursos a la misma velocidad, pero pueden tener diferentes tasas de crecimiento y capacidad de

carga. ¿Pueden las especies coexistir o hace una especie eventualmente conducir a la otra a la extinción? Es posible responder a esta pregunta sin llegar a la solución de las ecuaciones diferenciales.

Con $\alpha_{12} = \alpha_{21} = 1$ según corresponda para este ejemplo, las ecuaciones logísticas acopladas a (1.7) se convierten en

$$\frac{dN_1}{dt} = r_1 N_1 \left(1 - \frac{N_1 + N_2}{K_1}\right), \qquad \frac{dN_2}{dt} = r_2 N_2 \left(1 - \frac{N_1 + N_2}{K_2}\right) \tag{1.8}$$

Por lo antes dicho supongamos que $K_1 > K_2$. Entonces los únicos puntos fijos distintos del trivial $(N_1, N_2) = (0, 0)$ son $(N_1, N_2) = (K_1, 0)$ y $(N_1, N_2) = (0, K_2)$. La estabilidad puede ser calculada analíticamente por una expansión bidimensional en serie de Taylor, pero aquí un argumento mas simple puede ser suficiente. Consideremos en primer lugar $(N_1, N_2) = (K_1, \varepsilon)$ con ε pequeño. Dado que $K_1 > K_2$, observar a partir de (1.8) que la especie dos se extingue $N_2 < 0$. Por lo tanto $(N_1, N_2) = (K_1, 0)$ es un punto fijo estable. Ahora consideremos $(N_1, N_2) = (\varepsilon, K_2)$ con ε pequeño, nuevamente como $K_1 > K_2$, observar a partir de (1.8) que si $N_1 > 0$ la especie uno aumenta en número. Por lo tanto $(N_1, N_2) = (0, K_2)$ es un punto fijo inestable.

Hemos encontrado así que, dentro de nuestro acoplado modelo logístico, las especies que ocupan el mismo nicho y consumen recursos a la misma velocidad no pueden coexistir y que la especie con la mayor capacidad de carga sobrevivirá y conducirá a la otra especie a la extinción. Este es el llamado principio de exclusión competitiva también llamado K-selección (las especies con mayor capacidad de carga salen victoriosas). De hecho, los ecologistas también hablan de r-selección (las especies con las mayores tasas de crecimiento salen victoriosas). Nuestro modelo logístico acoplado no modela r-selección , lo que demuestra las limitaciones potenciales de un modelo matemático muy simple.

Para algunos valores de α_{12} y α_{21} nuestro modelo admite una solución de equilibrio estable en la que conviven dos especies: el cálculo de los puntos fijos y su estabilidad es mas complicado que el calculo solo hecho y se presentan solo los resultados. La convivencia estable de dos especies dentro de nuestro modelo es posible sólo si $\alpha_{12} K_2 < K_1$ y $\alpha_{21} K_1 < K_2$. Ya que resolviendo para los puntos fijos de las ecuaciones (1.7) y teniendo en cuenta que $N_1 \neq N_2 = 0$

$$\begin{aligned} K_1 &= N_1 + \alpha_{12} N_2 \\ K_2 &= \alpha_{21} N_1 + N_2 \end{aligned}$$

y multiplicando la primera ecuación por α_{21} se tiene el nuevo sistema

$$\begin{aligned} \alpha_{21} K_1 &= \alpha_{21} N_1 + \alpha_{21}\alpha_{12} N_2 \\ K_2 &= \alpha_{21} N_1 + N_2 \end{aligned}$$

que al resolverse se obtiene

$$\begin{aligned} N_1 &= K_1 - \alpha_{12} N_2 &> 0 \\ N_2 &= \tfrac{K_2 - \alpha_{21} K_1}{1 - \alpha_{21}\alpha_{12}} &> 0 \end{aligned}$$

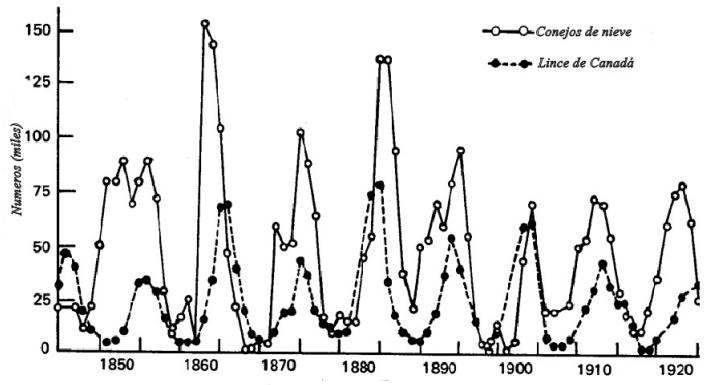


Figura 1.4: *Registros de la Compañía de la Bahía Hudson negociación-Pelt para la liebre y el lince.[Fundamentos de ecología,1953.]*

De donde tenemos que $K_1 > \alpha_{12}N_2$, $K_2 > \alpha_{21}K_1$. De manera análoga resolviendo para N_1 se obtiene que $K_1 > \alpha_{12}K_2$.

Cabe mencionar que el último sistema descrito no es el que mejor representa al conejo-lince ya que es positivo para el lince que hayan más conejos.

1.5.4. El Modelo Depredador-Presa de Lotka-Volterra

Lotka y Volterra propusieron de manera independiente en la década de 1920 un modelo matemático para la dinámica de población: depredador-presa de Lotka-Volterra que se ha convertido en un modelo ícono de la biología matemática.

Para desarrollar estas ecuaciones, supongamos que una población de depredadores se alimenta de una población de presas. Supongamos que el número de presas crece exponencialmente en ausencia de depredadores y que el número de depredadores drece exponencialmente en ausencia de presas. El contacto entre los depredadores y sus presas aumenta el número de depredadores y disminuye el número de presas.

Sean $U(t)$ y $V(t)$ el número de presas y depredadores respectivamente en un tiempo t. Para desarrollar un modelo de ecuaciones diferenciales acopladas, consideremos los tamaños de población en el tiempo $t + \Delta t$. El crecimiento exponencial de presas en ausencia de depredadores puede ser modelado por términos lineales habituales. El acoplamiento entre presa y depredador debe ser modelado con dos parámetros adicionales.

Escribamos el tamaño de la población en el tiempo $t + \Delta t$ como

$$\begin{aligned} U(t + \Delta t) &= U(t) + \alpha \Delta t U(t) - \gamma \Delta t U(t) V(t) \\ V(t + \Delta t) &= V(t) + e\gamma \Delta t U(t) V(t) - \beta \Delta t V(t) \end{aligned}$$

Los parámetros α y β son las tasas de natalidad promedio de la presa y el índice de mortalidad de los depredadores en ausencia de especies. Para el modelo de acoplamiento entre depredadores y presas aparece el parámetro γ que es la fracción de presas atrapadas por los

depredadores por unidad de tiempo; el número total de presas capturadas por los depredadores durante el tiempo Δt es $\gamma \Delta t UV$. La presa comida se convierte entonces en depredadores nacidos, con factor de conversión e, de modo que el número de depredadores durante el tiempo Δt aumenta en $e\gamma \Delta t UV$.

De la conversión de estas ecuaciones diferenciales cuando $\Delta t \to 0$, obtenemos las ecuaciones depredador-presa conocidas como ecuaciones de Lotka-Volterra

$$\frac{dU}{dt} = \alpha U - \gamma UV, \qquad \frac{dV}{dt} = e\gamma UV - \beta V. \tag{1.9}$$

Antes de analizar las ecuaciones de Lotka-Volterra, primero revisaremos los puntos fijos y el análisis de estabilidad lineal aplicada a lo que se llama un sistema autónomo de ecuaciones diferenciales. Por simplicidad, consideramos un sistema de sólo dos ecuaciones diferenciales de la forma

$$\dot{x} = f(x, y), \qquad \dot{y} = g(x, y) \tag{1.10}$$

Aunque nuestros resultados pueden generalizarse a sistemas más grandes. El sistema dado por (1.10) se dice que es autónomo cuando f y g no dependen explicitamente de la variable independiente t.

Los puntos fijos de este sistema se determinan mediante el establecimiento $\dot{x} = \dot{y} = 0$ y resolviendo para x e y. Supongamos que un punto fijo es (x_*, y_*). Para determinar su estabilidad lineal consideramos las condiciones iniciales para (x, y) cerca del punto fijo con pequeñas perturbaciones independientes en ambas direcciones es decir

$$x(0) = x_* + \varepsilon(0), \qquad y(0) = y_* + \delta(0)$$

Si la perturbación inicial crece en el tiempo, decimos que el punto fijo es inestable, si decrece, se dice que el punto fijo es estable. En consecuencia, dejamos

$$x(t) = x_* + \varepsilon(t), \qquad y(t) = y_* + \delta(t). \tag{1.11}$$

Y sustituyendo (1.10) en (1.11) para determinar la dependencia de ε y δ y dado que x_* y y_* son constantes tenemos

$$\dot{\varepsilon} = f(x_* + \varepsilon, y_* + \delta) \qquad \dot{\delta} = g(x_* + \varepsilon, y_* + \delta)$$

El análisis de estabilidad lineal procede del supuesto que las perturbaciones iniciales $\varepsilon(0)$ y $\delta(0)$ son lo suficientemente pequeñas para truncar la expansión de la serie de Taylor bidimensional de f y g sobre $\varepsilon = \delta = 0$ de primer orden en ε y δ.

Tener en cuenta que en general, la serie bidimensional de Taylor de una función $F(x, y)$ sobre el origen está dada por

$$F(x, y) = F(0, 0) + xF_x(0, 0) + yF_y(0, 0) + \frac{1}{2}\left[x^2 F_{xx}(0, 0) + 2xyF_{xy}(0, 0) + y^2 F_{yy}(0, 0)\right] + ...$$

donde los términos de la expansión pueden identificarse requiriendo que todas las derivadas parciales de la serie coincidan con las de $F(x, y)$ en el origen. Ahora expandiendo la serie de Taylor para $f(x_* + \varepsilon, y_* + \delta)$ y $g(x_* + \varepsilon, y_* + \delta)$ sobre $(\varepsilon, \delta) = (0, 0)$ se tiene que los términos constantes desaparecen ya que (x_*, y_*) es un punto fijo, y nos olvidamos de los términos con ordenes mayores de ε y δ. Por lo tanto

$$\dot{\varepsilon} = \varepsilon f_x(x_*, y_*) + \delta f_y(x_*, y_*)$$

$$\dot{\delta} = \varepsilon g_x(x_*, y_*) + \delta g_y(x_*, y_*)$$

que puede ser escrita en forma matricial como

$$\frac{d}{dt}\begin{pmatrix} \varepsilon \\ \delta \end{pmatrix} = \begin{pmatrix} f_x^* & f_y^* \\ g_x^* & g_y^* \end{pmatrix}\begin{pmatrix} \epsilon \\ \delta \end{pmatrix} \tag{1.12}$$

donde $f_x^* = f_x(x_*, y_*)$. La ecuación (1.12) es un sistema lineal de ecuaciones diferenciales, y su solución se obtiene (linealizando) asumiendo que

$$\begin{pmatrix} \varepsilon \\ \delta \end{pmatrix} = \exp(\lambda t)\mathbf{v}. \tag{1.13}$$

Tras la sustitución de (1.12) en (1.12) y cancelando $\exp(\lambda t)$, obtenemos el problema de valores propios de álgebra lineal.

$$\mathrm{J}^*\mathbf{v} = \lambda\mathbf{v} \qquad \text{con} \qquad \mathrm{J}^* = \begin{pmatrix} f_x^* & f_y^* \\ g_x^* & g_y^* \end{pmatrix}$$

donde λ es un valor propio (eigenvalor), $\mathbf{v}$ el correspondiente eigenvector y J^* la matriz jacobiana evaluada en el punto fijo. El eigenvalor se determina a partir de la ecuación característica

$$\det\left(J^* - \lambda I\right) = 0,$$

que es nada más resolver una ecuación cuadrática para λ.

A partir de la solución (1.13), el punto fijo es estable si para todo eigenvalor λ, $\mathbf{Re}\{\lambda\} < 0$, y es inestable si para al menos un λ, $\mathbf{Re}\{\lambda\} > 0$. Donde $\mathbf{Re}\{\lambda\}$ es la parte real del posible eigenvalor complejo λ.

Ahora retomando las ecuaciones de Lotka-Volterra (1.9). Las soluciones en el punto fijo se encuentran resolviendo $\dot{U} = \dot{V} = 0$, y las dos soluciones son

$$(U_*, V_*) = (0, 0) \quad \text{ó} \quad \left(\frac{\beta}{e\gamma}, \frac{\alpha}{\gamma}\right).$$

Para el punto fijo $(0, 0)$ es trivial ver que es inestable, ya que si la población de presas es inicialmente pequeña, esta crece exponencialmente. Para determinar la estabilidad del segundo punto fijo, escribimos la ecuación de Lotka-Volterra en la forma

$$\frac{dU}{dt} = F(U, V), \qquad \frac{dV}{dt} = G(U, V),$$

con

$$F(U, V) = \alpha U - \gamma UV, \qquad G(U, V) = e\gamma UV - \beta V.$$

Donde sus derivadas parciales son

$$\begin{array}{rclrcl} F_U &=& \alpha - \gamma V, & F_V &=& -\gamma U. \\ G_U &=& e\gamma V, & G_V &=& e\gamma U - \beta. \end{array}$$

El jacobiano en el punto fijo $(U_*, V_*) = \left(\dfrac{\beta}{e\gamma}, \dfrac{\alpha}{\gamma}\right)$ es

$$J^* = \begin{pmatrix} 0 & \dfrac{-\beta}{e} \\ \\ e\alpha & 0 \end{pmatrix}$$

y $\det(J^* - \lambda I) = \lambda^2 + \alpha\beta$, tiene las soluciones $\lambda_\pm = \pm i\sqrt{\alpha\beta}$, que son imaginarios puros. Cuando los eigenvalores del jacobiano de dos por dos son imaginarios puros, el punto fijo se denomina centro y la perturbación no crece ni decae, pero oscila. Aquí, la frecuencia angular de oscilación es $\omega = \sqrt{\alpha\beta}$, y el periodo de oscilación es $\dfrac{2\pi}{\omega}$.

Graficamos U y V frente a t (gráfica de series temporales), y V frente a U (diagrama de espacio fase) para ver como se comporta la solución. Para un sistema de ecuaciones no lineales, tales como (1.9), se requiere una solución numérica.

Las ecuaciones de Lotka-Volterra tienen cuatro parámetros libres α, β, γ y e. Las unidades relevantes aquí son el tiempo, el número de presas y el número de depredadores. El teorema de Buckingham Pi predice que las ecuaciones pueden reducir el número de parámetros libres de tres a uno, para una sola agrupación adimensional de parámetros. Desarrollando un argumento similar al de la ecuación logística, se encuentra una nueva representación de las ecuaciones para encontrar su solución o realizando el siguiente análisis, adimensionalizando el tiempo utilizando la frecuencia angular de oscilación y el número de presas y depredadores usando sus valores de punto fijo. Con signos de intercalación que denotan las variables adimensionales, denotamos

$$\hat{t} = \sqrt{\alpha\beta}t, \qquad \hat{U} = \frac{U}{U_*} = \frac{e\gamma}{\beta}U, \qquad \hat{V} = \frac{V}{V_*} = \frac{\gamma}{\alpha}V. \tag{1.14}$$

La sustitución de (1.14) en las ecuaciones de Lotka-Volterra (1.9) da lugar a las ecuaciones adimensionales

$$\frac{d\hat{U}}{d\hat{t}} = r(\hat{U} - \hat{U}\hat{V}), \qquad \frac{d\hat{V}}{d\hat{t}} = \frac{1}{r}(\hat{U}\hat{V} - \hat{V}),$$

con una sola agrupación adimensional de $r = \sqrt{\dfrac{\alpha}{\beta}}$. La especificación de r junto con las condiciones iniciales determina completamente la solución. Debe observarse aquí que la solución a largo plazo de las ecuaciones de Lotka-Volterra depende de las condiciones iniciales. Esta dependencia asintótica de las condiciones iniciales se considera inicialmente como un error del modelo.

Una solución numérica que utiliza `ode45.m` (resuelve un sistema de ED utilizando Runge-Kutta-Fehlberg, es decir, RK de orden 4 y 5) en MATLAB es construida en función de la integración de las ecuaciones diferenciales. El código siguiente genera la Figura (1.5),donde puede observarse como la población de depredadores disminuye la población de presas: un

aumento en el número de presas da como resultado un incremento retardado en el número de depredadores ya que los depredadores comen mas presas. Los diagramas de fase muestran claramente la periodicidad de la oscilación. Tener en cuenta que las curvas se mueven en sentido contrario: el número de presas aumentan cuando el número de depredadores disminuye y el número de presas disminuye cuando el número de depredadores aumenta.

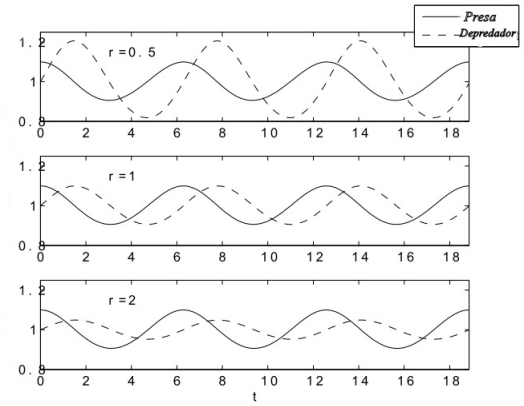
```
function  lotka_volterra
% plots  time  series  and  phase  space  diagrams
[t,UV]=ode45(@(t,UV)  lv_eq(t,UV,r(i)),[t0,tf],[1+eps  1+delta],
   options);
U=UV(:,1);  V=UV(:,2);
subplot(3,1,i);  plot(t,U,t,V,' --');
axis([0  6*pi,0.8  1.25]);  ylabel('predator,prey');
text(3,1.15,['r=',num2str(r(i))]);
end

xlabel('t');
subplot(3,1,1);  legend('prey', 'predator');
%phase  space  plot
xpos=[2.5  2.5  2.5];  ypos=[3.5  3.5  3.5];%for  annotating  graph
    for  i=1:length(r);
       for  eps=0.1:0.1:1.0;
         [t,UV]=ode45(@(t,UV)  lv_eq(t,UV,r(i)),[t0,tf],[1+eps  1+
            delta],options);
         U=UV(:,1);  V=UV(:,2);
         figure(2);subplot(1,3,i);  plot(U,V);  hold  on;
       end
       axis  equal;  axis([0  4  0  4]);
       text(xpos(i),ypos(i),['r=',num2str(r(i))]);
       if  i==1;  ylabel('predator');  end;
       xlabel('prey');
    end

function  dUV=lv_eq(t,UV,r)
dUV=zeros(2,1);
dUV(1)  =  r*(UV(1)-UV(1)*UV(2));
dUV(2)  =  (1/r)*(UV(1)*UV(2)-UV(2));
```

1.6. Optimización

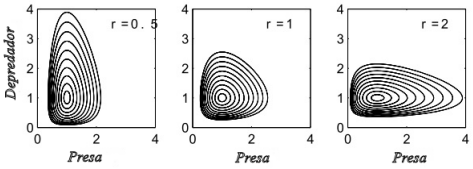
La optimización es una herramienta importante en la toma de decisiones y en el análisis de los sistemas físicos. Para usarlo, debemos primero identificar un objetivo, una medida cuantitativa del rendimiento del sistema en estudio.

Matemáticamente hablando, la optimización es la minimización o maximización de una función, sujeta a restricciones en sus variables. Utilizamos la siguiente notación:

- $\mathbf{x}$ es el vector de variables, también llamadas incógnitas o parámetros;



Soluciones de series de tiempo.



Diagramas de espacio fase.

Figura 1.5: *Soluciones de las ecuaciones de Lotka-Volterra sin dimensiones.*

- f es la función objetivo, una función de x que queremos maximizar o minimizar;
- **c** es el vector de limitaciones que las incógnitas deben satisfacer. Esta es una función del vector de las variables **x**. El número de componentes en **c** es el número de restricciones individuales que colocamos en las variables.

El problema de optimización puede entonces ser escrito como

$$\min_{x\epsilon\mathbb{R}^n} f(x) \qquad \text{sujeto a } \begin{cases} c_i(x) = 0, & i\epsilon\,\varepsilon \\ c_i(x) \geq 0, & i\epsilon\,\mathcal{I} \end{cases}$$

Aquí f y cada c_i son funciones escalares de las variables x, e $\mathcal{I}, \varepsilon$, son conjuntos de índices.

Ejemplo 4 . ***El problema del transporte.*** *Una empresa química tiene dos fabricas F_1 y F_2 y una docena de puntos de venta $R_1, ..., R_{12}$. Cada fábrica F_i puede producir a_i toneladas de un producto químico determinado cada semana; a_i es llamada la capacidad de la planta. Cada punto de venta R_j tiene una demanda semanal conocida de b_j toneladas del producto. El costo de envío de una tonelada de producto de la fábrica F_i para punto de venta R_j es c_{ij}.*

El problema es determinar cuánto del producto enviar desde cada fábrica para cada salida con el fin de satisfacer todos los requisitos y minimizar el costo. Las variables del problema son x_{ij}, $i = 1, 2$ $j = 1, 2, 3, ..., 12$ donde x_{ij} es el número de toneladas de producto enviado desde la fábrica F_i para el punto de venta R_j.

Podemos escribir el problema como

$$\text{mín} \sum_{ij} c_{ij} x_{ij}$$

sujeto a

$$\begin{aligned}
&\sum_{j=1}^{12} x_{ij} \leq a_i, \quad i = 1, 2 \\
&\sum_{i=1}^{2} x_{ij} \geq b_j, \quad j = 1, 2, ...12 \\
&x_{ij} \geq 0 \qquad\quad i = 1, 2;\ j = 1, 2, ...12.
\end{aligned}$$

En un modelo práctico para este problema, también queremos incluir los costos asociados con la fabricación y el almacenamiento del producto. Este tipo de problema se conoce como un problema de programación lineal, ya que la función objetivo y las restricciones son todas funciones lineales.

1.6.1. Optimización Continua y Discreta

El término genérico de optimización discreta se refiere a los problemas en los que la solución que buscamos es una de una serie de objetos en un conjunto finito. Por el contrario, los problemas de optimización continua encuentran una solución de un conjunto infinito estableciendo normalmente un conjunto de vectores con componentes reales. Problemas de optimización continuos son normalmente más fáciles de resolver, debido a que la suavidad de las funciones hace posible el uso de información objetiva y restricción en un punto particular x para deducir información sobre el comportamiento de la función en todos los puntos cercanos a x. La misma afirmación no se puede hacer sobre los problemas discretos donde el conjunto de posibles soluciones es demasiado grande como para hacer una búsqueda exhaustiva para el mejor valor en este conjunto finito.

Algunos modelos contienen variables que tienen permiso para variar de forma continua y otros que pueden alcanzar sólo valores enteros; nos referimos a estos problemas de programación entera como mixtos.

1.6.2. Optimización Restringida y no Restringida

Problemas de optimización sin restricciones surgen directamente en muchas aplicaciones prácticas. Si hay limitaciones naturales de las variables, a veces es seguro hacer caso omiso de ellos y asumir que no tienen ningún efecto sobre la solución óptima. Problemas sin restricciones surgen también como reformulaciones de los problemas de optimización con restricciones, en los que las restricciones se sustituyen por los términos de penalización en la función objetivo que tienen el efecto de violaciones de restricción desalentadores.

Problemas de optimización con restricciones surgen de modelos que incluyen restricciones explícitas sobre las variables. Estas limitaciones pueden ser los límites simples como $0 \leq x_1 \leq 100$, restricciones lineales más generales, tales como $\sum_i x_i \leq 1$, o desigualdades

lineales que representan relaciones complejas entre las variables.

Cuando tanto la función objetivo y todas las restricciones son funciones lineales de x, el problema es un problema de programación lineal. Ciencias de la Gestión e Investigación Operativa hacen un uso extensivo de los modelos lineales. Problemas de programación no lineal, en el que al menos algunas de las limitaciones o el objetivo son funciones no lineales, tienden a surgir de forma natural en las ciencias físicas y la ingeniería, y son cada vez más ampliamente utilizadas en la gestión y las ciencias económicas.

1.6.3. Optimización Local y Global

Los algoritmos de optimización más rápidos sólo buscan una solución local, un punto en el que la función objetivo es más pequeña que en todos los demás puntos factibles en su vecindad. No siempre es posible hallar el mejor de todos esos mínimos, es decir, la solución global. Soluciones globales son necesarias (o al menos altamente deseable) en algunas aplicaciones, pero son difíciles generalmente para identificar y aún más difícil de localizar. Un caso especial importante es la programación convexa en la que todas las soluciones locales son también soluciones globales. Problemas de programación lineal caen en la categoría de programación convexa. Sin embargo, los problemas no lineales generales, tanto limitados y sin restricciones, pueden poseer soluciones locales que no son soluciones globales.

1.6.4. Algoritmos de Optimización

Los algoritmos de optimización son iterativos. Comienzan con una estimación inicial de los valores óptimos de las variables y generan una secuencia de estimaciones mejoradas hasta que llegan a una solución. La estrategia utilizada para pasar de una iteración a la siguiente distingue a un algoritmo de otro. La mayoría de las estrategias hacen uso de los valores de la función objetivo f, las restricciones c y posiblemente, la primera y segunda derivada de estas funciones. Algunos algoritmos acumulan información recopilada en iteraciones anteriores, mientras que otros utilizan sólo información local desde el punto actual. Independientemente de estas especificaciones todos los buenos algoritmos deben poseer las siguientes propiedades:

- Robustez. Deben realizarse bien en una amplia variedad de problemas en su clase, para todas las opciones razonables de las variables iniciales.
- Eficiencia. No deben requerir demasiado tiempo en la computadora.
- Precisión. Ellos deben ser capaces de identificar una solución con precisión, sin estar demasiado sensibles a errores en los datos o a los errores de redondeo aritméticos que se producen cuando el algoritmo se ejecuta en un ordenador.

1.6.5. Convexidad

El concepto de convexidad es fundamental en la optimización; que implica que el problema es simplificado en varios aspectos. El término convexo se puede aplicar tanto a los conjuntos como a las funciones.

- $S \in \mathbb{R}^n$ es un conjunto convexo si el segmento de recta que une dos puntos cualesquiera de S se encuentra en su totalidad dentro de S. Formalmente, para dos puntos $x \in S$ y $y \in S$, tenemos $\alpha x + (1 - \alpha)y \in S$ para todo $\alpha \in [0, 1]$.
- f es una función convexa si su dominio es un conjunto convexo y si para cualquier par de puntos x e y en este dominio la gráfica de f está por debajo de la línea recta que conecta $(x, f(x))$ con $(y, f(y))$ en el espacio $\mathbb{R}^{n+1}$. Es decir, tenemos $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$ para todo $\alpha \in [0, 1]$.

Una función f se dice que es cóncava si $-f$ es convexa.

Como veremos mas adelante, los algoritmos de optimización sin restricciones suelen converger a un punto estacionario (maximizador, minimizador, o en el punto de inflexión) de la función objetivo f. Si sabemos que f es convexa, entonces podemos estar seguros de que el algoritmo ha convergido a un minimizador global.

Optimización tiene sus raíces en el cálculo de variaciones y el trabajo de Euler y Lagrange.

Capítulo 2

Modelos Epidemiológicos

2.1. Modelo SI (Susceptible-Infeccioso)

El modelo más simple de una enfermedad infecciosa clasifica a las personas como susceptibles e infecciosas (SI). Uno puede imaginar que las personas susceptibles son saludables y las personas infecciosas están enfermas. Sin embargo una persona susceptible puede convertirse en infecciosa por contacto con una persona infecciosa. Aquí, como en todos los modelos posteriores, se supone que la población de estudio está bien mezclada para que cada persona tenga la misma probabilidad de entrar en contacto con cualquier otra persona.

Derivamos la ecuación diferencial que rige al modelo SI considerando el número de personas que llegan a ser infecciosos durante el tiempo $\triangle t$. Sea $\beta \triangle t$ la probabilidad de que una persona infecciosa al azar infecta a una persona susceptible durante el tiempo $\triangle t$. Luego, definiendo con S a los susceptibles y con I a los infecciosos, el nuevo número esperado de infecciosos en la población total durante el tiempo $\triangle t$ es $\beta \triangle tSI$. Por lo tanto

$$I(t + \triangle t) = I(t) + \beta \triangle tS(t)I(t)$$

Y del limite cuando $\triangle t \longrightarrow 0$,

$$\frac{dI}{dt} = \beta SI. \tag{2.1}$$

Diagramamos (2.1) como

$$S \xrightarrow{\beta SI} I.$$

Ahora asumimos como N el tamaño de la población constante, dejando de lado los nacimientos y las muertes, por lo que S+I=N. Podemos eliminar S de (2.1) y volver a escribir la ecuación como

$$\frac{dI}{dt} = \beta NI\left(1 - \frac{I}{N}\right),$$

que puede ser reconocida como una ecuación logística, con tasa de crecimiento βN y capacidad de carga N. Por lo tanto $I \longrightarrow N$ cuando $t \longrightarrow \infty$ y toda la población se convertirá en infecciosa.

2.2. Modelo SIS

El modelo SI puede extenderse al modelo SIS, donde un infeccioso puede recuperarse y llegar a ser de nuevo susceptible. Suponemos que la probabilidad de que un infeccioso se recupere durante el tiempo $\triangle t$ está dada por $\gamma \triangle t$. Entonces el número total de personas infecciosas que se recuperan durante el tiempo $\triangle t$ está dada por $IX\gamma \triangle t$, y

$$I(t + \triangle t) = I(t) + \beta \triangle tS(t)I(t) - \gamma \triangle tI(t),$$

y cuando $\triangle t \longrightarrow 0$

$$\frac{dI}{dt} = \beta SI - \gamma I, \tag{2.2}$$

donde su diagrama es $S \underset{\overleftarrow{\gamma I}}{\overset{\beta SI}{\longrightarrow}} I$. Usando S + I = N y eliminando S de (2.2), definimos la relación básica reproductiva como

$$R_0 = \frac{\beta N}{\gamma}.$$

La ecuación (2.2) puede ser reescrita como $\frac{dI}{dt} = \gamma(R_0 - 1)I\left(1 - \frac{I}{N(1 - 1/R_0)}\right)$, que es de nuevo una ecuación logística, pero ahora con tasa de crecimiento $\gamma(R_0 - 1)$ y capacidad de carga $N(1-1/R_0)$. La enfermedad desaparecerá si la tasa de crecimiento es negativa, es decir $R_0 < 1$, y se convertirá en endémica si la tasa de crecimiento es positiva, es decir $R_0 > 1$. Para una enfermedad endémica con $R_0 > 1$, el número de personas infectadas se acerca a la capacidad de carga: $I \longrightarrow N(1 - 1/R_0)$ cuando $t \longrightarrow \infty$.

Podemos dar una interpretación biológica de la relación básica reproductiva R_0 . Sea $l(t)$ la probabilidad de que un individuo infectado inicialmente en $t = 0$ aún es infeccioso en el tiempo $\triangle t$. Dado que la probabilidad de ser infeccioso en el momento $t + \triangle t$ es igual a la probabilidad de ser infeccioso en el tiempo t multiplicado por la probabilidad de no recuperarse durante el tiempo $\triangle t$, tenemos

$$l(t + \triangle t) = l(t)(1 - \gamma \triangle t),$$

donde cuando $\triangle t \longrightarrow 0$,

$$\frac{dl}{dt} = -\gamma l.$$

Con condición inicial $l(0) = 1$,

$$l(t) = \exp(-\gamma t).$$

Ahora, el número esperado de infecciosos secundarios producidos por un solo infeccioso primario durante el período de tiempo $(t\,, t + \triangle t)$ viene dado por la probabilidad de que el infeccioso primario es todavía infeccioso en el tiempo t multiplicado por el número esperado de infecciosos secundarios producidos por el infeccioso durante el tiempo $\triangle t$; es decir, $l(t)XS(t)\beta \triangle t$. Suponemos que el número total de infecciosos secundarios por parte de un individuo infeccioso es pequeño en relación al tamaño de la población N.

Por lo tanto, el número esperado de infecciosos secundarios producidos por un único infeccioso primario introducido en una población completamente susceptible es

$$\begin{aligned}\int_0^\infty \beta l(t)S(t)dt &\simeq \beta N \int_0^\infty l(t)dt \\ &= \beta N \int_0^\infty \exp(-\gamma t)dt \\ &= \frac{\beta N}{\gamma} \\ &= R_0.\end{aligned}$$

Donde hemos aproximado $S(t) \simeq N$ durante el período de tiempo en el que el infeccioso sigue siendo infeccioso. Si una sola persona infectada se introduce en una población totalmente susceptible de más de un infeccioso secundario antes de recuperarse, entonces $R_0 > 1$ y la enfermedad se vuelve endémica.

2.3. Modelo de Enfermedad Epidémica SIR

El modelo SIR, publicado por primera vez por Kermack y McKendrick en 1927, es sin duda el más famoso modelo matemático para la propagación de una enfermedad infecciosa. Aquí, la gente se caracteriza en tres clases: susceptibles S, infecciosos I y removidos R. Individuos retirados ya no susceptibles ni infecciosos por cualquier razón son; por ejemplo, los que se han recuperado de la enfermedad y ahora son inmunes, o los que han sido vacunados o que han sido aislados del resto de la población o tal vez que han muerto a causa de la enfermedad. Al igual que en el modelo SIS, suponemos que infecciosos salen de la clase I con velocidad constante γ, pero en el modelo SIR se mueven directamente a la clase R. Dicho modelo puede ser diagramado como

$$S \xrightarrow{\beta SI} I \xrightarrow{\gamma I} R,$$

donde las correspondientes ecuaciones diferenciales relacionadas son

$$\frac{dS}{dt} = -\beta SI, \ \frac{dI}{dt} = \beta SI - \gamma I, \ \frac{dR}{dt} = \gamma I. \tag{2.3}$$

Con restricción de población constante I +S +R = N. Por conveniencia, adimensionalizamos (2.3) usando N como el tamaño de la población y γ^{-1} como el tiempo; es decir, dejar que

$$\hat{S} = S/N, \ \hat{I} = I/N, \ \hat{R} = R/N, \ \hat{t} = \gamma t,$$

y definir la relación básica reproductiva sin dimensiones como

$$R_0 = \frac{\beta N}{\gamma}. \tag{2.4}$$

Las ecuaciones adimensionales SIR entonces se dan por

$$\frac{d\hat{S}}{d\hat{t}} = -R_0\hat{S}\hat{I}, \ \frac{d\hat{I}}{d\hat{t}} = R_0\hat{S}\hat{I} - \hat{I}, \ \frac{d\hat{R}}{d\hat{t}} = \hat{I}, \tag{2.5}$$

con la restricción adimensional $\hat{S} + \hat{I} + \hat{R} = 1$.

Vamos a utilizar el modelo SIR para abordar dos cuestiones fundamentales: (1) ¿Bajo qué condiciones se produce una epidemia? (2) Si se produce una epidemia, ¿qué fracción de la población mezclada se enferma?

Sean $(\hat{S}_*, \hat{I}_*, \hat{R}_*)$ los puntos fijos de (2.5). Ajustando $d\hat{S}/d\hat{t} = d\hat{I}/d\hat{t} = d\hat{R}/d\hat{t} = 0$, observamos inmediatamente de la ecuación $d\hat{R}/d\hat{t}$ que $\hat{I} = 0$ y este valor fuerza a que todas las derivadas con respecto al tiempo sean iguales a cero para $\hat{S}$ y $\hat{R}$. Donde con $\hat{I} = 0$ se tiene $\hat{R} = (1 - \hat{S})$, y evidentemente todos los puntos fijos de (2.5) se dan por la familia de un parámetro $(\hat{S}_*, \hat{I}_*, \hat{R}_*)$=$(\hat{S}_*, 0, 1 - \hat{S}_*)$.

Una epidemia se produce cuando un pequeño número de infecciosos introducidos en una población susceptible se traduce en un número creciente de infecciosos. Podemos suponer una población inicial en un punto fijo de (2.5), perturbar este punto fijo mediante la introducción de un pequeño número de infecciosos, y determinar la estabilidad del punto fijo. Una epidemia se produce cuando el punto fijo es inestable. El problema de estabilidad lineal se puede resolver considerando sólo la ecuación para $d\hat{I}/d\hat{t}$ en (2.5).

Con $\hat{I} << 1$ y $\hat{S} \simeq \hat{S}_0$, se tiene

$$\frac{d\hat{I}}{d\hat{t}} = (R_0\hat{S}_0 - 1)\hat{I},$$

de manera que si se produce una epidemia $R_0\hat{S}_0 - 1 > 0$. Con la relación básica reproductiva dada por (2.4) y $\hat{S}_0 = S_0/N$ donde S_0 es el número inicial de individuos susceptibles y la epidemia ocurre si

$$R_0\hat{S}_0 = \frac{\beta S_0}{\gamma} > 1,$$

que se podría haber adivinado. Una epidemia se produce cuando un individuo infeccioso introducido en una población de individuos susceptibles S_0 infecta en promedio a más de una persona.

Ahora nos dirigimos a la segunda pregunta: Si se produce una epidemia, ¿qué fracción de la población se enferma? Por simplicidad, se supone que toda la población inicial es susceptible a la enfermedad, de modo que $\hat{S}_0 = 1$. Esperamos que la solución de las ecuaciones que gobiernan (2.5) se acerquen a un punto fijo asintóticamente en el tiempo (por lo que el número final de infecciosos será cero), y definimos este punto fijo para ser $(\hat{S}, \hat{I}, \hat{R})$=$(1 - \hat{R}_\infty, 0, \hat{R}_\infty)$, con $\hat{R}_\infty$ igual a la fracción de la población que se enferma. Para calcular $\hat{R}_\infty$ es más sencillo trabajar con una versión transformada de (2.5). Por la regla de la cadena, $d\hat{S}/d\hat{t} = (d\hat{S}/d\hat{R})(d\hat{R}/d\hat{t})$, tal que

$$\begin{aligned} \frac{d\hat{S}}{d\hat{R}} &= \frac{d\hat{S}/d\hat{t}}{d\hat{R}/d\hat{t}} \\ &= -R_0\hat{S} \end{aligned}$$

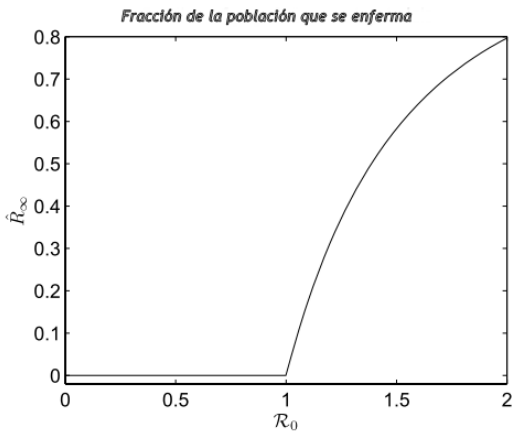


Figura 2.1: *Fracción de la población que se enferma del modelo SIR.*

que es separable. La separación y la integración de las condiciones iniciales y finales

$$\int_1^{\hat{1}-\hat{R}_\infty} \frac{d\hat{S}}{\hat{S}} = -R_0 \int_0^{\hat{R}_\infty} d\hat{R},$$

que tras la integración y la simplificación, los resultados se presentan en la siguiente ecuación transcendental para $\hat{R}_\infty$:

$$\hat{R}_\infty - \exp(-R_0\hat{R}_\infty) = 0,$$

una ecuación que se puede resolver numéricamente usando el método de Newton. Donde tenemos

$$\begin{aligned} F(\hat{R}_\infty) &= 1 - \hat{R}_\infty - \exp(-R_0\hat{R}_\infty), \\ F'(\hat{R}_\infty) &= -1 + R_0 \exp(-R_0\hat{R}_\infty); \end{aligned}$$

y usamos el método de Newton para resolver $F(\hat{R}_\infty) = 0$ iterando $\hat{R}_\infty^{(n+1)} = \hat{R}_\infty^n - \dfrac{F(\hat{R}_\infty^{(n)})}{F'(\hat{R}_\infty^{(n)})}$ para R_0 fijo y una condición inicial adecuada para $R_\infty^{(0)}$, que tomamos como unidad. A continuación se da el código para calcular R_∞ como una función de R_0, y el resultado se muestra en la figura anterior. Donde se produce una explosión en el número de infecciones si R_0 sobrepasa la unidad, y este rápido aumento es un ejemplo clásico de lo que se conoce más generalmente como fenómeno umbral.

```
function [R0, R_inf] = sir_rinf
 computes solution of R_inf using Newton's method from SIR model
nmax=10; numpts=1000;
R0 = linspace(0,2,numpts); R_inf = ones(1,numpts);
for i=1:nmax
    R_inf = R_inf - F(R_inf,R0)./Fp(R_inf,R0);
end
plot(R0,R_inf); axis([0 2 -0.02 0.8])
xlabel('$\mathcal{R}_0$', 'Interpreter', 'latex', 'FontSize',16)
ylabel('$\hat R_\infty$', 'Interpreter', 'latex', 'FontSize',16);
title('fraction of population that get sick')
```

```
subfunctions
function y = F(R_inf,R0)
y = 1 - R_inf - exp(-R0.*R_inf);
function y = Fp(R_inf,R0)
y = -1 + R0.*exp(-R0.*R_inf);
```

Capítulo 3

Fundamentos de Optimización sin Restricciones

En la optimización sin restricciones, minimizamos una función objetivo que depende de variables reales, sin restricciones en absoluto sobre los valores de estas variables. La formulación matemática es:

$$\min_x f(x),$$

donde $\mathbf{x} \in \mathbb{R}^n$ es un vector real con $n \geqslant 1$ y $f : \mathbb{R}^n \to \mathbb{R}$ es una función suave.

Por lo general, carecemos de una perspectiva global de la función f. Todo lo que sabemos son los valores de f y tal vez algunos de sus derivados en un conjunto de puntos $x_0, x_1, x_2, ...$

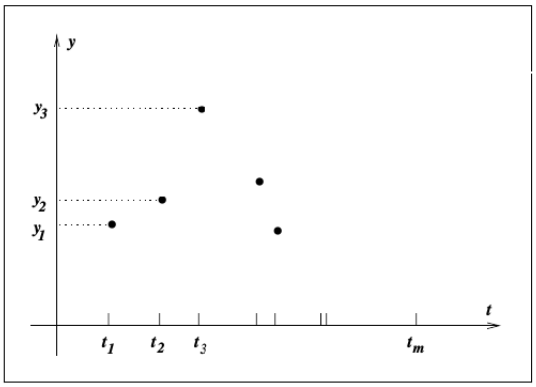


Figura 3.1: *El problema de ajuste con mínimos cuadrados.*

Ejemplo 5 . ***El problema de ajuste con mínimos cuadrados.*** *Supongamos que estamos tratando de encontrar una curva que se ajuste a algunos datos experimentales. La Figura 3.1 muestra las mediciones $y_1, y_2, ..., y_m$ de una señal tomada en los tiempos $t_1, t_2, ..., t_m$. A partir de los datos y el conocimiento de la aplicación, podemos deducir que la señal tiene un comportamiento exponencial y oscilatorio en ciertos tiempos, y optamos por modelarlo por la función*

$$\phi(t; \boldsymbol{x}) = x_1 + x_2 \exp^{-(x_3 - t)^2/x_4} + x_5 cos(x_6 t).$$

Los números reales x_i , $i = 1, 2, ..., 6$ son los parámetros del modelo. Nos gustaría elegir los valores del modelo $\phi(t_j; \mathbf{x})$ para hacerlos encajar con los datos observados y_j en la mayor medida posible. Para afirmar nuestro objetivo como un problema de optimización, agrupamos los parámetros x_i en un vector de incógnitas $\mathbf{x} = (x_1, x_2, ..., x_6)^T$,y definimos los residuales

$$r_j = y_j - \phi(t_j, \mathbf{x}), \ j = 1, ..., m.$$

Los cuales miden la discrepancia entre el modelo y los datos observados. Nuestra estimación de $\mathbf{x}$ se obtiene entonces resolviendo el problema

$$\min_{\mathbf{x} \in \mathbb{R}^6} f(\mathbf{x}) = r_1^2(\mathbf{x}) + ... + r_m^2(\mathbf{x}). \tag{3.1}$$

Este es un problema no lineal de mínimos cuadrados, un caso especial de optimización sin restricciones.

□

Supongamos que para los datos dados en la Figura 3.1 la solución óptima de (3.1) es de aproximadamente $\mathbf{x}^* = (1.1, 0.01, 1.2, 1.5, 2.0, 1.5)$ y el valor de la función correspondiente es $f(\mathbf{x}^*) = 0.34$. Debido a que el objetivo óptimo es distinto de cero, tiene que haber discrepancia entre las mediciones y_j observados y las predicciones del modelo $\phi(t_j, \mathbf{x}^*)$ para algunos de los valores de j. Por lo que el modelo no ha producido todos los puntos con exactitud. ¿Cómo, entonces, podemos verificar que $\mathbf{x}^*$ es de hecho un minimizador de f?. Para responder a esta pregunta, tenemos que definir el término "solución" y explicar cómo reconocer soluciones.

¿Cúal es la solución?

Generalmente, seríamos más felices si nos encontramos un minimizador global de f, un punto donde la función alcanza su mínimo valor. Una definición formal es:

Definición 7 *: Un punto x^* es un minimizador global si $f(x^*) \leqslant f(x)$, $\forall x$.*

Donde x se extiende sobre todo $\mathbb{R}^n$ (o al menos sobre el dominio de interés). El minimizador global puede ser difícil de hallar, porque nuestro conocimiento de f es por lo general sólo local. La mayoría de los algoritmos son capaces de hallar sólo un minimizador local, que es un punto donde f alcanza el mínimo valor en su vecindad. Formalmente, vamos a decir que:

Definición 8 *: Un punto x^* es un minimizador local si existe un vecindario $\mathcal{N}$ de x^* tal que $f(x^*) \leqslant f(x)$ $\forall x \in \mathcal{N}$.*

(Recordemos que un entorno de x^* es simplemente un conjunto abierto que contiene a x^*.) Un punto que satisface esta definición se denomina a veces débil minimizador local. Esta terminología distingue de un estricto minimizador local, que es el ganador absoluto en su vecindario.

Definición 9 *: Un punto x^* es un estricto minimizador local (también llamado fuerte minimizador local) si existe una vecindad $\mathcal{N}$ de x^* tal que $f(x^*) < f(x)$ $\forall x \in \mathcal{N}$ con $x \neq x^*$.*

Para la función constante $f(x) = 2$, cada punto x es un débil minimizador local, mientras que la función $f(x) = (x-2)^4$ tiene un estricto minimizador local en $x = 2$.

Un tipo un poco más exótico de minimizador local es definido de la siguiente manera.

Definición 10 *: Un punto x^* es un minimizador local aislado si existe una vecindad $\mathcal{N}$ de x^* tal que x^* es el único minimizador local en $\mathcal{N}$.*

Algunos minimizadores locales estrictos no son aislados, como lo ilustra la función

$$f(x) = x^4 cos(1/x) + 2x^4 \ , f(0) = 0.$$

Que es dos veces continuamente diferenciable y tiene un minimizador local estricto en $x^* = 0$. Sin embargo, existen minimizadores locales estrictos en muchos puntos cercanos a x_j que podemos etiquetarlos de tal manera que $x_j \to 0$ cuando $j \to \infty$.

Mientras que los minimizadores locales estrictos no son siempre aislados, es cierto que todos los minimizadores locales aislados son minimizadores estrictos. A veces tenemos conocimiento adicional "global"sobre f que puede ayudar en la identificación de los mínimos globales. Un caso especial importante es el de las funciones convexas, para las cuales cada minimizador local es también un minimizador global.

3.1. Teoremas Fuertes de Optimización sin Restricciones

De las definiciones dadas anteriormente, podría parecer que la única manera de determinar si un punto x^* es un mínimo local es examinar todos los puntos en sus inmediaciones, para asegurarse de que ninguno de ellos tiene un valor menor de la función. Cuando la función f es suave, sin embargo hay muchas formas prácticas más eficientes para identificar los mínimos locales. En particular, si f es dos veces continuamente diferenciable, podemos ser capaces de decir que x^* es un minimizador local (posiblemente un estricto minimizador local) mediante el examen de sólo el gradiente $\bigtriangledown f(x^*)$ y el Hessiano $\bigtriangledown^2 f(x^*)$.

La herramienta matemática utilizada para estudiar minimizadores de funciones suaves es el teorema de Taylor. Debido a que este teorema es central en nuestro análisis, lo declaramos ahora.

Teorema 6 *: **Teorema de Taylor.***
Supongamos que $f : \mathbb{R}^n \to \mathbb{R}$ es continuamente diferenciable y que $p \in \mathbb{R}^n$. Entonces se tiene que

$$f(x+p) = f(x) + \bigtriangledown f(x+tp)^T p,$$

para algun $t \in (0,1)$. Por otra parte, si f es dos veces continuamente diferenciable, tenemos que

$$\bigtriangledown f(x+p) = \bigtriangledown f(x) + \int_0^1 \bigtriangledown^2 f(x+tp) p dt,$$

tal que

$$f(x+p) = f(x) + \bigtriangledown f(x)^T p + \frac{1}{2} p^T \bigtriangledown^2 f(x+tp)p,$$

para algun $t \in (0,1)$.

Teorema 7 *: **Condiciones necesarias de primer orden.***
Si x^* *es un minimizador local y* f *es continuamente diferenciable en un entorno abierto de* x^**, entonces* $\bigtriangledown f(x^*) = 0$.

Prueba:
Supongamos por contradicción que $\bigtriangledown f(x^*) \neq 0$. Definimos el vector $p = - \bigtriangledown f(x^*)$ y observemos que $p^T \bigtriangledown f(x^*) = -|| \bigtriangledown f(x^*)||^2 < 0$. Debido a que $\bigtriangledown f$ es continuo cerca de x^*, hay un escalar $T > 0$ tal que

$$p^T \bigtriangledown f(x^* + tp) < 0 \ , \forall \bar{t} \in [0, T].$$

Para cualquier $\bar{t} \in (0, T]$ tenemos por el teorema de Taylor que

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t} p^T \bigtriangledown f(x^* + tp), \ \text{ para algun } t \in (0, \bar{t}).$$

Por lo tanto $f(x^* + \bar{t}p) < f(x^*)$ para todo $t \in (0, T]$. Hemos encontrado asi una dirección que se aleja de x^* a lo largo de la cual f disminuye, por lo que x^* no es un minimizador local lo cual es una contradicción.

□

Llamamos a x^* un punto estacionario si $\bigtriangledown f(x^*) = 0$. De acuerdo con el Teorema (7), cualquier minimizador local debe ser un punto estacionario.

Para el siguiente resultado recordemos que una matriz B es definida positiva si para todo $p \neq 0$ $p^T B p > 0$, y es semidefinida positiva si $p^T B p \geqslant 0$ $\forall p$.

Teorema 8 *: **Condiciones necesarias de segundo orden.***
Si x^* *es un minimizador local de* f *y* $\bigtriangledown^2 f$ *es continuo en un entorno abierto de* x^**, entonces* $\bigtriangledown f(x^*) = 0$ *y* $\bigtriangledown^2 f(x^*)$ *es semidefinido positivo.*

Prueba:
Sabemos por el Teorema (7) que $\bigtriangledown f(x^*) = 0$. Por contradicción, supongamos que $\bigtriangledown^2 f(x^*)$ no es semidefinido positivo. Entonces podemos elegir un vector p tal que $p^T \bigtriangledown^2 f(x^*)p < 0$ y como $\bigtriangledown^2 f$ es continuo cerca de x^*, hay un escalar $T > 0$ tal que $p^T \bigtriangledown^2 f(x^* + tp)p < 0$ para todo $t \in [0, T]$.

Al hacer un desarrollo en serie de Taylor alrededor de x^*, tenemos que para todo $\bar{t} \in (0, T]$ y algún $t \in (0, \bar{t})$ que

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t} p^T \bigtriangledown f(x^*) + \frac{1}{2} \bar{t}^2 p^T \bigtriangledown^2 f(x^* + tp)p < f(x^*).$$

Como en el Teorema (7), hemos encontrado una dirección de x^* a lo largo de f que está disminuyendo, y así de nuevo, x^* no es un minimizador local.

□

Describimos ahora condiciones suficientes, que son las condiciones en las derivadas de f en el punto z^* que garantiza que x^* es un minimizador local.

Teorema 9 *: **Condiciones suficientes de segundo orden.***
Supongamos que $\bigtriangledown^2 f$ es continuo en un entorno abierto de x^* y que $\bigtriangledown f(x^*) = 0$ y $\bigtriangledown^2 f(x^*)$ es definido positivo. Entonces x^* es un estricto minimizador local de f.

Prueba:
Debido a que el Hessiano es continuo y definido positivo en x^*, podemos elegir un radio $r > 0$ para que $\bigtriangledown^2 f(x)$ se mantenga positivo para todo x en la bola abierta $\mathcal{D} = \{z \mid ||z - x^*|| < r\}$. Tomando cualquier vector no nulo p con $||p|| < r$, tenemos que $x^* + p \in \mathcal{D}$ y así

$$\begin{array}{rcccccc} f(x^*+p) & = & f(x^*) & + & p^T \bigtriangledown f(x^*) & + & \frac{1}{2} \bigtriangledown^2 f(z)p \\ & = & f(x^*) & + & \frac{1}{2} \bigtriangledown^2 f(z)p, & & \end{array}$$

donde $z = x^* + tp$ para algun $t \in (0, 1)$. Dado que $z \in \mathcal{D}$, $p^T \bigtriangledown^2 f(z)p > 0$ y por lo tanto $f(x^* + p) > f(x^*)$, dando el resultado.

□

Teorema 10 *:*
Cuando f es convexa, cualquier minimizador local x^* es un minimizador global de f. Si además f es diferenciable, entonces cualquier punto estacionario x^* es un minimizador global de f.

Prueba:
Supongamos que x^* es un minimizador local, pero no global. Entonces podemos encontrar un punto $z \in \mathbb{R}^n$ con $f(z) < f(x^*)$. Ahora consideremos el segmento de recta que une x^* con z es decir,

$$x = \lambda z + (1 - \lambda)x^*, \ \textit{para algun } \lambda \in (0, 1]. \tag{3.2}$$

Por la propiedad de convexidad de f, tenemos

$$f(x) \leqslant \lambda f(z) + (1 - \lambda) f(x^*) < f(x^*). \tag{3.3}$$

Cualquier vecindad $\mathcal{N}$ de x^* contiene una pieza del segmento de línea (4.1), por lo que siempre habrán puntos $x \in \mathcal{N}$ en la que (4.2) se satisface. Por lo tanto, x^* no es un minimizador local.

Para la segunda parte del teorema, supongamos que x^* no es un minimizador global y elegimos z como anteriormente. Luego, a partir de la convexidad, tenemos

$$\begin{array}{rcl} \bigtriangledown f(x^*)^T (z - x^*) & = & \dfrac{d}{d\lambda} f(x^* + \lambda(z - x^*))|_{\lambda=0} \\ & = & \lim_{\lambda \to 0} \dfrac{f(x^* + \lambda(z - x^*)) - f(x^*)}{\lambda} \\ & \leqslant & \lim_{\lambda \to 0} \dfrac{\lambda f(z) + (1 - \lambda) f(x^*) - f(x^*)}{\lambda} \\ & = & f(z) - f(x^*) < 0. \end{array}$$

Por lo tanto $\bigtriangledown f(x^*) \neq 0$, por lo que x^* no es un punto estacionario.

□

Todos los algoritmos de minimización sin restricciones requieren que el usuario proporcione un punto de partida, que por lo general se denota por x_0. El usuario con conocimientos sobre la aplicación y el conjunto de datos puede estar en una buena posición para elegir x_0 y asi hacer una estimación razonable de la solución. De lo contrario, el punto de partida debe ser elegido de alguna manera arbitraria.

3.2. Tasa de Convergencia

Una de las medidas clave del rendimiento de un algoritmo es su velocidad de convergencia. Definimos ahora la terminología asociada con diferentes tipos de convergencia.

Definición 11 *: Sea $\{x_k\}$ una secuencia en $\mathbb{R}^n$ que converge a x^*. Decimos que la convergencia es Q-lineal si hay una constante $r \in (0,1)$ tal que*

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \leq r$$

para todo k suficientemente grande.

Esto significa que la distancia a la solución x^* disminuye en cada iteración por al menos un factor constante. Por ejemplo, la secuencia $1 + (0.5)^k$ converge Q-linealmente a 1. El prefijo "Q"significa çociente", porque este tipo de convergencia se define en términos del cociente de errores sucesivos.

Definición 12 *: La convergencia se dice que es Q-superlineal si*

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0.$$

Convergencia Q-Cuadrática, es una tasa de convergencia mas rapida, se obtiene si

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} \leq M,$$

para todo k suficientemente grande. Donde M es una constante positiva, no necesariamente inferior a uno.

También podemos definir índices más altos de convergencia (cúbica, cuártica, y así sucesivamente), pero estos son menos interesante en términos prácticos. En general, se dice que el Q-orden de convergencia es p (con $p > 1$) si existe una constante positiva M tal que

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} \leq M,$$

para todo k suficientemente grande.

Capítulo 4

Métodos de Búsqueda de Línea

Cada iteración de un método de búsqueda de línea calcula una dirección de búsqueda p_k y luego decide que tan lejos moverse a lo largo de esa dirección. La iteración está dada por

$$x_{k+1} = x_k + \alpha_k p_k, \tag{4.1}$$

donde el escalar positivo α_k se llama la longitud de paso. El éxito de un método de búsqueda de línea depende de las opciones eficaces tanto de la dirección p_k y la longitud de paso α_k. La mayoría de los algoritmos de búsqueda de línea requieren que p_k sea una dirección de descenso, para el cual

$$p_k^T \nabla f_k < 0,$$

porque esta propiedad garantiza que la función f puede ser reducida a lo largo de esta dirección. Por otra parte, la dirección de búsqueda a menudo tiene la forma

$$p_k = -B_k^{-1} \nabla f_k, \tag{4.2}$$

donde B_k es una matriz simétrica y no singular. En el método del descenso máximo, B_k es simplemente la matriz identidad I, mientras que en el método de Newton, B_k es el Hessiano exacto $\nabla^2 f(x_k)$. En los métodos cuasi-Newton, B_k es una aproximación del Hessiano, que se actualiza en cada iteración a través de una fórmula de bajo rango. Cuando p_k es definida por (4.2) y B_k es definida positiva, tenemos

$$p_k^T \nabla f_k = -\nabla f_k^T B_k^{-1} \nabla f_k < 0,$$

y por lo tanto p_k es una dirección de descenso.

En este capítulo, se discute cómo elegir α_k y p_k para promover la convergencia de puntos de partidas remotas. También estudiamos la tasa de convergencia del descenso máximo, cuasi-Newton, y los métodos de Newton. Damos ahora una cuidadosa consideración a la elección del parámetro de paso de longitud α_k.

4.1. Longitud de Paso

En el cálculo de α_k, nos enfrentamos a una solución de compromiso. Nos gustaría elegir α_k para dar una reducción substancial de f, pero al mismo tiempo no queremos pasar demasiado tiempo en hacer la elección. La opción ideal sería el minimizador global de la función

univariada $\phi(\cdot)$ definida por

$$\phi(\alpha) = f(x_k + \alpha p_k), \qquad \alpha > 0 \tag{4.3}$$

pero en general, es demasiado caro para identificar este valor (véase la figura(4.1)). Para encontrar incluso un minimizador local de ϕ a una precisión moderada generalmente se requieren demasiadas evaluaciones de la función objetivo f y posiblemente del gradiente ∇f. Estrategias más prácticas realizan una búsqueda de línea para identificar una longitud de paso que logra reducciones adecuadas en f a un costo mínimo.

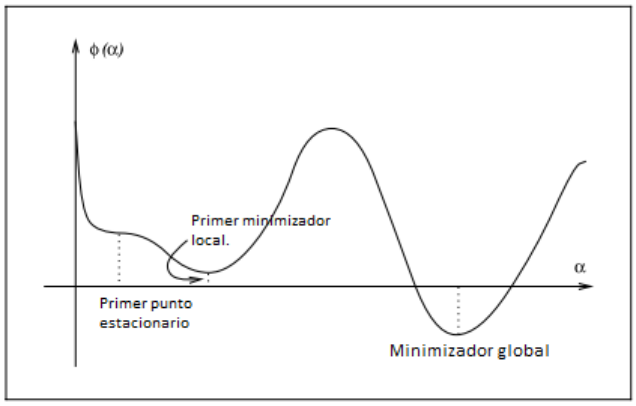


Figura 4.1: *El paso de longitud ideal es el minimizador global.*

Algoritmos de búsqueda de línea típicos prueban una secuencia de posibles valores para α, parando a aceptar uno de estos valores cuando se cumplen ciertas condiciones. La búsqueda de línea se realiza en dos etapas: una fase de acotamiento encuentra un intervalo que contiene longitudes de paso deseables, y una fase de bisección o interpolación la cual calcula una buena longitud de paso dentro de este intervalo.

Ahora discutimos varias condiciones de terminación para los algoritmos de búsqueda de línea y mostramos que los pasos eficaces de longitud no necesitan encontrarse cerca de minimizadores de la función univariada $\phi(\alpha)$ definida en (4.3).

Una condición simple que podríamos imponer a α_k es exigir una reducción de f, es decir,

$$f(x_k + \alpha p_k) < f(x_k).$$

Pero este requisito no es suficiente para producir la convergencia de x^* que es ilustrado en la Figura (4.2), para los que el valor de la función mínima es de $f^* = -1$, pero una secuencia de iteraciones x_k para los cuales $f(x_k) = 5/k$, $k = 0, 1, ...$ produce una disminución en cada iteración, pero tiene un valor de función limitadora en cero. La reducción insuficiente en f en cada paso hace que deje de converger hacia el minimizador de esta función convexa. Para evitar este comportamiento hay que cumplir una condición de disminución suficiente, concepto que discutiremos a continuación.

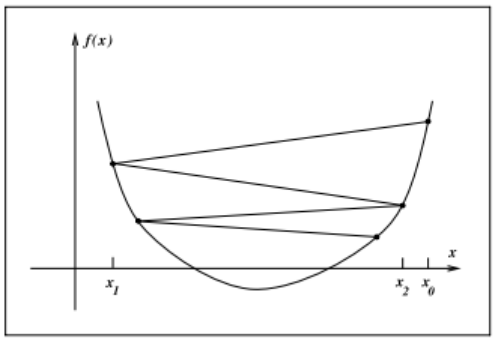


Figura 4.2: *Reducción insuficiente en f*

Las condiciones de Wolfe

Una condición de búsqueda de línea inexacta popular estipula que α_k en primer lugar debe dar disminución suficiente de la función objetivo f, medida por la siguiente desigualdad:

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k, \tag{4.4}$$

para alguna constante $c_1 \in (0, 1)$. En otras palabras, la reducción de f debe ser proporcional tanto a la longitud de paso α_k y a la derivada direccional $\nabla f_k^T p_k$. La desigualdad (4.4) se llama a veces la condición Armijo.

La condición de disminución suficiente se ilustra en la Figura (4.3). El lado derecho de (4.4), que es una función lineal, puede denotarse por $l(\alpha)$. La función $l(\cdot)$ tiene pendiente negativa $c_1 \nabla f_k^T p_k$, pero porque $c_1 \in (0, 1)$, se encuentra por encima de la gráfica de ϕ para pequeños valores positivos de α. La condición de disminución suficiente es que α es aceptable sólo si $\phi(\alpha) \leq l(\alpha)$. Los intervalos en los que se satisface esta condición se muestran en la Figura (4.3). En la práctica, se elige c_1 bastante pequeña, es decir $c_1 = 10^{-4}$.

La condición de disminución suficiente no es suficiente por sí misma para garantizar que el algoritmo hace un progreso razonable, ya que, como vemos en la Figura (4.3), que se satisface para todos los valores de α suficientemente pequeños. Para descartar pasos inaceptablemente cortos introducimos un segundo requisito, llamada la condición de curvatura, lo que requiere para satisfacer α_k :

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k,$$

para alguna constante $c_2 \in (c_1, 1)$, donde c_1 es la constante a partir de (4.4). Tenga en cuenta que el lado izquierdo es simplemente la derivada $\phi'(\alpha_k)$, por lo que la condición de curvatura asegura que la pendiente de ϕ en α_k es mayor que los tiempos de c_2 en la pendiente inicial $\phi'(0)$. Esto tiene sentido porque si la pendiente $\phi'(\alpha)$ es fuertemente negativa, tenemos una indicación de que podemos reducir significativamente f moviendo más allá en la dirección elegida.
Por otro lado, si $\phi(\alpha_k)$ es sólo ligeramente negativa o incluso positiva, es una señal de que no podemos esperar mayor disminución de f en esta dirección, así que tiene sentido para terminar la búsqueda de línea. La condición de curvatura se ilustra en la Figura (4.4). Los

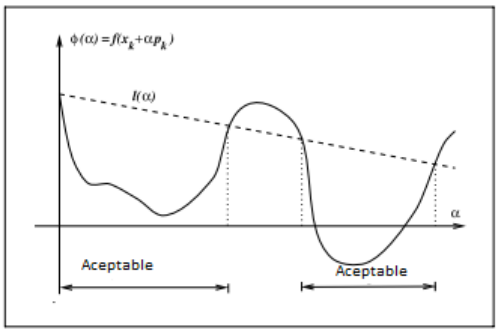


Figura 4.3: *Condición de disminución suficiente.*

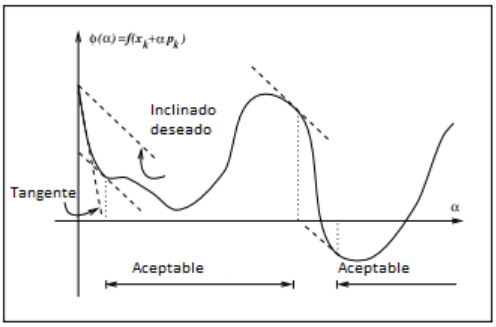


Figura 4.4: *La condición de curvatura.*

valores típicos de c_2 son 0.9 cuando la dirección de búsqueda p_k es elegida por un método de Newton o cuasi-Newton , y 0.1 cuando p_k se obtiene a partir de un método de gradiente conjugado no lineal.

Las condiciones de disminución y de curvatura suficientes se conocen colectivamente como las condiciones Wolfe ilustradas en la Figura (4.1), definidas por:

$$\begin{aligned} f(x_k + \alpha_k p_k) &\leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k, \\ \nabla f(x_k + \alpha_k p_k)^T p_k &\geq c_2 \nabla f_k^T p_k \end{aligned} \tag{4.5}$$

con $0 < c_1 < c_2 < 1$.

Una longitud de paso puede satisfacer las condiciones Wolfe sin estar particularmente cerca de un minimizador de ϕ como mostramos en la Figura (4.1). Podemos, sin embargo, modificar la condición de curvatura para forzar a α_k estar sobre al menos un amplio entorno de un minimizador local o punto de ϕ estacionario. Las fuertes condiciones Wolfe requieren a α_k para satisfacer

$$\begin{aligned} f(x_k + \alpha_k p_k) &\leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k, \\ |\nabla f(x_k + \alpha_k p_k)^T p_k| &\geq |c_2 \nabla f_k^T p_k|, \end{aligned} \tag{4.6}$$

con $0 < c_1 < c_2 < 1$.
No es difícil demostrar que existen longitudes de paso que cumplan las condiciones Wolfe para cada función f, que es suave y limitada por debajo.

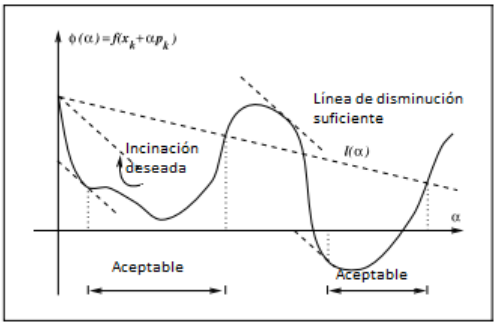


Figura 4.5: *Longitudes de paso que cumplan las condiciones Wolfe.*

Lema 1 *: Suponer que $f : \mathbb{R}^n \to \mathbb{R}$ es continuamente diferenciable y sea p_k una dirección de descenso en x_k; asumir que f está acotada inferiormente a lo largo del rayo $\{x_k + \alpha p_k | \alpha > 0\}$. Entonces, si $0 < c_1 < c_2 < 1$, existen intervalos de longitudes de paso que satisfacen las condiciones Wolfe (4.5) y las condiciones fuertes Wolfe (4.6).*

Prueba:
Notar que $\phi(\alpha) = f(x_k + \alpha p_k)$ está acotada por debajo de todo $\alpha > 0$. Como $0 < c_1 < 1$, la línea $l(\alpha) = f(x_k) + \alpha c_1 \nabla f_k^T p_k$ es limitada por debajo y por lo tanto debe intersecar la gráfica de ϕ al menos una vez. Dejar que $\alpha > 0$ sea el valor más pequeño de intersección de α, es decir,

$$\nabla f(x_k + \alpha^{'} p_k) = f_k + \alpha^{'} c_1 \nabla f_k^T p_k. \tag{4.7}$$

La condición de disminución suficiente (4.5) es claramente para todas las longitudes de paso menores de $\alpha^{'}$. Por el teorema del valor medio, existe $\alpha^{''} \in (0, \alpha^{'})$ tal que

$$\nabla f(x_k + \alpha^{'} p_k) - f_k = \alpha^{'} \nabla f(x_k + \alpha^{''} p_k)^T p_k. \tag{4.8}$$

Mediante la combinación de (4.5) y (4.6), obtenemos

$$\nabla f(x_k + \alpha^{''} p_k)^T p_k = c_1 \nabla f_k^T p_k > c_2 \nabla f_k^T p_k. \tag{4.9}$$

Esto cuando $c_1 < c_2$ y $\nabla f_k^T p_k < 0$. Por lo tanto, $\alpha^{''}$ satisface las condiciones Wolfe (4.5), y las desigualdades sujetan estrictamente en ambas ecuaciones. Por lo tanto, por nuestra suposición de suavidad en f, hay un intervalo alrededor de $\alpha^{''}$ para el cual las condiciones Wolfe se cumplen. Además, dado el término negativo en el lado izquierdo de (4.9), las condiciones fuertes de Wolfe (4.5) están en el mismo intervalo. Lo que demuestra lo requerido.

□

Las condiciones Wolfe son invariante en escala en un sentido amplio: Multiplicar la función objetivo por una constante o hacer un cambio afín de variables no les altera. Pueden ser utilizados en la mayoría de los métodos de búsqueda de línea, y son especialmente importantes en la implementación de métodos cuasi-Newton.

Las condiciones de Goldstein.

Al igual que las condiciones de Wolfe, las condiciones Goldstein aseguran que la longitud

de paso α logra suficiente disminución, pero no es demasiado corto. Las condiciones Goldstein también pueden expresarse como un par de desigualdades de la siguiente manera:

$$\nabla f(x_k) + (1-c)\alpha_k \nabla f_k^T p_k \leq f(x_k + \alpha_k p_k) \leq f(x_k) + c\alpha_k \nabla f_k^T p_k \tag{4.10}$$

con $0 < c < 1/2$.

La segunda desigualdad es la condición de disminución suficiente (4.4), mientras que la primera desigualdad se introduce para controlar la longitud del paso desde abajo; ver Figura (4.1). Una desventaja de las condiciones Goldstein a las condiciones Wolfe es que la primera desigualdad en (4.10) puede excluir todos los minimizadores de ϕ. Sin embargo, las condiciones de Goldstein y Wolfe tienen mucho en común y sus teorías de convergencia son bastante similares. Las condiciones Goldstein se utilizan a menudo en métodos de tipo Newton, pero no son muy adecuadas para los métodos cuasi-Newton que mantienen una aproximación del Hessiano definida positiva.

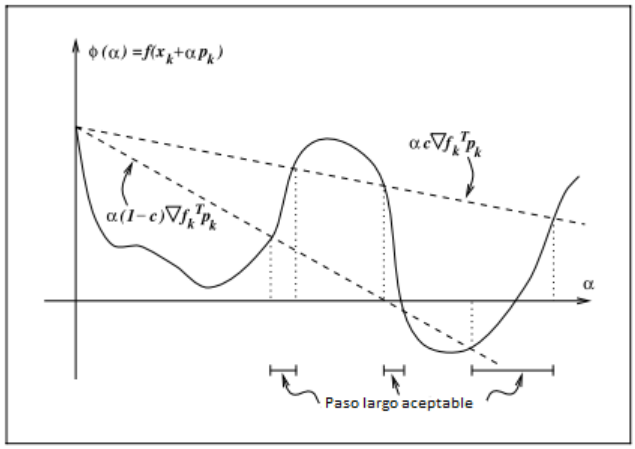


Figura 4.6: *Las condiciones de Goldstein.*

Ya hemos mencionado que la primera condición de disminución suficiente en las ecuaciones de (4.5) por sí sola no es suficiente para asegurar que los algoritmos hacen un progreso razonable a lo largo de la dirección de búsqueda dada. Sin embargo, si el algoritmo de búsqueda de línea elige su candidato paso de longitudes adecuadamente, mediante el uso de un enfoque llamado vuelta atrás, podemos prescindir de la segunda condición adicional de la ecuación (4.5) y utilizar sólo la condición de disminución suficiente para dar por terminado el procedimiento de búsqueda de línea. En su forma más básica, el retroceso procede como sigue.

Algoritmo 1 (Retroceso de línea de busqueda).
Elige $\bar{\alpha} > 0, \rho \in (0,1), c\epsilon(0,1)$; sea $\alpha \leftarrow \bar{\alpha}$;
Repetir hasta $f(x_k + \alpha p_k) \leq f(x_k) + c\alpha \nabla f_k^T p_k$
$\alpha \longrightarrow \rho\alpha$;
terminar(repetir)
Terminar con $\alpha_k = \alpha$.

En este procedimiento, el paso de longitud inicial $\bar{\alpha}$ es elegido para ser 1 en el método Newton y métodos cuasi-Newton, pero pueden tener valores diferentes en otros algoritmos como descenso máximo o gradiente conjugado. Una longitud de paso aceptable se encontrará después de un número finito de iteraciones, porque α_k eventualmente se convertirá en lo suficientemente pequeño que la condición de disminución suficiente posee (véase la Figura (4.3)). En la práctica, los factores de contracción ρ se permiten a menudo variar en cada iteración de la línea de búsqueda. Por ejemplo, puede ser elegido por interpolación cuidadosa, como se describe más adelante. Necesitamos garantizar sólo que en cada iteración tenemos $\rho \in [\rho_{\text{lo}}, \rho_{\text{hi}}]$, para algunas constantes fijas $0 < \rho_{\text{lo}} < \rho_{\text{hi}} < 1$.

El enfoque de retroceso garantiza que la longitud de paso α_k seleccionado es un valor fijo (la elección inicial $\bar{\alpha}$), o que es lo suficientemente corto para satisfacer la condición de disminución suficiente, pero no demasiado corto. La última afirmación sostiene porque el valor aceptado α_k esté dentro de un factor ρ del valor anterior α_k/ρ, que fue rechazado por violar la condición disminución suficiente, es decir, por ser demasiado largo.

Esta estrategia simple y popular para la terminación de una línea de búsqueda es muy adecuado para los métodos de Newton, pero es menos apropiado para cuasi-Newton y métodos de gradiente conjugado.

4.2. Convergencia de Métodos de Búsqueda de Línea.

Para obtener la convergencia global, debemos no sólo elegir bien la longitud del paso α_k, sino también elegir bien la búsqueda de direcciones p_k. Discutimos requisitos en la dirección de búsqueda en esta sección, centrándose en una propiedad clave: tomando θ_k como el ángulo entre p_k y el sentido de descenso más agudo ∇f_k, definido por:

$$\cos\theta_k = \frac{-\nabla f_k^T p_k}{\|\nabla f_k\|\|p_k\|}. \tag{4.11}$$

El siguiente teorema, debido a Guus Zoutendijk, tiene consecuencias de largo alcance. Se cuantifica el efecto de la etapa adecuadamente eligiendo longitudes α_k y muestra, por ejemplo, que el método del descenso máximo es globalmente convergente. Para otros algoritmos, describe hasta qué punto p_k puede desviarse de la dirección del descenso máximo y todavía producir una iteración global convergente. Varias condiciones de terminación de búsqueda de línea pueden ser utilizadas para establecer este resultado, pero por lo concreto vamos a considerar sólo las condiciones Wolfe (4.5). Aunque el resultado de Zoutendijk aparece a primera vista técnico y oscuro, su poder será pronto evidente.

Teorema 11 *: Considere cualquier iteración de la forma (4.1), donde p_k es una dirección de descenso y α_k satisface las condiciones Wolfe (4.5). Suponer que f está acotada inferiormente en $\mathbb{R}^n$ y que f es diferenciablemente continua en un conjunto abierto que contiene $\mathcal{N}$, el conjunto de nivel $\mathcal{L} \overset{def}{=} \{x : f(x) \leq f(x_0)\}$; donde x_0 es el punto de partida de la iteración. Suponer también que el gradiente ∇f es Lipschitz continua en $\mathcal{N}$, es decir, existe una constante $L > 0$ tal que*

$$\|\nabla f(x) - \nabla f(\tilde{x})\| \leq L\|x - \tilde{x}\|, \qquad \textit{para todo} \qquad x, \tilde{x} \in \mathcal{N}. \tag{4.12}$$

entonces

$$\sum_{k\geq 0} \cos^2\theta_k \|\nabla f_k\|^2 < \infty. \tag{4.13}$$

Prueba:
De la segunda condición Wolfe (4.5) y (4.1) tenemos que

$$(\nabla f_{k+1} - \nabla f_k)^T p_k \geq (c_2 - 1)\nabla f_k^T p_k,$$

mientras que la condición de Lipschitz (4.12) implica que $(\nabla f_{k+1} - \nabla f_k)^T p_k \leq \alpha_k L \|p_k\|^2$. Mediante la combinación de estas dos relaciones, obtenemos

$$\alpha_k = \frac{(c_2 - 1)}{L} \frac{\nabla f_k^T p_k}{\|p_k\|^2}.$$

Mediante la sustitución de esta desigualdad en la primera condición Wolfe (4.5), obtenemos

$$f_{k+1} \leq f_k - c_1 \frac{1 - c_2}{L} \frac{(\nabla f_k^T p_k)^2}{\|p_k\|^2}.$$

De la definición (4.11), podemos escribir esta relación como $f_{k+1} \leq f_k - c\cos^2\theta_k \|\nabla f_k\|^2$, donde $c = c_1(1 - c_2)/L$.

Sumando esta expresión sobre todos los índices de menos o igual a α_k, obtenemos

$$f_{k+1} \leq f_k - c \sum_{j=0}^{k} \cos^2\theta_j \|\nabla f_j\|^2.$$

Dado que f está acotada inferiormente, tenemos que $f_0 - f_{k+1}$ es menor que una constante positiva, para todo k, por lo tanto tomando limites en la ecuación anterior, obtenemos

$$\sum_{k=0}^{k} \cos^2\theta_k \|\nabla f_k\|^2 < \infty,$$

lo que concluye la prueba.

□

Resultados similares a este teorema se espera cuando se utilizan las condiciones Goldstein (4.10) o fuertes condiciones Wolfe (4.6) en lugar de las condiciones Wolfe. Por todas estas estrategias, de selección de la longitud de paso implica la desigualdad (4.13), lo que llamamos la condición Guus Zoutendijk. Tenga en cuenta que los supuestos del Teorema (11) no son demasiado restrictivas. Si la función f no está acotada inferiormente el problema de optimización no estaría bien definido. La hipótesis de suavidad (continuidad Lipschitz del gradiente) está implícita en muchas de las condiciones de suavidad que se utilizan en los teoremas de convergencia local y son a menudo satisfechas en la práctica.

La condición Guus Zoutendijk (4.13) implica que

$$\cos^2\theta_k \|\nabla f_k\|^2 \to 0. \tag{4.14}$$

Este límite se puede utilizar a su vez para obtener resultados de convergencia global para los algoritmos de búsqueda de línea. Si nuestro método para elegir la dirección de búsqueda p_k en la iteración (4.1) asegura que el ángulo θ_k definido por (4.11) está delimitado lejos de 90 grados, hay una constante positiva δ tal que

$$\cos\theta_k \geq \delta > 0 \qquad \text{para todo} \qquad k.$$

De ello se deduce inmediatamente a partir de (4.14) que

$$\lim_{k\to\infty} \|\nabla f_k\| = 0. \tag{4.15}$$

En otras palabras, podemos estar seguros de que las normas de gradiente $\|\nabla f_k\|$ convergen a cero, siempre que las direcciones de búsqueda nunca estén demasiado cerca de la ortogonalidad con el gradiente. En particular, el método del descenso máximo (para el que la dirección de búsqueda p_k es paralela a la negativa del gradiente) produce una secuencia de gradientes que convergen a cero, siempre que se utiliza una línea de búsqueda que satisfaga las condiciones Wolfe o Goldstein.

Usamos el término globalmente convergente para referirse a los algoritmos para el que concurre la propiedad (4.15), pero tenga en cuenta que este término se utiliza a veces en otros contextos para significar diferentes cosas. Para los métodos de búsqueda de línea de la forma general (4.1), el límite de (4.15) es el más fuerte resultado global de la convergencia que se puede obtener. No podemos garantizar que el método converge a un minimizador, sino sólo que se siente atraído por puntos estacionarios. Sólo haciendo requisitos adicionales en la dirección de búsqueda p_k para introducir información negativa para la curvatura del Hessiano $\nabla^2 f(x_k)$, por ejemplo podemos reforzar estos resultados para incluir la convergencia a un mínimo local.

Consideremos ahora el método de Newton como (4.1), (4.2) y se asume que las matrices B_k son definidas positivas con un número de condición uniformemente acotadas. Es decir, no es una constante M tal que

$$\|B_k\|\|B_k^{-1}\| \leq M, \qquad \text{para todo} \qquad k.$$

Es fácil demostrar a partir de la definición (4.11) que

$$\cos\theta_k \geq 1/M.$$

Mediante la combinación de esta cota con (4.14) encontramos que

$$\lim_{k\to\infty} \|\nabla f_k\| = 0. \tag{4.16}$$

Por lo tanto, hemos demostrado que los métodos de Newton y cuasi-Newton son globalmente convergente si las matrices B_k tiene un número de condición limitadas y son definidas positivas (lo que se necesita para asegurar que p_k sea una dirección de descenso), y si las longitudes de paso satisfacen la condiciones Wolfe.

Para algunos algoritmos, como los métodos de gradiente conjugado, van a ser capaces de probar el límite (4.15), pero sólo el resultado más débil

$$\liminf_{k\to\infty} \|\nabla f_k\| = 0. \tag{4.17}$$

En otras palabras, sólo una subsecuencia de las normas de gradiente $\|\nabla f_{k_j}\|$ converge a cero, en lugar de toda la secuencia. Este resultado también se puede demostrar mediante el uso de la condición de Zoutendijk (4.13), pero en lugar de una prueba constructiva, esbozamos una prueba por contradicción. Supongamos que (4.16) no se cumple, de manera que los gradientes permanecen limitados lejos de cero, es decir, existe $\gamma > 0$ tal que $\|\nabla f_k\| \geq \gamma$, para k suficientemente grande. Entonces a partir de (4.14) se concluye que $\cos\theta_k \to 0$, es decir, toda la secuencia $\cos\theta_k$ converge a 0. Para establecer (4.16), por lo tanto, es suficiente mostrar que una subsecuencia $\cos\theta_{k_j}$ está acotada lejos de cero.

Mediante la aplicación de esta técnica a prueba, podemos probar la convergencia global en el sentido de (4.16) o (4.18) para una clase general de algoritmos. Considere cualquier algoritmo para el que *i)* cada iteración produzca una disminución en la función objetivo, y *ii)* cada m iteración es un paso del descenso máximo, con la longitud del paso escogido para satisfacer las condiciones de Wolfe o Goldstein. Entonces, puesto que $\cos\theta_k = 1$ para los pasos gradiente descendente, el resultado (4.18) se mantiene. Por supuesto, podríamos diseñar el algoritmo para que haga algo "mejor"que el descenso más pronunciado en el otro $m-1$ iteraciones. Los ocasionales pasos gradientes descendentes no pueden avanzar mucho, pero al menos garantiza la convergencia global.

Tenga en cuenta que en toda esta sección hemos utilizado solamente el hecho de que la condición del Zoutendijk implica el límite (4.14). En capítulos posteriores se hará uso de la condición de suma acotada (4.13), lo que obliga a la sucesión $\cos^2\theta_k\|\nabla f_k\|^2$ a converger a cero a una velocidad suficientemente rápida.

4.3. Tasa de Convergencia.

Parecería que el diseño de algoritmos de optimización con buenas propiedades de convergencia es fácil, ya que todo lo que necesitamos para asegurarla es que la dirección de búsqueda p_k no tienda a ser ortogonal al gradiente ∇f_k, o que pasos gradiente descendente se tomen con regularidad. Podríamos simplemente calcular $\cos\theta_k$ en cada iteración y volvemos p_k hacia la dirección del descenso más agudo si $\cos\theta_k$ es más pequeño que algunos preseleccionados $\delta > 0$. La prueba de ángulo constante de este tipo asegura la convergencia global, pero son deseables por dos razones. En primer lugar, pueden impedir una rápida tasa de convergencia, porque el Hessiano puede estar mal condicionado y puede ser necesario para producir direcciones de búsqueda que son casi ortogonales a la pendiente o una elección inapropiada del parámetro δ puede causar dichas medidas a ser rechazadas. En segundo lugar, las pruebas de ángulo destruyen las propiedades de invariancia de métodos cuasi-Newton.

Estrategias algorítmicas que logran una rápida convergencia a veces pueden estar en conflicto con los requisitos de convergencia global, y viceversa. Por ejemplo, el método del descenso máximo es el algoritmo convergente global por excelencia, pero es bastante lento en la práctica, como veremos a continuación. Por otro lado, la iteración Newton converge rápidamente cuando se inicia lo suficientemente cerca de una solución, pero sus pasos puede incluso no ser direcciones de descenso fuera de la solución. El reto es diseñar algoritmos que incorporan ambas propiedades: buenas garantías de convergencia globales y un rápido ritmo de convergencia.

Comenzamos nuestro estudio de las tasas de convergencia de los métodos de búsqueda de línea considerando el enfoque más básico de todos: el método del descenso máximo.

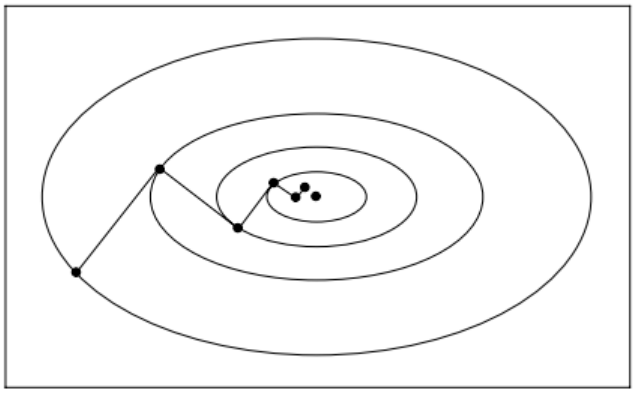


Figura 4.7: *Escalones más empinados de descenso.*

Tasa de convergencia y velocidad del descenso

Podemos aprender mucho sobre el método del descenso máximo considerando el caso ideal, en el que la función objetivo es cuadrática y los métodos de búsqueda de línea son exactos. Supongamos que

$$f(x) = \frac{1}{2}x^T Qx - b^T x, \tag{4.18}$$

Donde Q es simétrica y definida positiva. El gradiente está dada por $\nabla f(x) = Qx - b$ y el minimizador x^* es la única solución del sistema lineal $Qx = b$.

Es fácil calcular la longitud del paso α_k que minimice $f(x_k - \alpha\nabla f_k)$. Al diferenciar la función

$$f(x_k - \alpha\nabla f_k) = \frac{1}{2}(x_k - \alpha\nabla f_k)^T Q(x_k - \alpha\nabla f_k) - b^T(x_k - \alpha\nabla f_k)$$

con respecto a α, y el establecimiento de la derivada a cero, obtenemos

$$\alpha_k = \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k}.$$

Si utilizamos este α_k minimizador exacto, la iteración máxima pendiente para (4.18) viene dada por

$$x_{k+1} = x_k - \left(\frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k}\right)\nabla f_k. \tag{4.19}$$

Donde $\nabla f_k = Qx_k - b$, se obtiene como una expresión de forma cerrada para x_{k+1} en términos de x_k. En la Figura (4.3), se traza una secuencia típica de iteraciones generadas por el método del descenso máximo en una función objetivo cuadrática bidimensional. Los contornos de f son elipsoides cuyos ejes se encuentran a lo largo de los vectores propios ortogonales de Q. Para cuantificar la tasa de convergencia introducimos la norma ponderada $\|x\|_Q^2 = x^T Qx$. Mediante el uso de la relación $Qx^* = b$, podemos demostrar que

$$\frac{1}{2}\|x - x^*\|_Q^2 = f(x) - f(x^*), \tag{4.20}$$

por lo que esta norma mide la diferencia entre el valor objetivo actual y el valor óptimo. Mediante el uso de la igualdad (4.19) y observando que podemos derivar la igualdad dada por

$$\|x_{k+1} - x^*\|_Q^2 = \left\{1 - \frac{(\nabla f_k^T \nabla f_k)^2}{(\nabla f_k^T Q \nabla f_k)(\nabla f_k^T Q^{-1} \nabla f_k)}\right\} \|x_k - x^*\|_Q^2. \tag{4.21}$$

Esta expresión describe la disminución exacta de f en cada iteración, pero dado que el término dentro del paréntesis es difícil de interpretar, es más útil el límite en términos del número de condición del problema.

Teorema 12 . *Cuando el método del descenso máximo con la línea exacta de búsqueda (4.19) se aplican a la función cuadrática fuertemente convexa (4.18), la norma de error (4.20) satisface*

$$\|x_{k+1} - x^*\|_Q^2 \leq \left\{\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right\}^2 \|x_k - x^*\|_Q^2, \tag{4.22}$$

donde $0 < \lambda_1 < \lambda_2 < .... < \lambda_n$ *son los eigenvalores de* Q.

Las desigualdades (4.22) y (4.21) muestran que la función de los valores de f_k convergen al mínimo de f_* a una velocidad lineal. Como un caso especial de este resultado, vemos que la convergencia se logra en una iteración si todos los valores propios son iguales. En este caso, Q es un múltiplo de la matriz de identidad, por lo que los contornos en la Figura (4.3), son círculos y la dirección del descenso más agudo siempre apunta a la solución. En general, como el número de condición $k(Q) = \lambda_n/\lambda_1$ aumenta, los contornos de la cuadrática son más alargados y (4.22) implica que se degrada la convergencia. A pesar de que (4.22) es el peor de los casos unidos, da una indicación precisa del comportamiento del algoritmo cuando $n > 2$.

El comportamiento del método del descenso máximo de la velocidad de convergencia es esencialmente el mismo en funciones objetivas lineales generales. En el siguiente resultado asumimos que la longitud del paso es el minimizador global a lo largo de la dirección de búsqueda.

Teorema 13 . *Supongamos que* $f : \mathbb{R}^n \to \mathbb{R}$ *es dos veces continuamente diferenciable y que las iteraciones generadas por el método del descenso máximo con búsquedas exactas de líneas convergen en un punto* x^* *en el que la matriz Hessiana* $\nabla^2 f(x^*)$ *es definida positiva. Sea* r *cualquier escalar que satisface*

$$r \in \left\{\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}, 1\right\},$$

donde $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ *son los eigenvalores de* $\nabla^2 f(x^*)$. *Entonces, para todo* k *suficientemente grande, tenemos*

$$f(x_k + 1) - f(x^*) \leq r^2[f(x_k) - f(x^*)].$$

En general, no podemos esperar que para mejorar la tasa de convergencia se utilice una línea de búsqueda inexacta. Por lo tanto, el Teorema (13) muestra que el método del descenso máximo puede tener una tasa inaceptablemente lenta de convergencia, incluso cuando la matriz Hessiana está razonablemente bien condicionado. Por ejemplo, si $k(Q) = 800$ $f(x_1) = 1$, y $f(x^*) = 0$, el Teorema (13) sugiere que el valor de la función seguirá siendo aproximadamente 0.08 después de mil iteraciones del método del descenso más agudo con la búsqueda de línea exacta.

Capítulo 5

Métodos de Región Factible

Métodos de búsqueda de línea y métodos de región factible, generan pasos con la ayuda de un modelo cuadrático de la función objetivo, pero que utilizan este modelo de diferentes maneras. Métodos de búsqueda de línea lo utilizan para generar un sentido de búsqueda y a continuación, centrar sus esfuerzos en encontrar la longitud de paso α adecuada a lo largo de esta dirección. Métodos de región factible definen una región alrededor de la iteración actual en la que confían que el modelo es una representación adecuada de la función objetivo y luego eligen el paso a ser el minimizador aproximado del modelo en esta región. En efecto, eligen la dirección y la longitud del paso simultáneamente. Si un paso no es aceptable, reducen el tamaño de la región y encuentran un nuevo minimizador. En general, la dirección de la etapa cambia cada vez que se altera el tamaño de la región factible.

El tamaño de la región factible es fundamental para la eficacia de cada paso. Si la región es demasiado pequeña, el algoritmo pierde la oportunidad de dar un paso importante que se moverá mucho cerca del minimizador de la función objetivo. Si es demasiado grande, el minimizador del modelo puede estar lejos del minimizador de la función objetivo en la región, por lo que debe reducirse el tamaño de la región y volverse a intentar. En los algoritmos prácticos, elegimos el tamaño de la región de acuerdo con el desempeño del algoritmo durante las iteraciones anteriores. Si el modelo es siempre fiable, da buenos pasos y predice con precisión el comportamiento de la función objetivo a lo largo de estos pasos, el tamaño de la región factible se puede aumentar.

Vamos a suponer que la función modelo m_k que se utiliza en cada iteración x_k es cuadrática. Por otra parte, m_k se basa en la expansión de la serie de Taylor de f alrededor de x, que es

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

donde $f_k = f(x_k)$ y $g_k = \bigtriangledown f(x_k)$, y t es algún escalar en el intervalo (0,1). Mediante el uso de una aproximación de la matriz Hessiana B_k en el término de segundo orden, m_k se define de la siguiente manera:

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

donde B_k es alguna matriz simétrica. La diferencia entre $m_k(p)$ y $f(x_k + p)$ es $O(\|p\|^2)$, que es pequeño cuando p es pequeño.

Cuando B_k es igual a la matriz hessiana $\bigtriangledown^2 f(x_k)$, el error de aproximación en el modelo de la función m_k es $O(\|p\|^3)$, por lo que este modelo es especialmente preciso cuándo $\|p\|$ es pequeña.

Para obtener cada paso, buscamos una solución del subproblema

$$\min_{p\in\mathbb{R}^n} m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p \;\text{ sujeto a } \|p\| \leq \triangle_k, \tag{5.1}$$

donde $\triangle_k > 0$ es el radio de la región factible. En la mayoría de nuestras discusiones, definimos $\|\cdot\|$ ser la norma euclídea, de manera que la solución p_k^* de la ecuación anterior es el minimizador de m_k en la bola de radio $\triangle_k$. Por lo tanto, el enfoque de la región factible nos obliga a resolver una serie de subproblemas en la que la función objetivo y las restricciones (que se pueden escribir como $p^T p \leq \triangle_k^2$) son ambas cuadráticas.

Enfoque de región factible

Uno de los ingredientes clave en un algoritmo de región factible es la estrategia para la elección del radio $\triangle_k$ en cada iteración. Basamos esta elección por acuerdo entre la función del modelo m_k y la función objetivo f en iteraciones anteriores. Dado un paso p_k definimos el radio

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}; \tag{5.2}$$

donde el numerador se llama reducción real y el denominador es la reducción prevista (es decir, la reducción de f predicha por la función de modelo). Tener en cuenta que desde el paso p_k se minimiza el modelo m_k sobre una región que incluye $p = 0$, la reducción prevista será siempre no negativa. Por lo tanto, si ρ_k es negativo, el nuevo valor objetivo $f(x_k + p_k)$ es mayor que el valor actual $f(x_k)$, por lo que el paso debe ser rechazado. Por otro lado, si ρ_k es cercano a 1, existe una buena concordancia entre el modelo m_k y la función f en este paso, lo que es seguro para expandir la región factible para la próxima iteración. Si ρ_k es significativamente positivo pero menor que 1, no alteramos la región factible, pero si es cercana a cero o negativa, encogemos la región mediante la reducción de $\triangle_k$ en la siguiente iteración.

El siguiente algoritmo describe el proceso.

ALGORITMO 5.1 (Región Factible).
Dado $\hat{\triangle} > 0$, $\triangle_0 \in (0, \hat{\triangle})$, y $\eta \in [0, \frac{1}{4})$:
for $k = 1, 2, \ldots$
  Obtener p_k (aproximadamente) resolviendo (5.1);
  Evaluar ρ_k en (5.2);
  if $\rho_k < \frac{1}{4}$
    $\triangle_{k+1} = \frac{1}{4}\triangle_k$
  else
    if $\rho_k > \frac{3}{4}$ y $\|p_k\| = \triangle_k$
      $\triangle_{k+1} = \min(2\triangle_k, \hat{\triangle})$
    else
      $\triangle_{k+1} = \triangle_k$;

if $\rho_k > \eta$
  $x_{k+1} = x_k + p_k$
else
  $x_{k+1} = x_k$;
end (for).

Donde $\hat{\triangle}$ es un conjunto unido de las longitudes de paso. Teniendo en cuenta que el radio se incrementa sólo si $\|p_k\|$ alcanza el límite de la región factible. Si el paso se mantiene estrictamente dentro de la región, se infiere que el valor actual de $\triangle_k$ no está interfiriendo con el progreso del algoritmo, por lo que dejamos su valor sin cambios para la siguiente iteración.

Para convertir el algoritmo 5.1 en un algoritmo práctico, tenemos que centrarnos en resolver el subproblema de región factible (5.1). En la discusión de este asunto, abandonamos el subíndice de iteración k y reiteramos el problema (5.1) de la siguiente manera:

$$\min_{p\in\mathbb{R}^n} m(p) \stackrel{def}{=} f + g^T p + \frac{1}{2} p^T B p \;\text{ sujeto a } \|p\| \leq \triangle. \tag{5.3}$$

Un primer paso para caracterizar las soluciones exactas de (5.3) viene dada por el siguiente Teorema.

Teorema 14 *:*
El vector p^* es solución global del problema de región factible

$$\min_{p\in\mathbb{R}^n} m(p) = f + g^T p + \frac{1}{2} p^T B p \;\; \textit{sujeto a } \|p\| \leq \triangle,$$

si y sólo si p^* es factible y hay un escalar $\lambda \geq 0$ tal que las siguientes condiciones se cumplen:

$$(B + \lambda I)p^* = -g,$$

$$\lambda(\triangle - \|p^*\|) = 0,$$

$$(B + \lambda I) \textit{ es semidefinido positivo.}$$

5.1. Algoritmos Basados en el Punto Cauchy

El punto Cauchy

Como vimos en el capítulo anterior, los métodos de búsqueda de línea pueden ser convergentes a nivel global, incluso cuando la longitud óptima de paso no se utiliza en cada iteración. De hecho, la longitud de paso α_k sólo debe satisfacer criterios suaves. Una situación similar se aplica en los métodos de región factible. Aunque, en principio, buscamos la solución óptima del subproblema (5.1), es suficiente para los propósitos de convergencia global encontrar una solución aproximada p_k que se encuentra dentro de la región factible y de una reducción suficiente en el modelo. La reducción suficiente se puede cuantificar en términos del punto Cauchy, que denotamos por p_k^c y definimos en términos del siguiente procedimiento simple.

ALGORITMO 5.2 (Calculando el Punto Cauchy)
Encontrar el vector p_k^s que resuelve la versión lineal de (5.1), tal que,

$$p_k^s = \arg\min_{p\in\mathbb{R}^n} f_k + g_k^T p \quad \text{sujeto a } \|p\| \leq \triangle_k;$$

Calcular el escalar $\tau_k > 0$ que minimiza $m_k(\tau p_k^s)$ sujeto a satisfacer la región factible, tal que

$$\tau_k = \arg\min_{\tau\geq 0} m_k(\tau p_k^s) \quad \text{sujeto a } \|\tau p_k^s\| \leq \triangle_k;$$

donde $p_k^c = \tau_k p_k^s$.

Es fácil escribir una definición de forma cerrada del punto de Cauchy. Para empezar, la solución es simplemente

$$p_k^s = -\frac{\triangle_k}{\|g_k\|} g_k.$$

Para obtener τ_k explícitamente, consideramos los casos de $g_k^T B_k g_k \leq 0$ y $g_k^T B_k g_k > 0$ por separado. Para el primer caso, la función $m_k(\tau p_k^s)$ disminuye monótonamente con τ cuando $g_k \neq 0$, por lo que τ_k es simplemente el valor más grande que satisface la región factible vinculada con $\tau_k = 1$. Para el caso $g_k^T B_k g_k > 0, m_k(\tau p_k^s)$ es cuadrática convexa en τ, por lo que τ_k es o bien el minimizador sin restricciones de esta cuadrática o el valor límite de 1, que ocurra primero. En resumen, tenemos

$$p_k^c = -\tau_k \frac{\triangle_k}{\|g_k\|} g_k,$$

donde

$$\tau_k = \begin{cases} 1 & \text{si } g_k^T B_k g_k \leq 0; \\ \text{mín}(\|g_k\|^3/(\triangle_k g_k^T B_k g_k), 1) & \text{otro caso.} \end{cases}$$

El paso Cauchy es computacionalmente barato, para calcular la matriz no se requieren factorizaciones y es de importancia crucial para decidir si una solución aproximada de la región factible del subproblema es aceptable. Específicamente, un método de región factible será globalmente convergente si sus pasos p_k dan una reducción en el modelo m_k que es al menos algún múltiplo positivo fijo de la disminución alcanzada por el paso Cauchy.

Mejora en el punto Cauchy

El punto Cauchy no depende fuertemente de la matriz B_k, que se utiliza sólo en el cálculo de la longitud de paso. Convergencia rápida se puede esperar sólo si B_k juega un papel en la determinación de la dirección de la etapa, así como su longitud y si B_k contiene información válida acerca de la curvatura de la función.

Una serie de algoritmos de región factible calculan el punto Cauchy y luego tratan de mejorarlo. La estrategia de mejora suele diseñarse para que se elija $p_k^B = -B_k^{-1} g_k$ siempre que B_k es definida positiva y cuando B_k es la matriz Hessiana exacta $\bigtriangledown^2 f(x_k)$ o una aproximación Cuasi-Newton, esta estrategia puede producir convergencia superlineal.

Ahora consideraremos tres métodos para encontrar soluciones aproximadas a (5.1) que tienen las características descritas. A lo largo de esta sección nos centraremos en el funcionamiento interno de una sola iteración, por lo que simplificamos la notación dejando de lado el subíndice "k". En esta sección, denotamos la solución de (5.3) por $p^*(\triangle)$, para enfatizar la dependencia de $\triangle$.

El método pata de perro (Dogleg)

El primer enfoque que se discute va por el título descriptivo del método pata de perro. Se puede utilizar cuando B es definida positiva. Para motivar este método, comenzamos examinando el efecto del radio de la región factible $\triangle$ en la solución $p^*(\triangle)$ del subproblema (5.3). Cuando B es definida positiva, ya hemos señalado que el minimizador sin restricciones de m es $p^B = -B^{-1}g$. Cuando este punto es factible para (5.3), es obviamente una solución, por lo que tenemos

$$p^*(\triangle) = p^B, \text{ cuando } \triangle \geq \|p^B\|.$$

El método pata de perro encuentra una solución aproximada mediante la sustitución de la trayectoria curvada para $p^*(\triangle)$ con una trayectoria que consiste en dos segmentos de línea. El primer segmento de línea se extiende desde el origen hasta el minimizador de m a lo largo de la dirección del descenso más agudo, que es

$$p^U = -\frac{g^T g}{g^T B g} g.$$

Mientras que el segundo segmento de línea se extiende desde p^U *a* p^B. Formalmente, denotamos esta trayectoria por $\tilde{p}(\tau)$ para $\tau \in [0, 2]$, donde

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

El método pata de perro elige p para minimizar el modelo m a lo largo este camino, sin perjuicio de la región factible vinculada. El siguiente lema muestra que el mínimo a lo largo de la ruta pata de perro se puede encontrar fácilmente.

Lema 2 *: Sea B definida positiva. Entonces*

(i) *$\|\tilde{p}(\tau)\|$ es una función creciente de τ, y*

(ii) *$m(\tilde{p}(\tau))$ es una función decreciente de τ.*

Se deduce de este lema que el camino $\tilde{p}(\tau)$ intersecta la región factible límite $\|p\| = \triangle$ exactamente en un punto si $\|p^B\| \geq \triangle$ y en ningún otro caso. Como m es decreciente a lo largo del camino, el valor elegido de p estará en p^B si $\|p^B\| \leq \triangle$, de lo contrario en el punto de intersección de la pata de perro y la región factible límite. En este último caso, se calcula el valor apropiado de τ resolviendo la siguiente ecuación cuadrática escalar:

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

Minimización del subespacio bidimensional

Cuando B es definida positiva, la estrategia del método pata de perro se puede hacer un poco más sofisticada por intensificar la búsqueda de p para todo el subespacio bidimensional abarcado por p^U y p^B (equivalentemente a g y $-B^{-1}g$). El subproblema (5.3) se sustituye por

$$\min_p m(p) = f + g^T p + \frac{1}{2} p^T B p \;\text{ sujeto a } \|p\| \leq \triangle,\; p \in \text{gen}[g, B^{-1}g].$$

El cual es un problema de dos variables que es computacionalmente barato para resolver.(Después de una cierta manipulación algebraica se puede reducir a la búsqueda de raíces de un polinomio de cuarto grado.) Es evidente que el punto Cauchy es factible, por lo que la solución óptima de este subproblema al menos da reducción en m como en el punto de Cauchy, lo que resulta de la convergencia global del algoritmo.

5.2. Convergencia Global

Reducción obtenida por el punto Cauchy

En la discusión anterior de los algoritmos para resolver aproximadamente el subproblema de región factible, hemos subrayado en repetidas ocasiones que la convergencia global depende de la solución aproximada de obtener al menos tanta disminución de la función modelo m como del punto Cauchy. Comenzamos el análisis de convergencia global mediante la obtención de una estimación de la disminución de m lograda por el punto Cauchy. A continuación, utilizamos esta estimación para demostrar que la secuencia de los gradientes $\{g_k\}$ generada por el algoritmo 5.1 tiene un punto de acumulación en cero y de hecho converge a cero cuando η es estrictamente positivo.

Nuestro primer resultado principal es que el pata de perro y los algoritmos de minimización del subespacio de dos dimensiones producen soluciones aproximadas p_k del subproblema (5.1) que cumplen la siguiente estimación de disminución de la función modelo:

$$m_k(0) - m_k(p_k) \geqslant c_1 \|g_k\| \min\left(\triangle_k, \frac{\|g_k\|}{\|B_k\|}\right), \tag{5.4}$$

para alguna constante $c_1 \in (0, 1]$. La utilidad de esta estimación se hará evidente en las secciones siguientes. Por ahora, observamos que cuando $\triangle_k$ es el valor mínimo de (5.4), la condición ligeramente recuerda a la primera condición Wolfe: la reducción deseada en el modelo es proporcional al gradiente y al tamaño de paso.

Mostramos ahora que el punto Cauchy p_k^c satisface (5.4) con $c_1 = \frac{1}{2}$.

Lema 3 *: El punto Cauchy p_k^c satisface (5.4) con $c_1 = \frac{1}{2}$, tal que,*

$$m_k(0) - m_k(p_k^c) \geqslant \frac{1}{2} \|g_k\| \min\left(\triangle_k, \frac{\|g_k\|}{\|B_k\|}\right).$$

Para satisfacer (5.4), nuestra solución aproximada p_k ha de lograr una reducción que es al menos alguna fracción c_2 fija de la reducción lograda por el punto Cauchy.

Teorema 15 *:*
Sea p_k cualquier vector tal que $\|p_k\| \leq \triangle_k$ y $m_k(0) - m_k(p_k) \geq c_2(m_k(0) - m_k(p_k^C))$. Entonces p_k satisface (5.4) con $c_1 = c_2/2$. En particular, si p_k es la solución exacta p_k^* de (5.1), entonces esta satisface (5.4) con $c_1 = \frac{1}{2}$.

Tener en cuenta que pata de perro y los algoritmos de minimización del subespacio de dos dimensiones, ambos satisfacen (5.4) con $c_1 = \frac{1}{2}$, porque todos ellos producen soluciones aproximadas p_k para el que $m_k(p_k) \leq m_k(p_k^c)$.

Convergencia de los puntos fijos

Resultados de convergencia global para los métodos de región factible vienen en dos variedades, dependiendo de si fijamos el parámetro η en el algoritmo 5.1 a cero o a algún pequeño valor positivo. Cuando $\eta = 0$ podemos demostrar que la secuencia de los gradientes $\{g_k\}$ tiene un punto límite en cero.

En esta sección demostramos los resultados de convergencia global para ambos casos. Suponemos que a lo largo, los hessianos B_k aproximadamente están uniformemente acotados en norma y que f estará limitada en el conjunto

$$S \stackrel{def}{=} \{x | f(x) \leq f(x_0)\}.$$

Para su posterior consulta, definimos un entorno abierto de este conjunto por

$$S(R_0) \stackrel{def}{=} \{x | \|x - y\| < R_0 \text{ para todo } y \in S\},$$

donde R_0 es una constante positiva. También permitimos que la longitud de la solución aproximada p_k de (5.1) supere la región factible unidad, siempre que se mantenga dentro de un múltiplo fijo del límite; es decir,

$$\|p_k\| \leq \gamma \triangle_k, \tag{5.5}$$

para alguna constante $\gamma \geq 1$. El primer resultado se ocupa del caso cuando $\eta = 0$.

Teorema 16 *:*
Sea $\eta = 0$ en el algoritmo 5.1. Supongamos que $B_k \leq \beta$ para alguna constante β, que f está limitada en el conjunto de nivel S definido anteriormente y Lipschitz continuamente diferenciable en el vecindario $S(R_0)$ para algun $R_0 > 0$, y que todas las soluciones aproximadas de (5.1) satisfacen las desigualdades (5.4) y (5.5), para algunas constantes positivas c_1 y γ. Entonces

$$\liminf_{k \to \infty} \|g_k\| = 0.$$

Teorema 17 *:*
Sea $\eta \in (0, \frac{1}{4})$ en el algoritmo 5.1. Supongamos que $\|B_k\| \leq \beta$ para alguna constante β, que f está limitada en el conjunto de nivel S y continuamente diferenciable Lipschitz en $S(R_0)$ para algun $R_0 > 0$, y que todas las soluciones aproximadas p_k de (5.1) satisfacen las desigualdades (5.4) y (5.5) para algunas constantes positivas c_1 y γ. Entonces

$$\lim_{k \to \infty} g_k = 0.$$

5.3. Solución Iterativa del Subproblema

En esta sección, se describe un enfoque para encontrar una buena aproximación a costa de unas cuantas factorizaciones de la matriz B (típicamente tres factorizaciones), en comparación con una sola factorización para el pata de perro y los métodos bidimensionales de minimización de subespacio. Este enfoque se basa en la caracterización de la solución exacta dada en el Teorema 14, junto con una ingeniosa aplicación del método de Newton en una variable. Esencialmente, el algoritmo trata de identificar el valor de λ para los que la primera condición del Teorema 14 se satisface con las soluciónes de (5.3).

La caracterización del Teorema 14 sugiere un algoritmo para encontrar la solución p de (5.3). Cualquier $\lambda = 0$ satisface las condiciones del Teorema 14 con $\|p\| \leq \triangle$ y también se define

$$p(\lambda) = -(B + \lambda I)^{-1} g$$

para λ suficientemente grande y $B + \lambda I$ definida positiva y buscamos un valor $\lambda > 0$ tal que

$$\|p(\lambda)\| = \triangle.$$

El cual es un problema unidimensional para encontrar raíces en la variable λ. Para ver que existe un valor de λ con todas las propiedades deseadas, hacemos un llamado a los eigenvalores de B y los usamos para estudiar las propiedades de $\|p(\lambda)\|$. Puesto que B es simétrica, hay una matriz ortogonal Q y una matriz diagonal Λ tal que $B = Q\Lambda Q^T$, donde

$$\Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_n),$$

y $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ son los eigenvalores de B. Aquí $B + \lambda I = Q(\Lambda + \lambda I)Q^T$, y para $\lambda \neq \lambda_j$, se tiene

$$p(\lambda) = -Q(\Lambda + \lambda I)^{-1} Q^T g = -\sum_{j=1}^{n} \frac{q_j^T g}{\lambda_j + \lambda} q_j,$$

en donde q_j denota la columna j-esima de Q. Por lo tanto, por ortonormalidad de $q_1, q_2, ..., q_n$, se tiene

$$\|p(\lambda)\|^2 = \sum_{j=1}^{n} \frac{\left(q_j^T g\right)^2}{(\lambda_j + \lambda)^2}.$$

Esta expresión nos dice mucho acerca de $\|p(\lambda)\|$. Si $\lambda > -\lambda_1$, se tiene $\lambda_j + \lambda > 0$ para todo $j = 1, 2, ..., n$, y así $\|p(\lambda)\|$ es una función continua no creciente de λ en el intervalo $(-\lambda_1, \infty)$. De hecho, tenemos que

$$\lim_{\lambda \to \infty} \|p(\lambda)\| = 0.$$

Más aún, cuando $q_j^T g \neq 0$ tenemos que

$$\lim_{\lambda \to -\lambda_j} \|p(\lambda)\| = \infty.$$

La Figura 5.1 de $\|p(\lambda)\|$ contra λ es un caso en el que $q_1^T g, q_2^T g$, y $q_3^T g$ son todos distintos de cero. Tener en cuenta que las dos propiedades anteriores sostienen que $\|p(\lambda)\|$ es una función no creciente de λ en $(-\lambda_1, \infty)$. En particular, siempre que $q_1^T g \neq 0$ hay un único valor

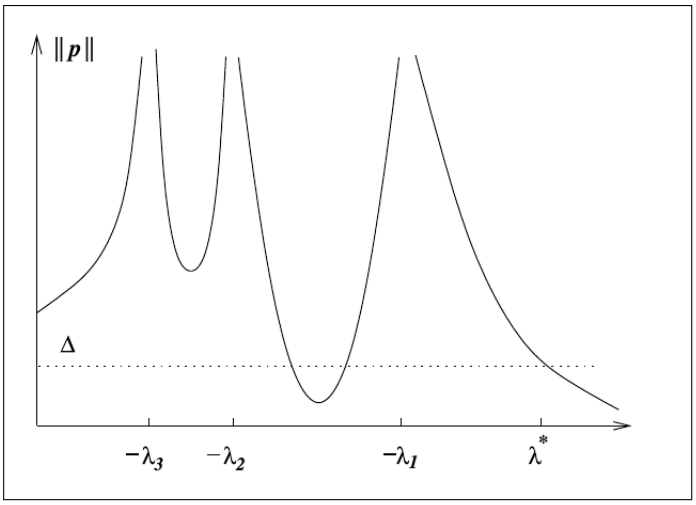


Figura 5.1: $\|p(\lambda)\|$ *como una función de* λ.

$\lambda^* \in (-\lambda_1, \infty)$ tal que $\|p(\lambda)\| = \triangle$.

Ahora esbozamos un procedimiento para identificar los $\lambda^* \in (-\lambda_1, \infty)$ para los que $\|p(\lambda^*)\| = \triangle$, que funciona cuando $q_1^T g \neq 0$. En primer lugar, tener en cuenta que cuando B es definida positiva y $\|B^{-1}g\| \leq \triangle$, el valor $\lambda = 0$ satisface las condiciones del Teorema 14, por lo que el procedimiento puede ser terminado inmediatamente con $\lambda^* = 0$. De lo contrario, se podría utilizar el método de busqueda de raíz de Newton para encontrar el valor de $\lambda > -\lambda_1$ que resuelve

$$\phi_1(\lambda) = \|p(\lambda)\| - \triangle = 0.$$

La desventaja de este enfoque puede ser la forma de considerar $\|p(\lambda)\|$ cuando λ es mucho menor, aunque próximo a $-\lambda_1$. Para tales λ, podemos aproximar ϕ_1 por una función racional, como sigue:

$$\phi_1(\lambda) \approx \frac{C_1}{\lambda + \lambda_1} + C_2,$$

donde $C_1 > 0$ y C_2 son constantes. Es evidente que esta aproximación (y por tanto ϕ_1) es altamente no lineal, por lo que el método de investigación de raíz Newton será poco fiable o lento. Se obtendrán mejores resultados si reformulamos el problema de manera que es casi lineal cerca del λ óptimo. Definiendo

$$\phi_2(\lambda) = \frac{1}{\triangle} - \frac{1}{\|p(\lambda)\|},$$

se puede mostrar usando que para λ ligeramente mayor que $-\lambda_1$, tenemos

$$\phi_2(\lambda) \approx \frac{1}{\triangle} - \frac{\lambda + \lambda_1}{C_3}$$

para algun $C_3 > 0$. Por lo tanto, ϕ_2 es casi lineal cerca de $-\lambda_1$ (véase la Figura 5.2) y el método de busqueda de la raíz de Newton será un buen desempeño, siempre que se mantenga

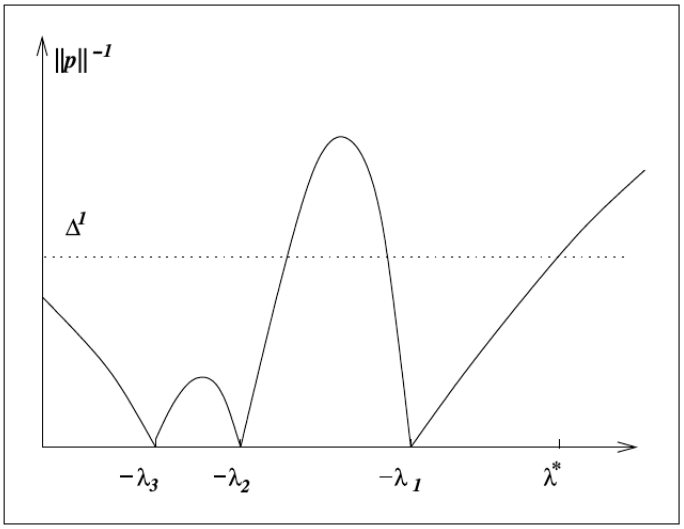


Figura 5.2: $\dfrac{1}{\|p(\lambda)\|}$ *como función de* λ.

$\lambda > -\lambda_1$. El método de investigación de raíz Newton aplicado a ϕ_2 genera una secuencia de iteraciones $\lambda^{(l)}$ mediante el establecimiento de

$$\lambda^{(l+1)} = \lambda^{(l)} - \frac{\phi_2(\lambda^{(l)})}{\phi_2'(\lambda^{(l)})}.$$

Después de una cierta manipulación elemental, esta fórmula de actualización se puede implementar de la siguiente manera práctica.

ALGORITMO 5.3 (Subproblema de Región Factible)
Dado $\lambda^{(0)}, \triangle > 0$:
for $l = 1, 2, \ldots$
Factor $B + \lambda^{(l)} I = R^T R$;
Resolver $R^T R p_l = -g, R^T q_l = p_l$;
Donde

$$\lambda^{(l+1)} = \lambda^{(l)} + \left(\frac{\|p_l\|}{\|q_l\|}\right)^2 \left(\frac{\|p_l\| - \triangle}{\triangle}\right);$$

end (for).

El trabajo principal en cada iteración de este método es la factorización Cholesky $B + \lambda^{(l)} I$. Versiones prácticas de este algoritmo que no dejan de iterar hasta que se obtiene la convergencia del λ óptimo con una alta precisión, se conforman con una solución aproximada que se puede obtener en dos o tres iteraciones.

5.4. Problemas de Mínimos Cuadrados

En problemas de mínimos cuadrados, la función objetivo f tiene la siguiente forma especial:

$$f(\mathbf{x}) = \frac{1}{2}\sum_{j=1}^{m} r_j^2(\mathbf{x}), \tag{5.6}$$

donde cada r_j es una función suave de $\mathbb{R}^n$ a $\mathbb{R}$. Nos referimos a cada r_j como un residuo, y suponemos a lo largo de esta seccción que $m \geq n$.

Problemas de mínimos cuadrados surgen en muchas áreas de aplicaciones y pueden de hecho ser la mayor fuente de problemas de optimización sin restricciones. Muchos de los que formulan un modelo parametrizado de una sustancia química, física, financiera o aplicación económica utilizan una función como la anterior para medir la discrepancia entre el modelo y el comportamiento observado del sistema. Al minimizar esta función, seleccionamos los valores de los parámetros que mejor coinciden con los datos del modelo. En esta sección se muestra cómo diseñar algoritmos de minimización eficientes, robustos explotando la estructura especial de la función f y sus derivados.

Para ver por qué la forma especial de f a menudo hace que los problemas de mínimos cuadrados sea más fácil de resolver que los problemas generales de minimización sin restricciones, primero colocamos las componentes individuales r_j de (5.6) en un vector residual $r : \mathbb{R}^n \to \mathbb{R}^m$ de la siguiente manera

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

Usando esta notación, podemos reescribir f como $f(\mathbf{x}) = \frac{1}{2}\|r(\mathbf{x})\|_2^2$. Los derivados de $f(\mathbf{x})$ se pueden expresar en términos del jacobiano $J(\mathbf{x})$, que es una matriz $m \times n$ de primeras derivadas parciales de los residuos, definido por

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

donde cada $\bigtriangledown r_j(\mathbf{x})$, $j = 1, 2, ..., m$ es el gradiente de r_j. El gradiente y Hessiano de f se pueden expresar como sigue:

$$\bigtriangledown f(\mathbf{x}) = \sum_{j=1}^{m} r_j(\mathbf{x}) \bigtriangledown r_j(\mathbf{x}) = J(\mathbf{x})^T r(\mathbf{x}), \tag{5.7}$$

$$\bigtriangledown^2 f(\mathbf{x}) = \sum_{j=1}^{m} \bigtriangledown r_j(\mathbf{x}) \bigtriangledown r_j(\mathbf{x})^T + \sum_{j=1}^{m} r_j(\mathbf{x}) \bigtriangledown^2 r_j(\mathbf{x})$$

$$\bigtriangledown^2 f(\mathbf{x}) = J(\mathbf{x})^T J(\mathbf{x}) + \sum_{j=1}^{m} r_j(\mathbf{x}) \bigtriangledown^2 r_j(\mathbf{x}). \tag{5.8}$$

En muchas aplicaciones, las primeras derivadas parciales de los residuos y por tanto, la matriz jacobiana $J(\mathbf{x})$, son relativamente fáciles de calcular. De este modo podemos obtener

el gradiente $\bigtriangledown f(\mathbf{x})$ usando la fórmula anterior. Usando $J(\mathbf{x})$, también podemos calcular el primer término de $J(\mathbf{x})^T J(\mathbf{x})$ en el Hessiano $\bigtriangledown^2 f(\mathbf{x})$ sin evaluar las segundas derivadas de las funciones r_j. Esta disponibilidad de parte de $\bigtriangledown^2 f(\mathbf{x})$ es el rasgo distintivo de los problemas de mínimos cuadrados. Por otra parte, este término $J(\mathbf{x})^T J(\mathbf{x})$ es a menudo más importante que el segundo término de la suma, ya sea porque los residuales r_j están cerca de la solución (es decir, $\bigtriangledown^2 r_j(\mathbf{x})$ es relativamente pequeño) o debido a pequeños residuos (es decir, los $r_j(\mathbf{x})$ son relativamente pequeños). La mayoría de los algoritmos no lineales de mínimos cuadrados explotan estas propiedades estructurales de la matriz Hessiana.

Los algoritmos más populares para reducir al mínimo (5.6) encajan en los marcos de búsqueda de línea y región factible que se describieron anteriormente. Se basan en los enfoques Newton y Cuasi-Newton descritos anteriormente, con las modificaciones que se aprovechan de la estructura particular de f.

5.4.1. Antecedentes

Discutimos un modelo parametrizado simple y mostramos cómo las técnicas de mínimos cuadrados se pueden utilizar para elegir los parámetros que mejor se adapten al modelo de los datos observados.

Ejemplo 6 . *Queremos estudiar el efecto de un determinado medicamento en un paciente. Obtenemos las muestras de sangre en determinados momentos después de que el paciente toma una dosis, y medimos la concentración del medicamento en cada muestra, tabulamos el tiempo t_j y la concentración y_j para cada muestra.*

Basándonos en nuestra experiencia previa en este tipo de experimentos, nos encontramos con que la siguiente función $\phi(\mathbf{x}; t)$ ofrece una buena predicción de la concentración en el tiempo t, para valores apropiados de los cinco parámetros dimensionales del vector $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$:

$$\phi(\mathbf{x}; t) = x_1 + tx_2 + t^2 x_3 + x_4 \exp^{-x_5 t}.$$

Elegimos el parámetro del vector $\mathbf{x}$ de manera que este modelo esté de acuerdo con nuestra observación, en algún sentido. Una buena manera de medir la diferencia entre los valores predichos por el modelo y las observaciones es la función de mínimos cuadrados siguiente:

$$\frac{1}{2} \sum_{j=1}^{m} [\phi(\mathbf{x}; t_j) - y_j]^2, \tag{5.9}$$

que resume los cuadrados de las discrepancias entre las predicciones y las observaciones en cada t_j. Esta función tiene precisamente la forma (5.6) si definimos

$$r_j(\mathbf{x}) = \phi(\mathbf{x}; t_j) - y_j.$$

Gráficamente, cada término en (5.9) representa el cuadrado de la distancia vertical entre la curva $\phi(\mathbf{x}; t)$ (trazada como una función de t) y el punto (t_j, y_j), para una elección fija del parámetro del vector $\mathbf{x}$. El minimizador $\mathbf{x}^*$ del problema de mínimos cuadrados es el vector de parámetro para el que se minimiza la suma de los cuadrados de las longitudes de las líneas de puntos en la Figura 5.3. Habiendo obtenido $\mathbf{x}^*$, utilizamos $\phi(\mathbf{x}^*; t)$ para estimar la concentración de medicamento que queda en el torrente sanguíneo del paciente en cualquier momento t.

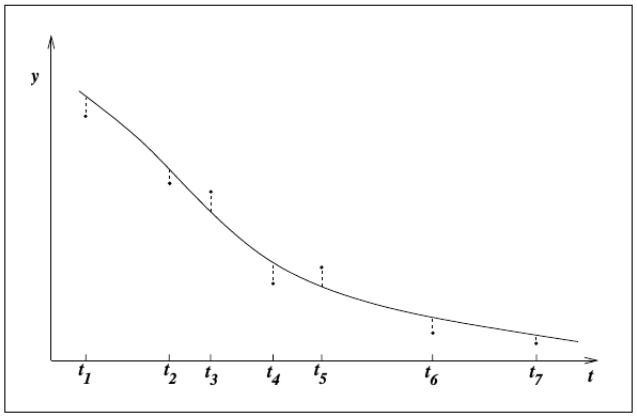


Figura 5.3: *Modelo 5.9 con desviaciones indicadas por líneas de puntos verticales.*

□

En los problemas generales de datos de ajuste del tipo que acabamos de describir, el t ordenada en el modelo de $\phi(\mathbf{x}; t)$ podría ser un vector en lugar de un escalar. (En el ejemplo anterior, por ejemplo, t podría tener dos dimensiones, la primera dimensión que representa el tiempo desde que se administra el fármaco y la segunda dimensión que representa el peso del paciente. Podríamos entonces utilizar las observaciones para toda una población de pacientes y no sólo un paciente, para obtener los "mejores"parámetros de este modelo).

La función de suma de cuadrados (5.9) no es la única manera de medir la discrepancia entre el modelo y las observaciones. Otras medidas comunes incluyen el valor máximo absoluto

$$\max_{j=1,2,...,m} |\phi(\mathbf{x}; t_j) - y_j|$$

y la suma de los valores absolutos

$$\sum_{j=1}^{m} |\phi(\mathbf{x}; t_j) - y_j|.$$

En algunas situaciones, hay motivaciones estadísticas para elegir el criterio de mínimos cuadrados. Cambiando la notación de la discrepancia entre el modelo y la observación por ϵ_j, es decir,

$$\epsilon_j = \phi(\mathbf{x}; t_j) - y_j.$$

A menudo es razonable suponer que los $\epsilon_j's$ son independientes e idénticamente distribuidos con una cierta varianza σ^2 y función de densidad de probabilidad $g_\sigma(\Delta)$. (Este supuesto será a menudo el caso, por ejemplo, cuando el modelo refleja con precisión el proceso real, y cuando los errores cometidos en la obtención de las mediciones y_j no contienen un sesgo sistemático.) Bajo este supuesto, la probabilidad de un conjunto particular de observaciones $y_j,\ j = 1, 2, ..., m$, dado que el vector de parámetro actual es $\mathbf{x}$, está dado por la función

$$p(\mathbf{y}; \mathbf{x}, \sigma) = \prod_{j=1}^{m} g_\sigma(\epsilon_j) = \prod_{j=1}^{m} g_\sigma(\phi(\mathbf{x}; t_j) - y_j).$$

Cuando asumimos que las discrepancias siguen una distribución normal, tenemos

$$g_\sigma(\epsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

De donde tendriamos que:

$$p(\mathbf{y}; \mathbf{x}, \sigma) = (2\pi\sigma^2)^{-m/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{m}[\phi(\mathbf{x}; t_j) - y_j]^2\right).$$

Para cualquier valor fijo de la varianza σ^2, es obvio que p se maximiza cuando se minimiza la suma de cuadrados (5.9). Para resumir: Cuando se supone que las discrepancias son independientes e idénticamente distribuidas con una función de distribución normal, la estimación de máxima verosimilitud se obtiene al minimizar la suma de cuadrados.

5.4.2. Problemas Lineales de Mínimos Cuadrados

Muchos modelos de problemas $\phi(\mathbf{x}; t)$ con datos de ajuste son funciones lineales de $\mathbf{x}$. En estos casos, los residuos $r_j(\mathbf{x})$ definidos anteriormente también son lineales, y el problema de minimización de (5.9) se llama un problema de mínimos cuadrados lineales. Podemos escribir el vector residual como $r(\mathbf{x}) = J\mathbf{x} - \mathbf{y}$ para alguna matriz J y vector $\mathbf{y}$ independiente de $\mathbf{x}$, de modo que el objetivo es

$$f(\mathbf{x}) = \frac{1}{2}\|J\mathbf{x} - \mathbf{y}\|^2, \tag{5.10}$$

donde $\mathbf{y} = r(0)$. También tenemos

$$\bigtriangledown f(\mathbf{x}) = J^T(J\mathbf{x} - \mathbf{y}), \quad \bigtriangledown^2 f(\mathbf{x}) = J^T J.$$

(Notar que el segundo término de $\bigtriangledown^2 f(\mathbf{x})$ desaparece, porque $\bigtriangledown^2 r_j = 0$ para todo $j = 1, 2, ..., m.$) Es fácil ver que $f(\mathbf{x})$ en (5.10) es convexa una propiedad que no necesariamente cumple el problema no lineal (5.6). El Teorema 10 nos dice que cualquier punto $\mathbf{x}$ para el que $\bigtriangledown f(\mathbf{x}^*) = 0$ es el minimizador global de f. Por lo tanto, $\mathbf{x}^*$ debe satisfacer el siguiente sistema de ecuaciones lineales:

$$J^T J\mathbf{x}^* = J^T\mathbf{y}. \tag{5.11}$$

Que son conocidas como las ecuaciones normales de (5.10).

Describimos brevemente tres principales algoritmos para el problema lineal sin restricciones de mínimos cuadrados. Asumimos en la mayor parte de nuestra discusión que $m \geq n$ y que J tiene rango de columna completo.

El primero y más evidente algoritmo es simplemente para formar y resolver el sistema de ecuaciones normales para (5.10) mediante el siguiente procedimiento de tres pasos:

- Calcular la matriz de coeficientes $J^T J$ y el lado derecho $J^T\mathbf{y}$;
- Calcular la factorización de Cholesky de la matriz simétrica $J^T J$;
- Realizar dos sustituciones triangulares con los factores de Cholesky para recuperar la solución $\mathbf{x}^*$.

La factorización de Cholesky

$$J^T J = \bar{R}^T \bar{R}$$

(donde $\bar{R}$ es triangular superior $n \times n$ con elementos diagonales positivos) se garantiza que existe cuando $m \geq n$ y J tiene rango n. Este método se utiliza con frecuencia en la práctica y es a menudo eficaz, pero tiene una desventaja significativa, es decir, que número de condición de $J^T J$ es el cuadrado del número de condición de J.

Un segundo enfoque se basa en una factorización QR de la matriz J. Dado que la norma euclidiana de cualquier vector no se ve afectada por las transformaciones ortogonales, tenemos

$$\|J\mathbf{x} - \mathbf{y}\| = \|Q^T(J\mathbf{x} - \mathbf{y})\| \tag{5.12}$$

para cualquier matriz ortogonal Q de $m \times m$. Supongamos que realizamos una factorización QR con giro de columna en la matriz J para obtener

$$J\Pi = Q \begin{bmatrix} R \\ 0 \end{bmatrix} = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} = Q_1 R, \tag{5.13}$$

donde

Π es una matriz de permutación $n \times n$ (aquí ortogonal);
Q es ortogonal $m \times m$;
Q_1 son las primeras n columnas de Q, Q_2 contiene las restantes $m - n$ columnas;
R es triangular inferior $n \times n$ con elementos positivos en la diagonal.

Mediante la combinación de (5.12) y (5.13), obtenemos

$$\begin{aligned} \|J\mathbf{x} - \mathbf{y}\|_2^2 &= \left\| \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} (J\Pi\Pi^T\mathbf{x} - \mathbf{y}) \right\|_2^2 \\ &= \left\| \begin{bmatrix} R \\ 0 \end{bmatrix} (\Pi^T\mathbf{x}) - \begin{bmatrix} Q_1^T\mathbf{y} \\ Q_2^T\mathbf{y} \end{bmatrix} \right\|^2 \\ &= \|R(\Pi^T\mathbf{x}) - Q_1^T y\|_2^2 + \|Q_2^T\mathbf{y}\|^2. \end{aligned}$$

No hay opción de $\mathbf{x}$ sobre el segundo término de esta última expresión, pero podemos minimizar $J\mathbf{x} - \mathbf{y}$ por la conducción del primer término a cero, es decir, mediante el establecimiento de

$$\mathbf{x}^* = \Pi^{-1} Q_1^T \mathbf{y} R^{-1}.$$

Un tercer enfoque, basado en la descomposición de valor singular (SVD) de J, se puede utilizar en estas circunstancias. Recordemos que el SVD de J viene dado por

$$J = U \begin{bmatrix} S \\ 0 \end{bmatrix} V^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} S \\ 0 \end{bmatrix} V^T = U_1 S V^T,$$

donde

U es ortogonal $m \times m$;
U_1 contiene las primeras n columnas de U, U_2 las restantes $m - n$ columnas;
V es ortogonal $n \times n$;
S es diagonal $n \times n$, con elementos en la diagonal $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$.

Siguiendo la misma lógica anterior, tenemos

$$\begin{aligned}\|J\mathbf{x}-\mathbf{y}\|^2 &= \left\|\begin{bmatrix} S \\ 0 \end{bmatrix}(V^T\mathbf{x}) - \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} y\right\|^2 \\ &= \|S(V^T\mathbf{x}) - U_1^T\mathbf{y}\|^2 + \|U_2^T\mathbf{y}\|^2.\end{aligned}$$

Una vez más, el óptimo se encuentra eligiendo $\mathbf{x}$ para hacer el primer término igual a cero; esto es,

$$\mathbf{x}^* = VS^{-1}U_1^T\mathbf{y}.$$

Denotando las i-esimas columnas de U y V por $u_i \in \mathbb{R}^m$ y $v_i \in \mathbb{R}^n$, respectivamente, tenemos

$$\mathbf{x}^* = \sum_{i=1}^{n} \frac{u_i^T\mathbf{y}}{\sigma_i} v_i. \tag{5.14}$$

De esta fórmula se obtiene información útil sobre la sensibilidad de $\mathbf{x}^*$. Cuando σ_i es pequeño, $\mathbf{x}^*$ es particularmente sensible a las perturbaciones en $\mathbf{y}$ que afectan $u_i^T\mathbf{y}$ y también a las perturbaciones en J que afectan a esta misma cantidad.

Los tres enfoques de arriba tienen su lugar. El algoritmo basado en Cholesky es particularmente útil cuando $m \gg n$ y es práctico almacenar J^TJ pero no J en sí mismo. El enfoque QR evita cuadratura del número de condición y por lo tanto puede ser más robusto numéricamente. Mientras que potencialmente el más caro, el enfoque SVD es el más robusto y fiable de todos.

5.4.3. Algoritmos de Mínimos Cuadrados para Problemas no Lineales

El método Gauss-Newton

Ahora describimos los métodos para minimizar la función objetivo no lineal (5.6) que aprovechan la estructura del gradiente ∇f (5.7) y el Hessiano $\nabla^2 f$ (5.8). El más simple de estos métodos, es el método Gauss-Newton que puede ser visto como un método de Newton modificado con la búsqueda de línea.

En lugar de resolver el estándar Newton de ecuaciones $\nabla^2 f(x_k)p = -\nabla f(x_k)$, nosotros resolvemos el siguiente sistema para obtener la dirección de búsqueda p_k^{GN} :

$$J_k^T J_k p_k^{GN} = -J_k^T r_k. \tag{5.15}$$

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

$$\nabla^2 f_k \approx J_k^T J_k, \tag{5.16}$$

la cual nos ahorra el problema de calcular los residuos individuales del Hessiano $\nabla^2 r_j, j = 0, 1, 2, ..., m$, que son necesarios en el segundo término de (5.8). De hecho, si calculamos el jacobiano J_k en el curso de la evaluación del gradiente $\nabla f_k = J_k^T r_k$, la aproximación (5.16)

no requiere evaluaciones de derivadas adicionales, y el ahorro en tiempo de cálculo puede ser muy importante en algunas aplicaciones. En segundo lugar, hay muchas situaciones interesantes en el que el primer término de (5.8) domina al segundo término (al menos cerca de la solución x^*), de modo 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 (5.8) será dominante cuando la norma de cada término de segundo orden (es decir, $|r_j(x)|\|\nabla^2 r_j(x)\|$) es significativamente menor que los valores propios de $J^T J$. Como se mencionó en la introducción, tendemos a ver este comportamiento cuando sean los residuos r_j pequeños o cuando están casi afín (de modo que los $\|\nabla^2 r_j\|$ son pequeños). En la práctica, muchos problemas de mínimos cuadrados tienen pequeños residuos en la solución, lo que lleva a una rápida convergencia local de Gauss-Newton.

Una tercera ventaja de Gauss-Newton es que cada vez J_k tiene rango completo y el gradiente ∇f_k es distinto de cero, la dirección p_k^{GN} es una dirección de descenso para f, y por lo tanto una dirección adecuada para una búsqueda de línea. A partir de (5.7) y (5.16) tenemos:

$$(p_k^{GN})^T \nabla f_k = (p_k^{GN})^T J_k^T r_k = -(p_k^{GN})^T J_k^T J_k p_k^{GN} = -\|J_k p_k^{GN}\|^2 \leq 0. \tag{5.17}$$

La desigualdad final es menos estricta $J_k p_k^{GN} = 0$, en cuyo caso tenemos por (5.15) y el rango completo de J_k que $J_k^T r_k = \nabla f_k = 0$; es decir, x_k es un punto estacionario. Por último, la cuarta ventaja de Gauss-Newton se deriva de la similitud entre las ecuaciones (5.15) y las ecuaciones normales (5.11) dada para el problema de mínimos cuadrados lineales. Esta conexión nos dice que p_k^{GN} es, de hecho, la solución del problema de mínimos cuadrados lineales.

$$\min_p \frac{1}{2}\|J_k p + r_k\|^2. \tag{5.18}$$

Por lo tanto, podemos encontrar la dirección de búsqueda mediante la aplicación de algoritmos de mínimos cuadrados lineales para el subproblema (5.18). De hecho, si se utilizan los algoritmos basados en QR o SVD, no hay necesidad de calcular la aproximación del Hessiano $J_k^T J_k$ en (5.15) de forma explícita; podemos trabajar directamente con el jacobiano J_k. Lo mismo ocurre si utilizamos una técnica de gradiente conjugado para resolver (5.18). Para este método necesitamos realizar multiplicaciones de matriz-vector con $J_k^T J_k$, que puede realizarse multiplicando por J_k y luego por J_k^T.

Si el número de residuos m es grande mientras el número de variables n es relativamente pequeño, puede ser prudente para almacenar el jacobiano J explícitamente. Una estrategia preferible puede ser para calcular la matriz $J^T J$ y el vector gradiente $J^T r$ evaluando r_j y ∇r_j sucesivamente para $j = 1, 2, ..., m$ y la realización de las acumulaciones

$$J^T J = \sum_{i=1}^{m} (\nabla r_j)(\nabla r_j)^T, \qquad J^T r = \sum_{i=1}^{m} r_j (\nabla r_j). \tag{5.19}$$

Los pasos de Gauss-Newton se pueden entonces calcular resolviendo el sistema (5.15) de ecuaciones normales directamente.

El subproblema (5.18) sugiere otra motivación para la dirección de búsqueda de Gauss-Newton. Podemos ver esta ecuación que se obtiene a partir de un modelo lineal para la

función del vector $r(x_k + p) \approx r_k + J_k p$ sustituido en la función $\frac{1}{2}\|\cdot\|^2$. En otras palabras, utilizar la aproximación $\frac{1}{2}\|r(x_k+p)\|^2 \approx \frac{1}{2}\|J_k p + r_k\|^2$, y seleccionar p_k^{GN} para ser el minimizador de esta aproximación. Las implementaciones del método Gauss-Newton suelen realizar una búsqueda de línea en la dirección p_k^{GN}, requiriendo el paso de longitud α_k para satisfacer las condiciones como las discutidas en el Capítulo 4, tales como las condiciones Armijo y Wolfe.

Convergencia del método Gauss-Newton

La teoría del capítulo 4 puede aplicarse para estudiar las propiedades de convergencia del método Gauss-Newton. Probamos resultados de convergencia global bajo el supuesto de que el jacobiano de $J(x)$ tienen sus valores singulares que estan acotados uniformemente fuera de cero en la región de interés; es decir, hay una constante $\gamma > 0$ tal que

$$\|J(x)z\| \geq \gamma\|z\|, \tag{5.20}$$

para todo x en un vecindario $\mathcal{N}$ del conjunto de nivel

$$\mathcal{L} = \{x | f(x) \leq f(x_0)\}, \tag{5.21}$$

donde x_0 es el punto de partida para el algoritmo. Asumiremos aquí y en el resto de los capítulos que $\mathcal{L}$ está acotada. Dicho resultado es consecuencia del teorema 12 en el capítulo 4.

Teorema 18 *: Supongamos que cada función residual r_j es Lipschitz continuamente diferenciable en un vecindario $\mathcal{N}$ del conjunto de nivel acotado (5.21), y que el jacobiano $J(x)$ satisface la condición de rango uniformemente completo (5.20) en $\mathcal{N}$. Entonces, si las iteraciones x_k se generan por el método de Gauss-Newton con longitudes de paso α_k que satisfacen (5.21), tenemos*

$$\lim_{k\to\infty} J_k^T r_k = 0.$$

Prueba.
En primer lugar, observamos que la vecindad $\mathcal{N}$ acota al conjunto de nivel $\mathcal{L}$ que puede elegirse suficientemente pequeño para que las siguientes propiedades se satisfagan para algunas constantes positivas L y β:

$$|r_j(x)| \leq \beta \qquad \text{y} \qquad \|\nabla r_j(x)\| \leq \beta,$$

$$|r_j(x) - r_j(\tilde{x})| \leq L\|x - \tilde{x}\| \qquad \text{y} \qquad \|\nabla r_j(x) - \nabla r_j(\tilde{x})\| \leq L\|x - \tilde{x},$$

para todo $x, \tilde{x} \epsilon \mathcal{N}$ y todo $j = 1, 2, ..., m$. Es fácil deducir que existe una constante $\bar{\beta} > 0$ tal que $\|J(x)^T\| = \|J(x)\| \leq \bar{\beta}$ para todo $x \epsilon \mathcal{L}$. Además, mediante la aplicación de los resultados referentes a la continuidad de Lipschitz productos y sumas al gradiente $\nabla f(x) = \sum_{j=1}^m r_j(x)\nabla r_j(x)$, podemos demostrar que ∇f es Lipschitz continua. Por lo tanto, las hipótesis del Teorema (11) se satisfacen.

Comprobamos que el siguiente ángulo θ_k entre la dirección de búsqueda p_k^{GN} y el gradiente negativo $-\nabla f_k$ está delimitada de manera uniforme lejos de $\pi/2$. A partir de (4.9), (5.17) y

(2.20) tenemos para $x = x_k \epsilon \mathcal{L}$ y $p^{GN} = p_K^{GN}$ que

$$\cos\theta_k = -\frac{(\nabla f)^T p^{GN}}{\|p^{GN}\|\|\nabla f\|} = \frac{\|Jp^{GN}\|^2}{\|p^{GN}\|\|J^T J p^{GN}\|} \geq \frac{\gamma^2\|p^{GN}\|^2}{\bar{\beta}^2\|p^{GN}\|^2} = \frac{\gamma^2}{\bar{\beta}^2} > 0.$$

Se deduce de (4.11) en el Teorema (11) que $\nabla f(x_k) \to 0$ dando el resultado.

□

Si J_k es de rango deficiente para algunos k (de modo que una condición como la (5.20) no se satisface), la matriz de coeficientes de (5.15) es singular. El sistema (5.15) todavía tiene una solución, debido a la equivalencia entre este sistema lineal y el problema de minimización (5.18).

De hecho, hay infinitas soluciones para p_k^{GN} en este caso; cada uno de ellos tiene la forma de (5.14). Sin embargo, ya no es una garantía que $\cos\theta_k$ esté delimitado de manera uniforme fuera de cero, por lo que no puede probar un resultado como el Teorema (18)

La convergencia de Gauss-Newton a una solución x^* puede ser rápida si el término principal $J_k^T J_k$ domina el término de segundo orden en el Hessiano (5.8). Supongamos que x_k está cerca de x^* y con esa suposición (5.20) se satisface.

El método Levenberg-Marquardt

Recordemos que el método Gauss-Newton es como el método de Newton con la búsqueda de línea, a excepción de que se utiliza la aproximación conveniente y a menudo efectiva (5.16) para el Hessiano. El método Levenberg-Marquardt se puede conseguir mediante el uso de la misma aproximación del Hessiano, pero reemplazando la búsqueda de acuerdo con una estrategia de región factible. El uso de una región factible evita una de las debilidades de Gauss-Newton, es decir, su comportamiento cuando el jacobiano es de rango deficiente, o casi deficiente ya que, las mismas aproximaciones del Hessiano se utilizan en cada caso, las propiedades de convergencia locales de los dos métodos son similares.

El método Levenberg-Marquardt puede ser descrito y analizado utilizando el marco de región factible del capítulo 5. (De hecho, el método Levenberg-Marquardt es a veces considerado como el progenitor del enfoque región factible para la optimización sin restricciones que se analizan en el capítulo 5.) Para una región factible esférica, el subproblema que hay que resolver en cada iteración es

$$\min_p \frac{1}{2}\|J_k p + r_k\|^2, \qquad \text{sujeto a } \|p\| \leq \triangle_k. \tag{5.22}$$

Dónde $\triangle_k > 0$ es el radio de la región factible. En efecto, estamos eligiendo la función modelo $m_k(\cdot)$ en (5.1) para ser

$$m_k = \frac{1}{2}\|r_k\|^2 + p^T J_k^T r_k + \frac{1}{2} p^T J_k^T J_k p. \tag{5.23}$$

Dejamos el contador de la iteración k durante el resto de esta sección y nos preocupamos por el subproblema (5.22). Los resultados del capítulo 5 nos permiten caracterizar la solución de (5.22) de la siguiente manera: Cuando la solución de las ecuaciones de Gauss-Newton (5.15)

se encuentran estrictamente dentro de la región factible (es decir, $\|p^{GN}\| < \triangle$), entonces este paso también resuelve el subproblema (5.22). De lo contrario, hay un $\lambda > 0$ tal que la solución $p = p^{LM}$ de (5.22) satisface $\|p\| = \triangle$, y

$$(J^T J + \lambda I)p = -J^T r. \tag{5.24}$$

Esta afirmación se verifica en el siguiente lema, que es una consecuencia directa del Teorema (14) del Capítulo 5.

Lema 4 *: El vector p^{LM} es una solución del subproblema de región factible*

$$\min_p \|Jp + r\|^2 \quad \textit{sujeto a} \quad \|p\| \leq \triangle$$

si y sólo si p^{LM} es factible y hay un escalar $\lambda \geq 0$ tal que

$$\begin{array}{ll} (J^T J + \lambda I)p^{LM} & = -J^T r, \\ \lambda(\triangle - \|p^{LM}\|) & = 0. \end{array} \tag{5.25}$$

Prueba:
A partir del Teorema (14), se cumple que $(J^T J + \lambda I)$ es semidefinida positiva ya que $J^T J$ es semidefinida positiva y $\lambda \geq 0$. Las dos restantes condiciones (5.25) se siguen a partir de las dos primeras condiciones del Teorema (14), respectivamente.

□

Tenga en cuenta que las ecuaciones (5.24) son sólo las ecuaciones normales para el siguiente problema lineal de mínimos cuadrados:

$$\min_p \frac{1}{2} \left\| \begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix} p + \begin{bmatrix} r \\ 0 \end{bmatrix} \right\|^2. \tag{5.26}$$

Al igual que en el caso de Gauss-Newton, la equivalencia entre (5.24) y (5.26) nos da una manera de resolver el subproblema sin calcular el producto matriz-matriz $J^T J$ y su factorización Cholesky.

Implementación del método Levenberg-Marquardt

Para encontrar un valor de λ que coincida aproximadamente con el $\triangle$ dado en el Lema(4), podemos utilizar el algoritmo de búsqueda de raíz que se describe en el capítulo 5. Es fácil salvaguardar este procedimiento: El factor Cholesky R se garantiza en la existencia de cada vez que la estimación actual $\lambda^{(l)}$ sea positiva, ya que el aproximado del Hessiano $B = J^T J$ es semidefinido positivo. Debido a la estructura especial de B, no necesitamos calcular la factorización Cholesky de $B + \lambda I$ desde cero en cada iteración del algoritmo 5.1. Más bien, se presenta una técnica eficaz para la búsqueda de la siguiente factorización QR de la matriz de coeficientes en (5.16):

$$\begin{bmatrix} R_\lambda \\ 0 \end{bmatrix} = Q_\lambda^T \begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix} \tag{5.27}$$

(Q_λ Ortogonal, R_λ triangular superior). El factor triangular superior R_λ satisface $R_\lambda^T R_\lambda = (J^T J + \lambda I)$. Podemos ahorrar tiempo en la computadora en el cálculo de la factorización (5.27)

mediante el uso de una combinación de Householder y transformaciones Givens. Supongamos que utilizamos transformaciones Householder para calcular la factorización QR de J solo como

$$J = Q \begin{bmatrix} R \\ 0 \end{bmatrix}. \tag{5.28}$$

Entonces tenemos

$$\begin{bmatrix} R \\ 0 \\ \sqrt{\lambda} I \end{bmatrix} = \begin{bmatrix} Q^T & \\ & I \end{bmatrix} \begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix} \tag{5.29}$$

La matriz de la izquierda en esta fórmula es triangular superior excepto por los n términos distintos de cero de la matriz λI. Estos pueden ser eliminados por una secuencia de $n(n+1)/2$ rotaciones dadas, en el que se utilizan los elementos diagonales de la parte triangular superior para eliminar los zeros de λI y el relleno en términos que surgen en el proceso. Los primeros pocos pasos para este proceso son los siguientes:

- rotar la fila n de $\mathbf{R}$ con la fila n de $\sqrt{\lambda} I$, para eliminar el elemento (n, n) de $\sqrt{\lambda} I$.
- rotar la fila $(n-1)$ de $\mathbf{R}$ con la fila $(n-1)$ de $\sqrt{\lambda} I$, para eliminar el elemento $(n-1, n-1)$ de esta última matriz. Este paso introduce relleno en la posición $(n-1, n)$ de $\sqrt{\lambda} I$, que se elimina mediante la rotación de la fila n de $\mathbf{R}$ con la fila $(n-1)$ de $\sqrt{\lambda} I$, para eliminar el elemento de relleno en la posición $(n-1, n)$;
- rotar la fila $(n-2)$ de $\mathbf{R}$ con la fila $(n-2)$ de $\sqrt{\lambda} I$, para eliminar la diagonal $(n-2)$ en esta última matriz. Este paso introduce la fila en la $(n-2, n-1)$ y $(n-2, n)$ posiciones , que eliminamos y asi sucesivamente.

Si reunimos todas las rotaciones Givens en una matriz $\bar{Q}_\lambda$, se obtiene a partir de (5.29) que

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

y por lo tanto (5.27) cumple con

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

La ventaja de este enfoque combinado es que cuando el valor de λ se cambia en el algoritmo de búsqueda de raíces, sólo tenemos que recalcular $\bar{Q}_\lambda$ en lugar de la factorización (5.29). Esta característica puede ahorrar mucho la computación en el caso cuando $m \gg n$, se requieren $O(n^3)$ operaciones para recalcular $\bar{Q}_\lambda$ y R_λ para cada valor de λ, después de que el costo inicial de $O(mn^2)$ operaciones necesarias para calcular Q en (5.28).

Problemas de mínimos cuadrados son a menudo mal escalados. Algunas de las variables tienen valores de aproximadamente: 10^4, mientras que otras variables podrían ser de orden de 10^{-6}. Si se ignoran tales variaciones de ancho, los algoritmos anteriormente pueden encontrar dificultades numéricas o producir soluciones de mala calidad. Una forma de reducir

los efectos de una mala escala es utilizar una región factible elipsoidal en lugar de una región esférica factible. Analíticamente, el subproblema de región factible se convierte en

$$\min_p \frac{1}{2}\|J_k p + r_k\|^2, \quad \text{sujeto a} \quad \|D_k p\| \leq \triangle_k, \tag{5.30}$$

donde D_k es una matriz diagonal con entradas diagonales positivas. En lugar de (5.24), la solución de (5.30) satisface una ecuación de la forma

$$(J_k^T J_k + \lambda D_k^2) p_k^{LM} = -J_k^T r_k, \tag{5.31}$$

y equivalentemente, resuelve el problema de mínimos cuadrados lineales

$$\min_p \left\| \begin{bmatrix} J_k \\ \sqrt{\lambda} D_k \end{bmatrix} p + \begin{bmatrix} r_k \\ 0 \end{bmatrix} \right\|^2. \tag{5.32}$$

Las diagonales de la matriz D_k pueden cambiar de una iteración a otra, cuando reunimos información sobre el rango típico de valores para cada componente de x. Si la variación de estos elementos se mantiene dentro de ciertos límites, entonces la teoría de la convergencia para el caso esférica continúa llevandose a cabo, con modificaciones menores. Por otra parte, la técnica descrita anteriormente para el cálculo de R_λ no necesita modificación.

Capítulo 6

Aplicaciones

6.1. Estimación de los parámetros β y γ en el modelo SIR

Para comenzar replanteamos el modelo SIR:

$$\frac{dS}{dt} = -\beta SI,\ \frac{dI}{dt} = \beta SI - \gamma I,\ \frac{dR}{dt} = \gamma I.$$

Donde se describe el movimiento entre susceptibles, infecciosos y recuperados y además β y γ son parámetros desconocidos.

Entonces primero veamos la función de optimización (*fminsearch*) de MATLAB para luego ver la implementación de Levenberg-Marquardt en Octave y así poder observar la variabilidad entre ambos algoritmos.

Primero hacemos el llamado de la función `sir_ode.m`, que contiene las tres ecuaciones diferenciales.

```
function dydt = sir_ode(t,y,p)
   beta = p(1);
   gamma = p(2);
   S = y(1);
   I = y(2);
   R = y(3);
   dydt = [-beta*I*S; beta*I*S - gamma*I; gamma*I];
```

% Como entrada en la ventana se tiene a `data` y `time`; `data` es el número de individuos infectados en un tiempo `time`.

```
data = [3 6 25 73 222 294 258 237 191 125 69 27 11 4];
time = 1:14;
```

% Condiciones iníciales: 760 susceptibles, 3 infectados, 0 recuperados.

```
y0 = [760 3 0];
```

% `tspan` tiene el tiempo para resolver el sistema de EDOs. Necesitamos resolver en el tiempo para nuestro conjunto de datos.

```
tspan = time;
```

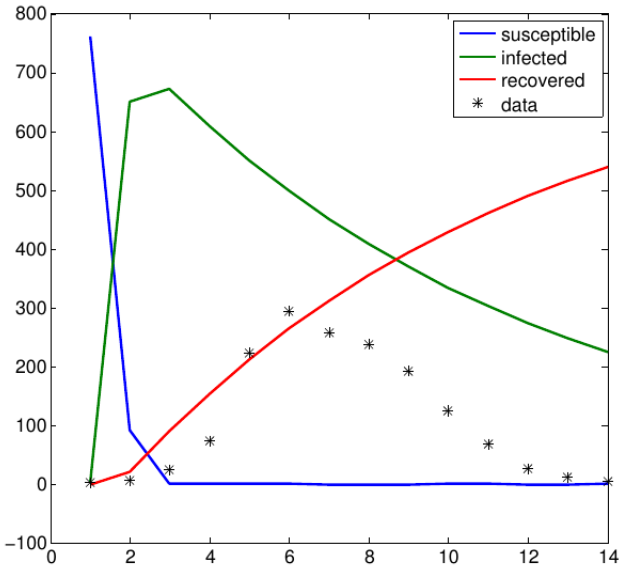
Ahora pedimos a MATLAB resolver el sistema SIR para los valores de los parámetros dados.

% Resuelve el sistema de EDO suponiendo unos buenos valores de los parámetros.

```
p0 = [.01 .1];
[t,y] = ode45(@sir_ode,tspan,y0,[],p0);
```

Para ver cual es el resultado de aprobación en MATLAB (para las variables y y t), graficamos la solución para los valores de los parámetros ($\beta = 0.01$ y $\gamma = 0.1$).

```
figure(1), clf;
plot(t, y, 'linewidth', 2);
hold on;
plot(time, data, 'k*', 'markersize', 10);
legend('susceptible','infected','recovered','data')
set(gca, 'FontSize', 15)
```



Donde podemos observar que los datos reales estan bien alejados de la aproximación. Por lo que para hacer parecidos los datos decidimos cuantificar con un modelo apto los datos dando nosotros los parámetros escogidos. Por lo que decidimos utilizar la suma de cuadrados de las discrepancias.

```
function disc = sir_discrepancy(p, data, tspan, y0)
[t,y] = ode45(@sir_ode,tspan,y0,[],p);
I = y(:,2);
disc = sum((I-data').^2);
```

Ahora necesitamos disminuir la discrepancia, para ello hacemos uso de optimización numérica de los parámetros usando la función *fminsearch* de MATLAB que encuentra el minimizador local de la función en estudio.

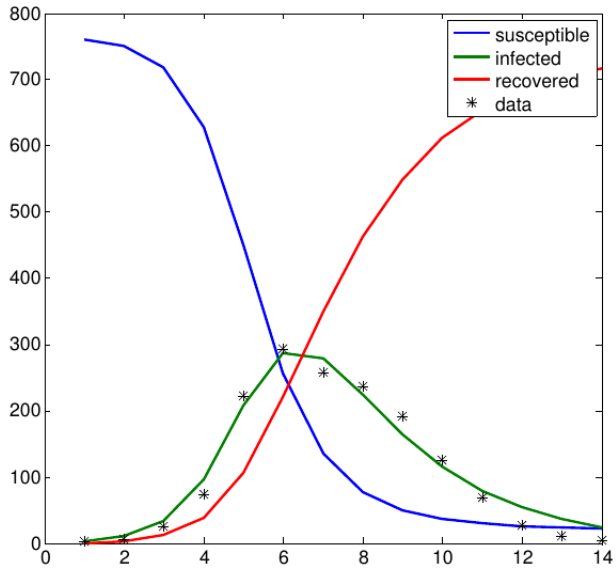
Nuestra función `discrepancy` toma varias entradas (`p, data, tspan, y0`), mientras que la función *fminserch* optimiza solo con una entrada (`p0`).

```
function p_opt = sir_optimize(data, tspan, y0, p0)
p_opt = fminsearch(@sir_disc_nested, p0);
    function disc = sir_disc_nested(p)
      disc = sir_discrepancy(p, data, tspan, y0);
    end
end
```

El cual podemos correrlo.

```
p_opt = sir_optimize(data, tspan, y0, p0);
```

Ahora teniendo los valores optimos de $\beta = 0.0022$ y $\gamma = 0.4485$ resolvemos el sistema SIR de EDO y graficamos su solución, la cual es mucho mejor.



El código en Octave de Levenberg-Marquardt se presenta en los anexos, por lo que aquí solo presentamos los resultados obtenidos. Para finalizar mostramos las tablas de comparación de ambos algoritmos del tiempo (en segundos) que tardan para resolver el problema de optimización de mínimos cuadrados. En donde se muestra que *fminsearch* de MATLAB tiene mayor rango de convergencia que Levenberg-Marquardt de Octave.

Parámetros $\downarrow (\gamma)$, $\longrightarrow (\beta)$	**0.01**	**0.005**	**0.0**	**-0.01**
0.1	0.3312	0.3125	0.2641	0.3156
0.05	0.3891	0.3312	**0.2594**	0.2641
0.0	0.3828	0.4844	0.3250	0.3922

Tabla 1: *Función fminsearch de MATLAB.*

Parámetros $\downarrow(\gamma)$, $\longrightarrow(\beta)$	**0.01**	**0.005**	**0.0**	**-0.01**
0.1	1.0468	0.8812	3.2436	0.2780
0.05	1.2530	0.9720	3.1656	0.2780
0.0	1.7250	1.0374	3.1408	0.2718

Tabla 2: *Método Levenberg-Marquardt con Octave.*

En la segunda tabla podemos notar que la primera columna no se parece a la primera, la segunda igual, la tercera se aleja más mientras que en la cuarta, aunque los valores son pequeños pero converge a otro mínimo. Por lo que concluimos que *fminsearch* es la mejor opción para el caso.

Conclusiones.

El método más óptimo para la estimación de parámetros en modelos epidemilógicos, en particular en el modelo SIR resulto ser la función de optimización de problemas de mínimos cuadrados *fminsearch* de MATLAB que presentó mayor tasa de convergencia en comparación con el método Levenberg-Marquardt de Octave.

Para estudiar métodos de optimización es de mucha importancia tener claro los conceptos básicos y teoremas fundamentales de optimización para poder encontrar la mejor solución del problema que se tiene en estudio.
Programar optimización en modelos epidemiológicos es más complejo de lo que habíamos visto, ya que tenemos problemas de escala. También hemos estudiado que podríamos aplicar Levenberg-Marquardt de una mejor forma, al insertar una matriz D_k para no estar buscando en una circunferencia sino que en una elipse, pero para esto necesitaríamos más tiempo y tal vez con estos cambios podríamos hacer que Levenberg-Marquardt pueda ser mejor que *fminsearch* de MATLAB.

También el programa de optimización es bastante sensible a los cambios; nosotros tenemos los parámetros que están cercanos a cero y eso si es otra dificultad porque pequeños cambios nos pueden lanzar a otros mínimos. Específicamente Levenberg-Marquardt nos está lanzando a otros lugares mientras que *fminsearch* nos está manteniendo en el lugar correcto.

MATLAB y Octave son programas que tienen un enfoque parecido en términos de programación. Sin embargo Octave es un programa que es libre, mientras que MATLAB tiene licencia pagada y comprar una licencia en MATLAB profesional cuesta $1000, también tiene sus paquetes de optimización, pero hay que pagar otros $200 a $300. Entonces a pesar de que la función *leasqr* de Octave nos da resultados inferiores a los de Matlab, pero son aceptables. Es cierto que la región de convergencia es más pequeña y que se está tardando mucho más tiempo, sin embargo es un recurso que es libre. Y esto dependiendo de los recursos financieros que una persona pueda tener o no: si no tiene $1000 entonces la mejor opción es Octave a pesar que a la hora de los resultados hay que ser más cuidadosos, pero es la única opción y si tiene los recursos, entonces utilizar MATLAB (*fminsearch*) es mucho mejor, que está basado en Simplex para problemas no lineales.

La implementación de estos algoritmos es bastante sofisticada, sólo hay que ver el que está ocupando Octave de Levenberg-Marquardt, que son 800 líneas, tal vez son 400 líneas efectivas de código, sin embargo la implementación de éstos, solamente la implementación ya es casi un trabajo de tesis de licenciatura.

Las condiciones Wolfe y Goldstein son condiciones que aseguran que la longitud de paso α_k logra suficiente disminución de f. Además se observa que las condiciones Wolfe son condiciones invariantes en escala, pues al multiplicar la función objetivo por una constante o hacer un cambio afín de variables no les altera.

Las matemáticas en epidemiología y en general en las ciencias biológicas constituyen, además de una herramienta, una forma de pensar y estructurar predicciones, descripciones y explicaciones de procesos. Por ello, tanto en epidemiología como en otras áreas del conocimiento biológico, las matemáticas son utilizadas para modelar.

Bibliografía

[1] Jeffrey R, and Chasnov. ***Mathematical Biology*** . 2nd ed. The Hong Kong University of Science and Technology: n.p., 2009-2014.

[2] Edwards, C., Henry y Penney, and David E. ***Ecuaciones Diferenciales y problemas con valores en la frontera***. 4th ed. PEARSON EDUCACION, S.A., México,: n.p., 2009.

[3] Jorge Nocedal Stephen, and J. Wright. ***Numerical Optimization*** . 2nd ed. N.p.: n.p., n.d.

[4] German Andres Oliverios Patiño. ***Trabajo De Investigación Realizado En Texas A&M University***. N.p.: n.p., n.d.

[5] Wikipedia contributors. ***"Buckingham*** Π ***theorem."*** *Wikipedia, The Free Encyclopedia.* Wikipedia, The Free Encyclopedia, 6 Apr. 2015. Web. 27 Apr. 2015.

[6] Wikipedia contributors. ***"Levenberg–Marquardt algorithm."*** *Wikipedia, The Free Encyclopedia.* Wikipedia, The Free Encyclopedia, 23 Mar. 2015. Web. 27 Apr. 2015.

[7] Wikipedia contributors. ***"Lotka–Volterra equation."*** *Wikipedia, The Free Encyclopedia.* Wikipedia, The Free Encyclopedia, 20 Apr. 2015. Web. 27 Apr. 2015.

[8] http://www.rcim.sld.cu/revista 13/articulos htm/modelosir.htm.

[9] Kenneth Levenberg (1944). .ª Method for the Solution of Certain Non-Linear Problems in Least Squares". Quarterly of Applied Mathematics 2: 164–168.

[10] Donald Marquardt (1963). .ªn Algorithm for Least-Squares Estimation of Nonlinear Parameters". SIAM Journal on Applied Mathematics 11 (2): 431–441. doi:10.1137/0111030.

[11] Jose Pujol (2007). "The solution of nonlinear inverse problems and the Levenberg-Marquardt method". Geophysics (SEG) 72 (4): W1–W16. doi:10.1190/1.2732552.

[12] D.G.LUENBERGER: Programación lineal y no lineal. Addison- Wesley Iberoamericana, 1989.

[13] Ángel Santos, Transparencias de Métodos Numéricos, Curso 2002/2003.

Anexos.

A continuación se muestra el código en Octave de Levenberg-Marquard.

```
function [xf, S, cnt] = LMFsolve(varargin)
% LMFSOLVE  Solve a Set of Nonlinear Equations in Least-Squares
   Sense.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% A solution is obtained by a shortened Fletcher version of the
% Levenberg-Maquardt algoritm for minimization of a sum of squares
%  of equation residuals.
%
% [Xf, Ssq, CNT] = LMFsolve(FUN,Xo,Options)
% FUN     is a function handle or a function M-file name that
   evaluates
%         m-vector of equation residuals,
% Xo      is n-vector of initial guesses of solution,
% Options is an optional set of Name/Value pairs of control
   parameters
%         of the algorithm. It may be also preset by calling:
%         Options = LMFsolve('default'), or by a set of Name/Value
   pairs:
%         Options = LMFsolve('Name',Value, ... ), or updating the
   Options
%                   set by calling
%         Options = LMFsolve(Options,'Name',Value, ...).
%
%    Name   Values {default}          Description
% 'Display'     integer     Display iteration information
%                            {0}  no display
%                             k   display initial and every k-th
   iteration;
% 'FunTol'      {1e-7}      norm(FUN(x),1) stopping tolerance;
% 'XTol'        {1e-7}      norm(x-xold,1) stopping tolerance;
% 'MaxIter'     {100}       Maximum number of iterations;
% 'ScaleD'                  Scale control:
%               value        D = eye(m)*value;
%               vector       D = diag(vector);
%                {[]}        D(k,k) = JJ(k,k) for JJ(k,k)>0, or
%                                   = 1 otherwise,
%                                     where JJ = J.'*J
% Not defined fields of the Options structure are filled by default
   values.
```

```
%
% Output Arguments:
% Xf        final solution approximation
% Ssq       sum of squares of residuals
% Cnt       >0          count of iterations
%           -MaxIter,   did not converge in MaxIter iterations

% Example:  Rosenbrock valey inside circle with unit diameter
%   R  = @(x) sqrt(x'*x)-.5;     %   A distance from the radius r=0.5
%   ros= @(x) [ 10*(x(2)-x(1)^2); 1-x(1); (R(x)>0)*R(x)*1000];
%   [x,ssq,cnt]=LMFsolve(ros,[-1.2,1],'Display',1,'MaxIter',50)
% returns   x = [0.4556; 0.2059],  ssq = 0.2966,  cnt = 18.
%
% Note:   Users with old MATLAB versions (<7), which have no
   anonymous
% functions implemented, should call LMFsolve with named function
   for
% residuals. For above example it is
%   [x,ssq,cnt]=LMFsolve('rosen',[-1.2,1]);
% where the function rosen.m is of the form
%   function r = rosen(x)
%%   Rosenbrock valey with a constraint
%   R = sqrt(x(1)^2+x(2)^2)-.5;
%%   Residuals:
%   r = [ 10*(x(2)-x(1)^2)  %   first part
%         1-x(1)            %   second part
%         (R>0)*R*1000.     %   penalty
%       ];
% Reference:
% Fletcher, R., (1971): A Modified Marquardt Subroutine for
   Nonlinear Least
% Squares. Rpt. AERE-R 6799, Harwell

% Miroslav Balda,
% balda AT cdm DOT cas DOT cz
% 2007-07-02    v 1.0
% 2008-12-22    v 1.1 * Changed name of the function in LMFsolv
%                     * Removed part with wrong code for use of
   analytical
%                       form for assembling of Jacobian matrix
% 2009-01-08    v 1.2 * Changed subfunction printit.m for better one
   , and
%                       modified its calling from inside LMFsolve.
%                     * Repaired a bug, which caused an inclination
   to
%                       istability, in charge of slower convergence.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%   OPTIONS
    %%%%%%%
```

```
%               Default Options
if nargin==1 && strcmpi('default',varargin(1))
   xf.Display  = 0;          %   no print of iterations
   xf.MaxIter  = 100;        %   maximum number of iterations allowed
   xf.ScaleD   = [];         %   automatic scaling by D = diag(diag(J
      '*J))
   xf.FunTol   = 1e-7;       %   tolerace for final function value
   xf.XTol     = 1e-4;       %   tolerance on difference of x-
      solutions
   return

%               Updating Options
elseif isstruct(varargin{1}) % Options=LMFsolve(Options,'Name','
   Value',...)
    if ~isfield(varargin{1},'Display')
        error('Options Structure not correct for LMFsolve.')
    end
    xf=varargin{1};           %   Options
    for i=2:2:nargin-1
        name=varargin{i};     %   Option to be updated
        if ~ischar(name)
            error('Parameter Names Must be Strings.')
        end
        name=lower(name(isletter(name)));
        value=varargin{i+1}; %   value of the option
        if strncmp(name,'d',1), xf.Display  = value;
        elseif strncmp(name,'f',1), xf.FunTol   = value(1);
        elseif strncmp(name,'x',1), xf.XTol     = value(1);
        elseif strncmp(name,'m',1), xf.MaxIter  = value(1);
        elseif strncmp(name,'s',1), xf.ScaleD   = value;
        else   disp(['Unknown Parameter Name --> ' name])
        end
    end
    return
    %               Pairs of Options
elseif ischar(varargin{1})   % check for Options=LMFSOLVE('Name',
   Value,...)
   Pnames=char('display','funtol','xtol','maxiter','scaled');
   if strncmpi(varargin{1},Pnames,length(varargin{1}))
      xf=LMFsolve('default'); % get default values
      xf=LMFsolve(xf,varargin{:});
      return
   end
end

%   LMFSOLVE(FUN,Xo,Options)
    %%%%%%%%%%%%%%%%%%%%%%%%%

FUN=varargin{1};             %   function handle
if ~(isvarname(FUN) || isa(FUN,'function_handle'))
```

```
   error('FUN Must be a Function Handle or M-file Name.')
end

xc=varargin{2};              %   Xo

if nargin>2                  %   OPTIONS
    if isstruct(varargin{3})
        options=varargin{3};
    else
        if ~exist('options','var')
            options = LMFsolve('default');
        end
        for i=3:2:size(varargin,2)-1
            options=LMFsolve(options, varargin{i},varargin{i+1});
        end
    end
else
    if ~exist('options','var')
        options = LMFsolve('default');
    end
end
x  = xc(:);
lx = length(x);

r  = feval(FUN,x);             % Residuals at starting point
%~~~~~~~~~~~~~~~~~
S  = r'*r;
epsx = options.XTol(:);
epsf = options.FunTol(:);
if length(epsx)<lx, epsx=epsx*ones(lx,1); end
J = finjac(FUN,r,x,epsx);
%~~~~~~~~~~~~~~~~~~~~~~
nfJ = 2;
A = J.'*J;                   % System matrix
v = J.'*r;

D = options.ScaleD;
if isempty(D)
    D = diag(diag(A));       % automatic scaling
    for i = 1:lx
        if D(i,i)==0, D(i,i)=1; end
    end
else
    if numel(D)>1
        D = diag(sqrt(abs(D(1:lx)))); % vector of individual scaling
    else
        D = sqrt(abs(D))*eye(lx);     % scalar of unique scaling
    end
end
```

```
Rlo = 0.25;
Rhi = 0.75;
l=1;      lc=.75;     is=0;
cnt = 0;
ipr = options.Display;
printit(ipr,-1);                  %   Table header
d = options.XTol;                 %   vector for the first cycle
maxit = options.MaxIter;          %   maximum permitted number of
   iterations
while cnt<maxit && ...            %   MAIN ITERATION CYCLE
    any(abs(d) >= epsx) && ...        %%%%%%%%%%%%%%%%%%%%
    any(abs(r) >= epsf)
    d  = (A+l*D)\v;               %   negative solution increment
    xd = x-d;
    rd = feval(FUN,xd);
%   ~~~~~~~~~~~~~~~~~~
    nfJ = nfJ+1;
    Sd = rd.'*rd;
    dS = d.'*(2*v-A*d);           %   predicted reduction

    R  = (S-Sd)/dS;
    if R>Rhi                      %   halve lambda if R too high
        l = l/2;
        if l<lc, l=0; end
    elseif R<Rlo                  %   find new nu if R too low
        nu = (Sd-S)/(d.'*v)+2;
        if nu<2
            nu = 2;
        elseif nu>10
            nu = 10;
        end
        if l==0
            lc = 1/max(abs(diag(inv(A))));
            l  = lc;
            nu = nu/2;
        end
        l = nu*l;
    end
        cnt = cnt+1;
    if ipr~=0 && (rem(cnt,ipr)==0 || cnt==1) %   print iteration?
        printit(ipr,cnt,nfJ,S,x,d,l,lc)
    end

    if Sd<S
        S = Sd;
        x = xd;
        r = rd;
        J = finjac(FUN,r,x,epsx);
%       ~~~~~~~~~~~~~~~~~~~~~~~~
        nfJ = nfJ+1;
```

```
        A = J'*J;
        v = J'*r;
    end
end %   while

xf = x;                             %   final solution
if cnt==maxit
    cnt = -cnt;
end                                 %   maxit reached
rd = feval(FUN,xf);
nfJ = nfJ+1;
Sd = rd.'*rd;
if ipr, disp(' '), end
printit(ipr,cnt,nfJ,Sd,xf,d,l,lc)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%   FINJAC        numerical approximation to Jacobi matrix
%   %%%%%%
function J = finjac(FUN,r,x,epsx)
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
lx=length(x);
J=zeros(length(r),lx);
for k=1:lx
    dx=.25*epsx(k);
    xd=x;
    xd(k)=xd(k)+dx;
    rd=feval(FUN,xd);
%   ~~~~~~~~~~~~~~~~
    J(:,k)=((rd-r)/dx);
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


function printit(ipr,cnt,res,SS,x,dx,l,lc)
%        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~  Printing of intermediate
    results
%  ipr <  0  do not print lambda columns
%      =  0  do not print at all
%      >  0  print every (ipr)th iteration
%  cnt = -1  print out the header
%         0  print out second row of results
%        >0  print out first row of results
if ipr~=0
   if cnt<0                 %   table header
      disp('')
      disp(char('*'*ones(1,75)))
      fprintf('  itr  nfJ   SUM(r^2)        x           dx');
      if ipr>0
          fprintf('           l           lc');
```

```
      end
      fprintf('\n');
      disp(char('*'*ones(1,75)))
      disp('')
   else                        %   iteration output
      if rem(cnt,ipr)==0
          f='%12.4e ';
          if ipr>0
             fprintf(['%4.0f %4.0f ' f f f f f '\n'],...
                 cnt,res,SS, x(1),dx(1),l,lc);
          else
             fprintf(['%4.0f %4.0f ' f f f '\n'],...
                 cnt,res,SS, x(1),dx(1));
          end
          for k=2:length(x)
             fprintf([blanks(23) f f '\n'],x(k),dx(k));
          end
      end
   end
end
```