OPTIMIZACIÓN NUMÉRICA

Departamento de Matemática Aplicada
Universidad de Salamanca

Luis Ferragut Canals

15-mayo-2017

Índice general

Capítulo 1

Elementos de Cálculo Diferencial en Espacios Normados

1.1. Espacios Normados

En esta sección y las siguientes consideraremos espacios vectoriales sobre el cuerpo de los reales.

Definición 1.1 Un Espacio Normado es un espacio vectorial E en el que se ha definido una aplicación

$$\begin{aligned} ||.|| : \quad E &\longrightarrow \mathcal{R} \\ v &\longrightarrow ||v|| \end{aligned}$$

verificando las propiedades,

- N1: $||v|| \geq 0 \quad \forall\, v \in E$ y $||v|| = 0$ solo si $v = 0$.
- N2: $||\lambda v|| = |\alpha|||v|| \quad \forall \lambda \in \mathbb{R} \quad \forall v \in E$.
- N3: $||u + v|| \leq ||u|| + ||v|| \quad \forall u, v \in E$ (Desigualdad Triangular).

La aplicación $||.||$ se llama norma y $||v||$ se lee norma del vector v. Todo espacio normado es evidentemente un espacio métrico con la distancia $d(u, v) = ||v - u||$. Todos los conceptos métricos y topológicos tendrán aquí su significado. Hablaremos pues de conjuntos cerrados y conjuntos abiertos , sucesiones convergentes, sucesiones de Cauchy, conjuntos compactos, etc.

Ejemplos: Sobre el espacio vectorial $E = \mathbb{R}^d$ podemos definir las siguientes normas

1. Norma l_2: $||v||_2 = (\sum v_i^2)^{1/2}$

2. Norma l_1: $||v||_1 = \sum |v_i|$
3. Norma l_∞: $||v||_\infty = \max |v_i|$
4. Norma $l_p,\ p \geq 1$: $||v||_p = (\sum |v_i|^p)^{1/p}$

En el espacio vectorial de las funciones continuas en un intervalo cerrado $[a, b]$, $E = C[a, b]$

1. Norma L^2: $||v||_{0,2,[a,b]} = (\int_a^b |v(x)|^2\, dx)^{1/2}$
2. Norma L^1: $||v||_{0,1,[a,b]} = \int_a^b |v(x)|\, dx$
3. Norma L^∞: $||v||_{0,\infty,[a,b]} = \max_{x\in[a,b]} |v(x)|$
4. Norma L^p: $||v||_{0,p,[a,b]} = (\int_a^b |v(x)|^p\, dx)^{1/p} \quad p \geq 1$

Ejercicios:

1. Verificar que l_1, l_∞ son efectivamente una norma en $\mathbb{R}^d$
2. Verificar que L^1, L^∞ son efectivamente una norma en $C[a, b]$

Dejamos para más adelante la verificación de que l_2 y L^2 son normas.

Recordemos algunas definiciones y propiedades fundamentales.

Nociones de topología en espacios normados

Definición 1.2 Sea E un espacio normado. Una bola (abierta) de centro $a \in E$ y radio $r \in \mathbb{R}$ es el conjunto

$$\mathcal{B}_r(a) = \{v \in E; \quad ||v - a|| < r\}$$

Definición 1.3 Entorno: Un conjunto es un entorno de un punto $a \in E$ si contiene a una bola de centro a.

Definición 1.4 Abierto: Un abierto es un conjunto que es entorno de todos sus puntos.

Definición 1.5 Cerrado: Un conjunto es cerrado si es complementario de un abierto.

Definición 1.6 Conjunto acotado: Un conjunto en E es acotado si existe una bola que lo continene.

Definición 1.7 Conjunto compacto: Un conjunto es compacto si de todo recubrimirnto abierto se puede obtener un subrecubrimiento finito.

Comentario 1.1 En espacios de dimensión finita (por ejemplo $\mathbb{R}^d$) los conjuntos compactos coinciden con los conjuntos que son cerrados y acotados.

Definición 1.8 Noción de límite de una función y continuidad en un punto $a \in A$. Sea $A \subset E$ un abierto de un espacio normado E. F otro espacio normado. $f : A \to F$ una función.

$$\lim_{v \to a} f(v) = b \Leftrightarrow \forall \varepsilon > 0, \exists \delta \, \text{tal que si} \, ||v - a|| < \delta \Rightarrow ||f(v) - b|| < \varepsilon$$

Diremos que f es continua en un punto $a \in A$ si verifica

$$\lim_{v \to a} f(v) = f(a)$$

Tenemos las siguientes propiedades:

Propiedad 1.1 Sea $f : E \to F$, f es continua (en todos los puntos) si para todo conjunto abierto $U \subset F$, $f^{-1}(U)$ es un conjunto abierto de de E.

Propiedad 1.2 Sea $A \subset E$ un abierto y $f : A \to F$. f es continua si para todo abierto $U \subset F$ existe algún conjunto abierto $V \subset E$ tal que $f^{-1}(U) = V \cap A$.

Propiedad 1.3 Sea $K \subset E$ un conjunto compacto t $f : K \to F$ una función continua entonces $f(K)$ es un conjunto compacto de F.

Definición 1.9 Norma equivalentes: Sea E un espacio normado en el que hemos definido dos normas $||.||$ y $|||.|||$. Diremos que las dos normas son equivalentes si existen dos constantes $C_1 > 0$ y $C_2 > 0$ tales que

$$C_1||v|| \leq |||v||| \leq C_2||v|| \quad \forall v \in E \tag{1.1}$$

Comentario 1.2 Dos normas equivalentes generan la misma topología.

Propiedad 1.4 En espacios de dimensión finita todas las normas son equivalentes

Definición 1.10 Sean E y F dos espacios vectoriales. Una aplicación $T : E \to F$ se dice que es lineal si

$$\begin{aligned} T(u+v) &= T(u) + T(v) \quad \forall u, vinE \\ T(\lambda v) &= \lambda T(v) \quad \forall \lambda \in \mathbb{R} \, \forall v \in E \end{aligned}$$

Para las aplicaciones lineales continuas tenemos la siguiente caracterización:

Teorema 1.1 : Caracterización de las aplicaciones lineales continuas:

Sean E y F dos espacios normados y $T : E \to F$ una aplicación lineal. Entonces si T es continua en el origen $u = 0 \in E$ es continua en todo punto y existe una constante $M \geq 0$ tal que

$$||Tv|| \leq M||v|| \quad \forall v \in E \tag{1.2}$$

Comentario 1.3 Mientras no de lugar a confusión designaremos las diferentes normas de diferentes espacios normados mediante la misma notación $||.||$.

Comentario 1.4 A la más pequeña de las constantes M verificando (1.2) se le designa mediante $||T||$ y es efectivamente una norma sobre el espacio de aplicaciones lineales continuas $\mathcal{L}(E;F)$ de E en F. La norma $||T||$ está caracterizada por

$$||T|| = \sup_{v\in E;\, v\neq 0} \frac{||Tv||}{||v||} \tag{1.3}$$

Comentario 1.5 En espacios de dimensión finita todas las aplicaciones lineales son continuas

Definición 1.11 Aplicaciones bilineales

Sean E y F espacios normados. Una aplicación bilineal es una aplicación

$$\begin{aligned} B : E \times E &\rightarrow F \\ u, v &\rightarrow B(u,v) \end{aligned}$$

que verifica

$$B(\lambda u + \lambda v, w) = \lambda B(u,w) + \mu B(v,w) \quad \forall \lambda, \mu \in \mathbb{R},\ \forall u, v, w \in E$$

$$B(u, \lambda v + \mu w) = \lambda B(u,w) + \mu B(v,w) \quad \forall \lambda, \mu \in \mathbb{R},\ \forall u, v, w \in E$$

Comentario 1.6 Designaremos mediante $|||[u,v]|||$ una norma del vector $[u,v]$ en el espacio vectorial producto $E \times E$. Podemos generar normas en el espacio producto $E \times E$ a partir de cualquier norma en E. Por ejemplo son normas en el espacio producto las siguientes, donde $||.||$ designa una norma en E:

- Norma l_2: $|||[u,v]|||_2 = (||u||^2 + ||v||^2)^{1/2}$
- Norma l_1: $|||[u,v]|||_1 = ||u|| + ||v||$
- Norma l_∞: $|||[u,v]|||_\infty = \max ||u||, ||v||$
- Norma $l_p,\ p \geq 1$: $|||[u,v]|||_p = (||u||^p + ||v||^p)^{1/p}$

Todas las normas anteriores son equivalentes.

Teorema 1.2 : Caracterización de las aplicaciones bilineales continuas:

Sean E y F dos espacios normados y $B : E \times E \rightarrow F$ una aplicación bilineal. Entonces si B es continua en el origen $[u,v] = [0,0] \in E \times E$ es continua en todo punto y existe una constante $M \geq 0$ tal que

$$||B(u,v)|| \leq M||u||.||v|| \quad \forall u, v \in E \tag{1.4}$$

Comentario 1.7 A la más pequeña de las constantes M verificando (1.2) se le designa mediante $||B||$ y es efectivamente una norma sobre el espacio de aplicaciones bilineales continuas $\mathcal{B}(E \times E; F)$ de $E \times E$ en F. La norma $||T||$ está caracterizada por

$$||B|| = \sup_{u,v \in E;\, u,v \neq 0} \frac{||B(u,v)||}{||u||.||v||} \tag{1.5}$$

Comentario 1.8 En espacios de dimensión finita todas las aplicaciones bilineales son continuas.

1.2. Espacios Euclídeos o Prehilbertianos

Definición 1.12 Un Espacio Eucl'ideo es un espacio vectorial en el que se ha definido una aplicación

$$\begin{aligned} (.,.): \quad E \times E &\longrightarrow \mathbb{R} \\ u, v &\longrightarrow (u,v) \end{aligned}$$

que llamaremos producto escalar, verificando las propiedades siguientes:

- P1: $(u,v) = (v,u) \quad \forall u, v \in E$ (Simetría).
- P2.1: $(\lambda u, v) = \lambda(u,v) \quad \forall \lambda \in \mathbb{R}, \forall u, v \in E$
- P2.2: $(u+v, w) = (u,v) + (u,w) \quad \forall u, v, w \in E$
- P3: $(v,v) \geq 0 \quad \forall v \in E;\ (v,v) = 0$ solo si $v = 0$ (Definida positiva)

Comentario 1.9 De la propiedad P1 y P2 se deduce

$$(u, \lambda v) = \lambda(u,v) \quad \forall \lambda \in \mathbb{R}, \forall u, v \in E$$

$$(u, v+w) = (u,v) + (u,w) \quad \forall u, v, w \in E$$

Por tanto el producto escalar es una aplicación bilineal.

Todo espacio eucĺdeo es un espacio normado con la norma $||v|| = (v,v)^{1/2}$. En efecto, las propiedades N1 y N2 se verifican de forma inmediata. Para verificar N3 necesitamos primero la siguiente:

Propiedad 1.5 Desigualdad de Cauchy-Schwarz

$$|(u,v)| \leq ||u||.||v|| \quad \forall u, v \in E \tag{1.6}$$

Demostración: Por la propiedad P3

$$(u - \lambda v, u - \lambda v) \geq 0 \quad \forall \lambda \in \mathbb{R}\, u,\, v \in E$$

de donde

$$f(\lambda) = (u - \lambda v, u - \lambda v) = ||u||^2 - 2\lambda(u, v) + \lambda^2||v||^2 \geq 0$$

$f(.)$ es un polinomio de segundo grado en la variable λ que toma valores mayores o igual que 0. Por tanto el discriminante asociado tiene que ser menor o igual que 0, es decir

$$(u, v)^2 - ||v||^2.||u||^2 \leq 0$$

reordenando y tomando la raiz cuadrada positiva obtenemos la desigualdad buscada (1.6). ■

La desigualdad N3 es ahora fácil de obtener, en efecto,

$$||u + v||^2 = ||u||^2 + 2(u, v) + ||v||^2 \leq ||u||^2 + 2||u||.||v|| + ||v||^2 = (||u|| + ||v||)^2$$

tomando la raiz cuadrada positiva obtenemo N3.

Ejemplos:

1. $(u, v) = \sum u_i v_i$ en $\mathbb{R}^d$ La norma l_2 es la norma asociada a este producto escalar.
2. $(u, v) = \int_a^b u(x)v(x)\,dx$ en $C[a, b]$ La norma L^2 es la norma asociada a este producto escalar.

1.3. Funciones diferenciables y diferencial de una función

Concepto de Diferencial de una función en un punto

Definición 1.13 Sean E y F, espacios normados, A un subconjunto abierto en E no vacío y $f : A \to F$ una aplicación de A en F. Decimos que f es diferenciable en un punto $a \in A$ si existe una aplicación lineal continua designada mediante la notación $Df(a) : E \to F$ verificando

$$\lim_{h \to 0} \frac{||f(a + h) - f(a) - Df(a)h||}{||h||} = 0 \tag{1.7}$$

Es fácil verificar que la aplicación $Df(a)$ si existe es única y la llamaremos diferencial de f en a. Así pues $Df(a) \in \mathcal{L}(E; F)$.

Comentario 1.10 Una aplicación diferenciable en un punto, es continua en este punto.

Diremos que una aplicación $f : A \to F$ es diferenciable en A si es diferenciable en todos los puntos de A.

Ejemplos:

1. Diferencial de una aplicación constante: Una aplicación constante $f(v) = b \forall v \in A$ es diferenciable en todo punto de A y la diferencial es la aplicación nula $Df(a) = O \in \mathcal{E}; \mathcal{F}$ donde $O : E \to F$, con $O(v) = 0 \; \forall v \in E$.

2. Sea $I \subset \mathbb{R}$ un intervalo abierto, $f : I \to \mathbb{R}$ derivable en un punto $a \in I$, entonces f es diferenciable en a y la diferencial es

$$\begin{aligned} Df(a) : \mathbb{R} &\to \mathbb{R} \\ x &\to f'(a)x \end{aligned}$$

3. Diferencial de una aplicación lineal continua: Sea $f \in \mathcal{L}(E; F)$. Entonces f es diferenciable en E y $Df(a) = f$ para todo $a \in E$. En efecto,

$$\frac{||f(a+h) - f(a) - f(h)||}{||h||} = \frac{||0||}{||h||} = 0$$

4. Diferencial de una aplicación bilineal continua: Sea $B : E \times E \to F$ una aplicación bilineal continua. B es diferenciable en todos los puntos de $E \times E$ y la diferencial en un punto $[a, b]$ es

$$\begin{aligned} DB(a,b) : E \times E &\to F \\ [u, v] &\to B(u, b) + B(a, v) \end{aligned}$$

   En efecto,

$$\frac{||B(u+h, v+k) - B(u,v) - B(u,k) - B(h,v)||}{||[h,k]||} = \frac{||B(h,k)||}{||[h,k]||} \leq \frac{||B||.||h||.||k||}{||u|| + ||v||}$$

   donde hemos aplicado la contunuidad de B y hemos tomado como norma en $E \times E$ la norma $||[u, v]|| = ||u|| + ||v|| \; \forall u, v \in E$. Finalmente tomando el límite cuando $||h|| \to 0$ y $||k|| \to 0$ resulta

$$\lim_{h \to 0, k \to 0} \frac{||B||.||h||.||k||}{||h|| + ||k||} \leq \lim_{h \to 0, k \to 0} \frac{||B||.(||h|| + ||k||)^2}{||h|| + ||k||} = 0$$

5. Diferencial de una aplicación cuadrática: Sea $B : E \times E \to F$ una aplicación bilineal continua y simétrica (es decir, $B(u, v) = B(v, u) \; \forall u, v \in E$) y sea $J : E \to F$ una aplicación definida por $J(v)) = B(v, v) \forall v \in E$. Entonces J es diferenciable en todo punto $a \in E$ y la diferencial es

$$\begin{aligned} DJ(a) : E &\to F \\ v &\to 2B(a, v) \end{aligned}$$

6. Un caso particular del ejemplo 4 es el producto escalar: Sea E un espacio prehilbertiano con producto escalar (u, v). Tenemos poniendo $B(u, v) = (u, v)$,

$$\begin{aligned} DB(a,b) : E \times E &\to F \\ [u, v] &\to (u, b) + (a, v) \end{aligned}$$

7. Un caso particular del ejemplo 5 son las funcionales cuadráticas: Sea E un espacio prehilbertiano con producto escalar $(.,.)$ y sea $J : E \to \mathbb{R}$ definida por $J(v) = (v, v)$. J es diferenciable en E y

$$\begin{aligned} DJ(a) : E &\to \mathbb{R} \\ v &\to 2(a, v) \end{aligned}$$

8. Un caso concreto combinación de los anteriores es: Sea M una matriz cuadrada en $\mathbb{R}^{d\times d}$ y $b \in \mathbb{R}^d$, sea $J : \mathbb{R}^d \to \mathbb{R}$ definida por $J(x) = \frac{1}{2}(Mx, x) - (b, x)\ \forall x \in \mathbb{R}^d$, siendo $(.,.)$ el producto escalar habitual en $\mathbb{R}^d$. J es diferenciable en $\mathbb{R}^d$ y

$$\begin{aligned} DJ(a) : \mathbb{R}^d &\to \mathbb{R} \\ x &\to (Ma, x) - (b, x) \end{aligned}$$

9. Un ejemplo análogo al anterior pero en un espacio de dimensión infinita es el siguiente: Sea $E = C[a, b]$ el espacio de funciones continuas en $[a, b]$ con el producto escalar $(u, v) = \int_a^b u(x)v(x)\,dx$, sea $f \in E$ y $J : E \to \mathbb{R}$ definida por $J(v) = \frac{1}{2}\int_a^b v(x)^2\,dx - \int_a^b f(x)v(x)\,dx$. Entonces J es diferenciable en E y

$$\begin{aligned} DJ(u) : E &\to \mathbb{R} \\ v &\to \int_a^b u(x)v(x)\,dx - \int_a^b f(x)v(x)\,dx \end{aligned}$$

Una herramienta fundamental en el cálculo de diferenciales es la regla de la cadena

Teorema 1.3 Sean E, F, S espacios normados. $A \subset E$ un abierto $f : A \to F$, f diferenciable en $a \in A$. Sea un abierto $U \in F$ que contiene a $f(a)$ y sea $g : U \to S$ una función diferenciable en $f(a)$. Entonces la función compuesta $g \circ f : A \to S$ es diferenciable en a y la diferencial viene dada por

$$D(g \circ f)(a) = Dg(f(a) \circ Df(a)$$

Ejemplos:

1. Sea $f : \mathbb{R}^2 \to \mathbb{R}$ definida por

$$f(x, y) = \int_0^{x+y} g(z)\,dz$$

donde $g : \mathbb{R} \to \mathbb{R}$ es una función integrable. Tenemos que f es diferenciable en todo punto de $\mathbb{R}^2$ y la diferencial en un punto $(a, b) \in \mathbb{R}^2$ es

$$Df(a, b) = g(a + b) \circ s$$

donde $s : \mathbb{R}^2 \to \mathbb{R}$ es la función suma, $s(x + y) = x + y$. En efecto, aplicando la regla de la cadena, f se puede escribir como $f = G \circ s$ donde $G : \mathbb{R} \to \mathbb{R}$ es la función $G(x) = \int_0^x g(z)\,dz$.

G es derivable, su derivada en un punto a es $G'(a) = g(a)$, y la diferencial $DG(a) : \mathbb{R} \to \mathbb{R}$ es la aplicación $DG(a)x = g(a)x \ \forall x \in \mathbb{R}$. La función suma s es lineal continua, por tanto su diferencial en cualquier punto es la misma función s. Aplicando la regla de la cadena

$$Df(a,b) = DG(s(a,b) \circ s = DG(a+b) \circ s$$

es decir $Df(a,b)$ es la aplicación

$$\begin{aligned} Df(a,b) : \mathbb{R}^2 &\to \mathbb{R} \\ x, y &\to g(a+b).(x+y) \end{aligned}$$

2. Sea $f : \mathbb{R}^2 \to \mathbb{R}$ definida por

$$f(x,y) = \int_0^{xy} g(z)\, dz$$

donde $g : \mathbb{R} \to \mathbb{R}$ es una función integrable. Tenemos que f es diferenciable en todo punto de $\mathbb{R}^2$. En efecto, f se puede poner como la composición de $p : \mathbb{R}^2 \to \mathbb{R}$, función producto de dos números reales $p(x,y) = xy$ y la función G introducida en el ejemplo anterior. Por la regla de la cadena tendremos

$$Df(a,b) = DG(ab) \circ Dp(a,b)$$

Como p es una función bilineal continua $Dp(a,b)(x,y) = ay + bx \ \forall x, y \in \mathbb{R}$. Finalmente

$$\begin{aligned} Df(a,b) : \mathbb{R}^2 &\to \mathbb{R} \\ x, y &\to g(ab).(ay+bx) \end{aligned}$$

Diferenciales de orden superior

Sean E y F, espacios normados, A un subconjunto abierto en E no vacío y $f : A \to F$ una aplicación de A en F. Supongamos que f es diferenciable en todos los puntos de A, de modo que $Df(a) \in \mathcal{L}(E;F)$. Consideremos la siguiente aplicación:

$$\begin{aligned} Df : A &\to \mathcal{L}(E;F) \\ x &\to Df(x) \end{aligned}$$

Si la aplicación diferencial anterior Df es a su vez diferenciable en un punto a, su diferencial que denotaremos $D^2f(a)$ es un elemento de $\mathcal{L}\big(E;\mathcal{L}(E;F)\big)$ y se llama diferencial segunda de la aplicación f en el punto a. Utilizando el isomorfismo canónico entre $\mathcal{L}\big(E;\mathcal{L}(E;F)\big)$ y el espacio de aplicaciones bilineales continuas de $E \times E$ en F, identificaremos $D^2f(a)$ con una aplicación bilineal continua de $E \times E$ en F. Escribiremos pues,

$$D^2f(a)(u,v) = \big(D^2f(a)(u)\big)(v) \ \forall u, v \in E$$

Comentario 1.11 Se puede demostrar que la aplicación $D^2f(a)$ es simétrica, es decir

$$D^2f(a)(u,v) = D^2f(a)(v,u)\ \forall u, v \in E$$

Reiterando la construccion anterior, podemos definir de forma análoga, la diferencial de orden n de una función $D^nf(a)$ como un elemento del espacio de las aplicaciones n-multilineales de $E \times E \times ...(n)... \times E$ en F. Tendremos,

$$D^nf(a)(u_1, ..., u_n) = \Big(...\big(D^nf(a)(u1)\big)(u_2)...\Big)(u_n)\ \forall u_1, u_2, ..., u_n \in E$$

Formulas de Taylor

Teorema 1.4 Formula de Taylor para funciones dos veces diferenciables.

Sea $A \subset E$ un abierto, $f : A \to F$ y $[a, a+h]$ un segmento cerrado contenido en A.

1. Formula de Taylor-Young: Si f es diferenciable en A y dos veces diferenciable en a, entonces

$$f(a+h) = f(a) + Df(a)h + \frac{1}{2}D^2f(a)(h,h) + ||h||^2\varepsilon(h), \quad \lim_{h\to 0}\varepsilon(h) = 0 \tag{1.8}$$

2. Formula de Taylor-Maclaurin: Si f es diferenciable en A con continuidad y dos veces diferenciable en $(a, a+h)$ y $F = \mathbb{R}$, entonces

$$f(a+h) = f(a) + Df(a)h + \frac{1}{2}D^2f(a+\theta h)(h,h), \quad 0 < \theta < 1 \tag{1.9}$$

3. Formula de Taylor con resto integral: Si f es dos veces diferenciable en A con continuidad y F es un espacio completo, entonces

$$f(a+h) = f(a) + Df(a)h + \int_0^1 (1-t)\big(D^2f(a+th)(h,h)\big)\,dt \tag{1.10}$$

De manera genearal para funciones n veces diferenciables tenemos, por ejemplo, la formula de Taylor con resto de Taylor-Young

$$f(a+h) = f(a) + Df(a)h + ... + \frac{1}{n!}D^nf(a)(h,h,...,h) + ||h||^n\varepsilon(h), \quad \lim_{h\to 0}\varepsilon(h) = 0 \tag{1.11}$$

Derivadas direccionales

Sean E y F, espacios normados, A un subconjunto abierto en E no vacío y $f : A \to F$ una aplicación de A en F.

Definición 1.14 Derivada Direccional: Sea $v \in E$. Si existe el límite en F siguiente

$$\lim_{\lambda \to 0} \frac{f(a+\lambda v) - f(a)}{\lambda} \tag{1.12}$$

lo llamaremos derivada de f en el punto a en la dirección v y lo designaremos mediante la notación $D_v f(a)$.

Propiedad 1.6 Si f es diferenciable en a es fácil ver que

$$D_v f(a) = Df(a)v$$

Tenemos por tanto que si una función es diferenciable en a admite derivadas direccionales en a en cualquier dirección v. El recíproco no es cierto.

Cálculo diferencial en $\mathbb{R}^d$

Consideraremos ahora funciones definidas en $\mathbb{R}^d$ a valores reales. Sea un abierto $A \subset \mathbb{R}^d$, y f una función $f : A \to \mathbb{R}^p$ diferenciable en un punto $a \in A$. La diferencial de $Df(a)$ es una aplicación lineal de $\mathbb{R}^d$ en $\mathbb{R}^p$. Eligiendo bases en $\mathbb{R}^d$ y en $\mathbb{R}^p$ la diferencial $DF(a)$ tendrá una representación matricial que denotamos $f'(a)$ y que llamaremos matriz jacobiana de f en el punto a y que será una matriz de p filas y d columnas.

En particular para una función definida en un abierto $A \subset \mathbb{R}^d$, $f : A \to \mathbb{R}$, diferenciable en $a \in A$, la diferencial de $Df(a)$ es una aplicación lineal de $\mathbb{R}^d$ en $\mathbb{R}$. Si elegimos una base en $\mathbb{R}^d$, por ejemplo, la base canónica $e_i,\ i = 1, ..., d$ con $e_i = (0, .., 0, 1, 0, .., 0)^t$, podemos representar $Df(a)$ como la matriz fila

$$f'(a) = (Df(a)e_1,\ \dots, Df(a)e_d)$$

Para una función $f : A \subset \mathbb{R}^d \to \mathbb{R}^p,\ f(v) \in \mathbb{R}^p$, y elegida una base en $\mathbb{R}^p$ podemos considerar las componentes de $f(v)$, $f(v) = (f_1(v),\ \dots, f_d(v))^t \in \mathbb{R}^p,\ f_i(v) \in \mathbb{R},\ i = 1, ..., p$. Las funciones $f_i : \mathbb{R}^d \to \mathbb{R}$ las llamaremos funciones componentes de f. De forma inmediata, tenemos que f es diferenciable en a si y solo si las funciones componentes f_i son diferenciables en a. La matriz jacobiana es entonces

$$f'(a) = \begin{pmatrix} Df_1(a)e_1 & \dots & Df_1(a)e_d \\ & \dots & \\ Df_p(a)e_1 & \dots & Df_p(a)e_d \end{pmatrix}$$

Consideremos en $\mathbb{R}^d$ la base canónica $e_i,\ i = 1, ..., d$. Y sea una función $f : A \subset \mathbb{R}^d \to \mathbb{R}$ que admita derivadas direccionales. La derivada en la dirección e_i de f en un punto $a \in \mathbb{R}^d$, $D_{e_i} f(a) \in \mathbb{R}$ se llama derivada parcial i-ésima de f en el punto a y la designamos mediante las notaciones $\frac{\partial f}{\partial x_i}(a)$, o bien $\partial_i f(a)$. Tendremos

$$\partial_i f(a) = \lim_{\lambda \to 0} \frac{f(a_1,\ \dots a_{i-1},\ a_i + \lambda,\ a_{i+1},\ \dots, a_d) - f(a_1,\ \dots, a_d)}{\lambda}$$

Si f es diferenciable en a, tenemos $\partial_i f(a) = Df(a)e_i$. Y la matriz jacobiana viene dada por

$$f'(a)) = (\partial_1 f(a), \ldots, \partial_d f(a))$$

En el caso de funciones $f : A \subset \mathbb{R}^d \to \mathbb{R}^p$, la derivada direccional en un punto $D_v f(a)$ es un elemento de $\mathbb{R}^p$, cuyas componentes, una vez elegida una base en $\mathbb{R}^p$ son las derivadas direccionales de las funciones componentes $D_v f_i(a)$ y en particular tendremos para las derivadas parciales

$$\partial_i f(a) = \begin{pmatrix} \partial_i f_1(a) \\ \ldots \\ \partial_i f_p(a) \end{pmatrix}$$

y para la matriz jacobiana

$$f'(a) = \begin{pmatrix} \partial_1 f_1(a) & \ldots & \partial_d f_1(a) \\ & \ldots & \\ \partial_1 f_p(a) & \ldots & \partial_d f_p(a) \end{pmatrix}$$

Comentario 1.12 Regla de la cadena y matrices jacobianas: Consideremos ahora $A \subset \mathbb{R}^d$ un abierto $f : A \to \mathbb{R}^p$, f diferenciable en $a \in A$. Sea un abierto $U \in F$ que contiene a $f(a)$ y sea $g : U \to \mathbb{R}^s$ una función diferenciable en $f(a)$. Entonces la función compuesta $g \circ f : A \to S$ sabemos que es diferenciable en a y que la diferencial en a es $D(g \circ f)(a) = Dg(f(a) \circ Df(a)$. Utilizando matrices jacobianas, la regla de la cadena se escribirá

$$(g \circ f)'(a) = g'(f(a).f'(a)$$

es decir la matriz jacobiana en a de la función compuesta $g \circ f$ es el producto de la matriz jacobiana de g en $f(a)$ y de la matriz jacobiana de f en a.

Comentario 1.13 Matriz jacobiana y gradiente de una función: Sea $f : A \subset \mathbb{R}^d \to \mathbb{R}$ una función diferenciable. Consideremos la $Df(a)$ que es una aplicación lineal continua de $\mathbb{R}^d$ en R, es decir un elemento del espacio $\mathcal{L}(\mathbb{R}^d; \mathbb{R})$, es decir el espacio dual de $\mathbb{R}^d$. Vamos a ver que podemos considerar también $Df(a)$ como un vector de $\mathbb{R}^d$. En efecto, podemos identificar el espacio dual $\mathcal{L}(\mathbb{R}^d; \mathbb{R})$ con $\mathbb{R}^d$ de la siguiente forma: Dado $u \in \mathbb{R}^d$ le asociamos un elemento $T_u \in \mathcal{L}(\mathbb{R}^d; \mathbb{R})$ mediante

$$T_u(v) = (u, v) \; \forall v \in \mathbb{R}^d$$

donde $(.,.)$ es el producto escalar en $\mathbb{R}^d$. Es fácil ver que dado $u \in \mathbb{R}^d$, la aplicación T_u así definida es lineal, por lo tanto un elemento del dual. Además la aplicación

$$\begin{array}{rcl} \mathbb{R}^d & \to & \mathcal{L}(\mathbb{R}^d; \mathbb{R}) \\ u & \to & T_u \end{array}$$

es biyectiva y es de hecho una isometría, es decir, $||u|| = ||T_u|| \; \forall u \in \mathbb{R}^d$. Si consideramos $Df(a)$ como un vector de $\mathbb{R}^d$ a través de la anterior identificación lo llamaremos vector gradiente y lo designamos mediante $\nabla f(a)$. En la base canónica se escribe

$$\nabla f(a) = \begin{pmatrix} \partial_1 f(a) \\ \ldots \\ \partial_d f(a) \end{pmatrix}$$

y tendremos la diferentes formas de expresar el valor de $Df(a)v$:

$$Df(a)v = f'(a).v = (\nabla f(a), v)$$

que se lee: $Df(a)v$ es el valor de la diferencial de f en a aplicado al vector v que es igual a $f'(a).v$ que es el producto de la matriz fila $f'(a)$ por el vector columna v y que finalmente es igual a $(\nabla f(a), v)$ que es el producto escalar del vector $\nabla f(a)$ por el vector v.

Diferenciales de orden superior en $\mathbb{R}^d$

Sea un abierto $A \subset \mathbb{R}^d$, y f una función $f : A \to \mathbb{R}$ diferenciable dos veces en un punto $a \in A$. Tenemos $D^2f(a) \in \mathcal{L}\big(\mathbb{R}^d; \mathcal{L}(\mathbb{R}^d; R)\big)$ o bien $D^2f(a) \in \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R})$, identificando el espacios de aplicaciones lineales $\mathcal{L}\big(\mathbb{R}^d; \mathcal{L}(\mathbb{R}^d; R)\big)$ con el espacio de aplicaciones bilineales $\mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R})$. Para determinar la aplicación $D^2f(a)$ basta conocer su imagen al aplicarla a los elementos de una base $e_i,\ i = 1, \ldots, d$ de $\mathbb{R}^d$. En efecto conocidos los valores de $D^2f(a)(e_i, e_j)$ el valor de $D^2f(a)(u, v)$, para cualquier par de vectores $u = \sum_i u_i,\ v = \sum_i v_i$ es

$$D^2f(a)(u, v) = \sum_{i,j} u_i v_j D^2f(a)(e_i, e_j)$$

La matriz

$$H(a) = \begin{pmatrix} D^2f(a)(e_1, e_1) & \ldots & D^2f(a)(e_1, e_d) \\ & \ldots & \\ D^2f(a)(e_d, e_1) & \ldots & D^2f(a)(e_d, e_d) \end{pmatrix}$$

se llama matriz Hessiana de f en el punto a y es la representación matricial de la Diferencial segunda de f en a.

El cálculo efectivo de $D^2f(a)(u, v)$, se expresa matricialmente como

$$v^t.H(a).u$$

donde aquí v^t es el vector fila formado por las componentes de v en la base elegida y u es el vector columna formado por las componentes de u en la base elegida.

Comentario 1.14 En el caso de una función $f : A \subset \mathbb{R}^d \to \mathbb{R}^p$ con p componenets el concepto anterior se aplica a cada una de las funciones componentes de f.

Derivadas parciales de orden superior y cálculo práctico de la matriz Hessiana

La matriz Hessiana se puede calcular fácilmente con ayuda del concepto de derivada parcial segunda. Sea $f : A \subset \mathbb{R}^d \to \mathbb{R}$ con derivadas parciales en A. La función

$$\begin{aligned} \partial_i f : \mathbb{R}^d &\to \mathbb{R} \\ x &\to \partial_i f(x) \end{aligned}$$

Se llama función derivada parcial i-ésima de f. Si esta función admite sus derivadas parciales $\partial_j(\partial_i f)(a)$ en un punto $a \in A$ este número se llama derivada parcial segunda de f en el punto a. La notaciones habituales son:

$$\partial_j(\partial_i f)(a) = \partial^2_{ji} f(a) = \frac{\partial^2 f}{\partial x_j \partial x_i}(a)$$

Teorema 1.5 Para $f : A \subset \mathbb{R}^d \to \mathbb{R}$ dos veces diferenciable se tiene

$$D^2 f(a)(e_i, e_j) = \partial^2_{ji} f(a) = \partial^2_{ij} f(a)$$

Comentario 1.15 La matriz Hessiana es simétrica.

Análogamente tendremos para las diferenciales de orden n de una función $f : A \subset \mathbb{R}^d \to \mathbb{R}$ en un punto a, la diferencial $D^n f(a)$ es una aplicación multilineal de $\mathcal{M}(\mathbb{R}^d \times \mathbb{R}^d \times ...(n)... \times \mathbb{R}^d; \mathbb{R})$ que vendrá determinada por los d^n valores

$$D^n f(a)(e_{i_1}, , ..., e_{i_n}),\ i_1, ..., i_n = 1, ..., d$$

Se puede demostrar que

$$D^n f(a)(e_{i_1}, , ..., e_{i_n}) = \partial^n_{i_1,...,i_n} f(a),\ i_1, ..., i_n = 1, ..., d$$

Capítulo 2

Fundamentos de la Optimización

2.1. Introducción

En el espacio euclídeo $\mathbb{R}^d$ con la norma euclídea que denotamos $||.||$ y que deriva del correspondiente producto escalar que denotaremos mediante $(.,.)$, consideremos

- K un subconjunto cerrado de $\mathbb{R}^d$
- $J : K \longrightarrow \mathbb{R}$ una función sobre K
- Consideraremos el problema siguiente: Hallar $u \in K$ tal que

$$J(u) = \inf_{v \in K} J(v)$$

- Si $K = \mathbb{R}^d$ el problema anterior se llama de optimización sin restricciones.
- ¿ En qué condiciones el problema anterior tiene solución?
- Si tiene solución ¿ es ésta única?
- Se dice que u realiza el mínimo global de J.

Vamos a tratar de responder a las anteriores preguntas.

Teorema 2.1 de Weierstrass

Sea K un conjunto compacto de $\mathbb{R}^d$, no vacío y $J : K \longrightarrow \mathbb{R}$ continua en K. Entonces el problema:

Hallar $u \in K$ tal que

$$J(u) = \inf_{v \in K} J(v)$$

tiene solución.

Demostración:

K compacto y J continua implica que $J(K)$ es compacto en $\mathbb{R}$, por tanto $J(K)$ está acotado inferiormente y existe un extremo inferior, es decir, existe un número $\alpha > -\infty$ tal que $\alpha = \text{ínf}_{v\in K} J(v)$.

Consideremos una sucesión minimizante $(u_n)_n$, es decir, $u_n \in K$ tal que $J(u_n) \to^{n\to\infty} \alpha$. Una tal sucesión la podemos siempre construir tomando por ejemplo $0 < \varepsilon < 1$, $u_n \in K$ tal que

$$\alpha \leq J(u_n) \leq \alpha + \varepsilon^n$$

$(u_n)_n$ es una sucesión en el compacto K, por lo tanto existe una subsucesión convergente $(u_\nu)_\nu$. Sea u el límite de (u_ν). Gracias a la continuidad de $J(.)$ tendremos $\alpha = \lim_{\nu\to\infty} J(u_\nu) = J(u)$. ■

Una variante del anterior teorema de Weierstrass es el siguiente:

Teorema 2.2 Sea $J : \mathbb{R}^d \to \mathbb{R}$ continua, verificando la propiedad (coercividad)

$$\lim_{||v||\to\infty} J(v) = \infty$$

Entonces existe $u \in \mathbb{R}^d$ tal que

$$J(u) = \inf_{v\in\mathbb{R}} J(v)$$

Demostración:

Llamemos $\alpha = \text{ínf}_{v\in\mathbb{R}^d} J(v)$. En principio α pordría tomar el valor $-\infty$. Consideremos una sucesión $(u_n)_n$ minimizante de elementos de $\mathbb{R}^d$, es decir,

$$J(u_n) \longrightarrow \inf_{v\in\mathbb{R}^d} J(v) = \alpha$$

$$\alpha \leq J(u_n) \leq \alpha + \varepsilon^n$$

si α es finito y si $\alpha = -\infty$ entonces tomaremos u_n de modo que $J(u_n) \to -\infty$.

La sucesión $(u_n)_n$ está acotada, pues si no fuera así, resultaría

$$\lim_{||u_n||\to\infty} J(u_n) = \infty$$

en contra de la elección de $(u_n)_n$.

Podemos pues extraer una subsucesión convergente. Sea $(u_\nu)_\nu$ tal que $\lim u_\nu = u$; como J es continua resulta $\lim J(u_\nu) = J(u)$ y α es un número finito. ■

Ejercicio: Dar 3 contraejemplos en los que el problema anterior no tenga solución debido a que:

1. K no sea compacto.
2. J no sea continua.
3. J no sea coerciva.

2.1.1. Extremos relativos y diferenciabilidad

En este apartado consideraremos la noción de extremo relativo y la relacionaremos con las nociones de diferencial.

Definición 2.1 Sea $A \subset \mathbb{R}^d$ un conjunto abierto y $J : A \longrightarrow \mathbb{R}$. Se dice que $J(.)$ tiene un mínimo relativo en $u \in A$ si existe un entorno U de u tal que

$$\forall v \in U \quad J(u) \leq J(v)$$

Análogamente $J(.)$ tiene un máximo relativo en $u \in A$ si existe un entorno U de u tal que

$$\forall v \in U \quad J(u) \geq J(v)$$

Teorema 2.3 Condición necesaria de extremo relativo

A abierto de $\mathbb{R}^d$, $J : A \longrightarrow \mathbb{R}$ diferenciable en A. Si $J(.)$ tiene un extremo relativo en $u \in A$ (máximo o mínimo) entonces la diferencial de $J(.)$ en u es la aplicación nula, es decir, $DJ(u) = 0$, es decir

$$DJ(u)v = 0 \quad \forall v \in \mathbb{R}^d$$

o bien identificando $DJ(u)$ con un elemento de $\mathbb{R}^d$, $\nabla J(u) = 0$.

Nota 2.1 Referido a la base canónica de $\mathbb{R}^d$, $DJ(u)$ se escribirá $J'(u)$, es decir, la matriz Jacobiana de $J(.)$ en el punto u. En este caso la matriz jacobiana en un matriz fila. La transpuesta de esta matriz es un vector columna que se llama vector gradiente de $J()$ en u y se escribe $\nabla J(u)$. Cuando identifiquemos $DJ(u)$ con un vector de $\mathbb{R}^d$ escribiremos indistintamente $DJ(u)v$ o $(\nabla J(u), v)$ que también podremos escribir con notación matricial como $J'(u).v$

Demostración:

- Caso $d = 1$: Sea A abierto de $\mathbb{R}$, $u \in A$, $f : A \subset \mathbb{R} \longrightarrow \mathbb{R}$, $f(u) \leq f(v) \quad \forall v \in U$, entorno de u. Tendremos, por ejemplo

$$f'(u) = \lim_{h \to 0^+} \frac{f(u+h) - f(u)}{h} \geq 0$$

$$f'(u) = \lim_{h \to 0^-} \frac{f(u+h) - f(u)}{h} \leq 0$$

de donde $f'(u) = 0$.

- Caso general $d > 1$

  Fijemos h con norma suficientemente pequeña de modo que $u+th \in U \quad \forall t \in]-1,+1[$. Introducimos ahora la función

$$f :]-1,+1[\longrightarrow \mathbb{R}$$
$$t \longrightarrow J(u+th)$$

  Sea $f = J \circ g$ donde

$$g : \mathbb{R} \longrightarrow \mathbb{R}^d$$
$$t \longrightarrow u+th$$

  Si J tiene un mínimo relativo en u, es decir, $J(u) \leq J(u+th)$, entonces $f(0) \leq f(t) \quad \forall t \in]-1,1[$. Resulta $f(.)$ tiene un mínimo relativo en 0 y por lo tanto $f'(0) = 0$, es decir, aplicando la regla de la cadena

$$f'(t) = DJ(g(t)) \circ Dg(t) = J'(u+th).g'(t) = J'(u+th).h$$

  de modo que

$$f'(0) = J'(u).h = 0$$

  Como la dirección h es cualquiera resulta $J'(u) = 0$. ■

Vamos ahora a tener en cuenta las derivadas segundas, para obtener una condición necesaria y suficiente de mínimo relativo.

Teorema 2.4 Condición necesaria de mínimo relativo.

Sea A un abierto de $\mathbb{R}^d$, $J : A \longrightarrow \mathbb{R}$ dos veces diferenciable en $u \in A$; Entonces si $J(.)$ admite un mínimo relativo en u, entonces la diferencial segunda de $J(.)$ en u verifica,

$$D^2J(u)(h,h) \geq 0 \quad \forall h \in \mathbb{R}^d$$

o de manera equivalente, la matriz Hessiana de $J(.)$ en u es semidefinida positiva.

Demostración:

Utilizaremos el desarrollo de Taylor con resto de Taylor-Young.

$$J(u+h) = J(u) + DJ(u)(h) + \frac{1}{2}D^2J(u)(h,h) + \varepsilon(h)||h||^2$$

donde $\varepsilon(h) \stackrel{h\to 0}{\longrightarrow} 0$. Como $DJ(u) = 0$ resulta,

$$0 \leq J(u+h) - J(u) = \frac{1}{2}D^2J(u)(h,h) + \varepsilon(h)||h||^2$$

Sustituyendo h por th con $t \in \mathbb{R}$ de modo que $u + th \in A$

$$0 \leq J(u+th) - J(u) = \frac{t^2}{2} D^2 J(u)(h,h) + t^2 \varepsilon(h)||h||^2)$$

Dividiendo por $\frac{t^2}{2}$

$$0 \leq D^2 J(u)(h,h) + 2\varepsilon(h)||h||^2$$

pasando al límite cuando $t \to 0$, $\varepsilon(th) \to 0$ lo que implica $D^2 J(u)(h,h) \geq 0$.■

Vamos a buscar ahora condiciones suficientes de mínimo relativo. Necesitamos primero una definición y un lema.

Definición 2.2 Sea $B : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ una aplicación bilineal. Se dice que B es definida positiva si

$$B(v,v) \geq 0 \quad \forall v \in \mathbb{R}^d$$

y

$$B(v,v) = 0, \quad \text{si y solo si } v = 0$$

Lema 2.1 Sea $B : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ una aplicación bilineal y definida positiva, entonces existe una constante $\alpha > 0$ tal que

$$B(v,v) \geq \alpha ||v||^2$$

Demostración:

Consideremos el conjunto $S = \{v \in \mathbb{R}^d;\ ||v|| = 1\}$ que es compacto (cerrado y acotado) y la aplicación $J : v \to J(v) = B(v,v)$ que es continua sobre el compacto S, entonces alcanza el mínimo en el compacto (teorema de Weierstrass). Sea u el punto donde alcanza este mínimo, es decir,

$$\alpha = J(u) = \min_{v \in S} J(v)$$

tendremos $\alpha \neq 0$ pues $u \in S \Rightarrow u \neq 0$. Finalmente $\forall v \in \mathbb{R}^d \quad v \neq 0$, $\frac{v}{||v||} \in S$ y

$$J(\frac{v}{||v||}) = B(\frac{v}{||v||}, \frac{v}{||v||}) \geq \alpha \Rightarrow B(v,v) \geq \alpha ||v||^2$$

. ■

Teorema 2.5 Condición suficiente de mínimo relativo.

Sea A abierto de $\mathbb{R}^d$ y $J : A \longrightarrow \mathbb{R}$ dos veces diferenciable, tal que $DJ(u) = 0$ en $u \in A$.

Si la función $J(.)$ es tal que su diferencial segunda en u es definida positiva entonces J tiene un mínimo relativo en u.

Demostración:

Consideremos el desarrollo de Taylor-Young en un entorno de u.

$$J(u+h) = J(u) + DJ(u)(h) + \frac{1}{2}D^2J(u)(h,h) + \varepsilon(h)||h||^2$$

donde $\varepsilon(h) \stackrel{h\to 0}{\longrightarrow} 0$

$$J(u+h) - J(u) = \frac{1}{2}D^2J(u)(h,h) + \varepsilon(h)||h||^2$$

$$J(u+h) - J(u) \geq \frac{1}{2}\alpha||h||^2 + \varepsilon(h)||h||^2$$

$$J(u+h) - J(u) \geq \frac{1}{2}(\alpha + 2\varepsilon(h))||h||^2$$

Para $||h||$ suficientemente pequeño tendremos $\frac{1}{2}(\alpha + 2\varepsilon(h)) > 0$, es decir $J(u+h) - J(u) \geq 0$.■

Ejercicio: verificar que $J : A \longrightarrow \mathbb{R}$ donde

$$A = \{(x,y) \in \mathbb{R}^2 \; x^2 + y^2 < 1\}$$

$$J(x,y) = \log(x^2 + y^2 + 1)$$

tiene un mínimo relativo en $(0,0)$.

2.2. Extremos y convexidad

Vamos a introducir la convexidad en el estudio de los extremos de funciones. Recordemos primero las nociones de conjunto convexo y de función convexa.

Definición 2.3 Conjunto convexo.

Un conjunto $K \subset \mathbb{R}^d$ se dice que es convexo si

$$\forall u, v \in K, \quad \text{se verifica} \quad \lambda u + (1-\lambda)v \in K \quad \forall \lambda \in [0,1]$$

Definición 2.4 Función convexa y estrictamente convexa.

Sea $K \subset \mathbb{R}^d$ un conjunto convexo. Una función $J : K \subset \mathbb{R}^d \longrightarrow \mathbb{R}$ se dice que es convexa en K, si para todo $\lambda \in [0,1]$ y para todo $u, v \in K$ se verifica

$$J(\lambda u + (1-\lambda)v) \leq \lambda J(u) + (1-\lambda)J(v)$$

Además se dice que la función es estrictamente convexa si para todo $\lambda \in]0,1[$ y para todo $u, v \in K$ se verifica

$$J(\lambda u + (1-\lambda)v) < \lambda J(u) + (1-\lambda)J(v)$$

Vamos a estudiar ahora algunas propiedades de la funciones convexas y diferenciables.

Teorema 2.6 Sea $J : A \subset \mathbb{R}^d \longrightarrow \mathbb{R}$, donde A es un abierto de $\mathbb{R}^d$, J diferenciable en A. Sea K una parte convexa de A.

1. $J(.)$ es convexa en K si y solo si

$$J(v) \geq J(u) + (\nabla J(u), v - u) \quad \forall u, v \in K$$

2. $J(.)$ es estrictamente convexa en K si y solo si

$$J(v) > J(u) + (\nabla J(u), v - u) \quad \forall u, v \in K$$

Ejercicio: Interpretar geométricamente las anteriores desigualdades.

Demostración:

1. Si $J(.)$ es convexa en K y $u, v \in K$ entonces para $\lambda \in]0, 1[$

$$u + \lambda(v - u) \in K$$

$$J(u + \lambda(v - u)) \leq (1 - \lambda)J(u) + \lambda J(v)$$

$$J(u + \lambda(v - u)) - J(u) \leq \lambda(J(v) - J(u))$$

$$\frac{J(u + \lambda(v - u)) - J(u)}{\lambda} \leq J(v) - J(u)$$

haciendo $\lambda \to 0^+$

$$(\nabla J(u), v - u) \leq J(v) - J(u)$$

Recíprocamente, sea $J(.)$ verificando

$$J(v) \geq J(u) + (\nabla J(u), v - u) \quad \forall v, u \in K$$

Sea $\lambda \in]0, 1[$ y tomemos primero en el lugar de u, $v + \lambda(u - v)$, resulta

$$J(v) \geq J(v + \lambda(u - v)) - \lambda(\nabla J(v + \lambda(u - v)), u - v)$$

Ahora en el lugar de v tomemos u y en el lugar de u tomemos $v + \lambda(u - v)$, resulta

$$J(u) \geq J(v + \lambda(u - v)) + (1 - \lambda)(\nabla J(v + \lambda(u - v)), u - v)$$

Multiplicando la primera por $(1 - \lambda)$ y la segunda por λ y sumando

$$(1 - \lambda)J(v) + \lambda J(u) \geq J(v + \lambda(u - v))$$

es decir

$$J(\lambda u + (1 - \lambda)v) \leq \lambda J(u) + (1 - \lambda)J(v)$$

2. Si $J(.)$ es estrictamente convexa, el razonamiento anterior no es aplicable pues la desigualdad estricta se pierde al pasar al límite. Procedemos entonces de la siguiente manera: Sea $\lambda \in]0,1[$ y $\omega \in]0,1[$ verificando $\omega > \lambda$, resulta

$$u + \lambda(v-u) = u - \frac{\lambda}{\omega}u + \frac{\lambda}{\omega}u + \lambda(v-u) = \frac{\omega-\lambda}{\omega}u + \frac{\lambda}{\omega}(u + \omega(v-u))$$

de donde

$$J(u+\lambda(v-u)) = J(\frac{\omega-\lambda}{\omega}u + \frac{\lambda}{\omega}(u+\omega(v-u)) \leq \frac{\omega-\lambda}{\omega}J(u) + \frac{\lambda}{\omega}J(u+\omega(v-u))$$

$$J(u+\lambda(v-u)) - J(u) \leq \frac{\lambda}{\omega}(J(u+\omega(v-u) - J(u))$$

$$\frac{J((u+\lambda(v-u)) - J(u)}{\lambda} \leq \frac{J(u+\omega(v-u)) - J(u)}{\omega} < \frac{(1-\omega)J(u) + \omega J(v) - J(u)}{\omega} = J(v) - J(u)$$

pasando al límite cuando $\lambda \to 0$, resulta

$$(\nabla J(u), v-u) \leq \frac{J(u+\omega(v-u)) - J(u)}{\omega} < J(v) - J(u)$$

La recíproca es idéntica al caso anterior.

■

Corolario 2.1 Sea $J : \mathbb{R}^d \to \mathbb{R}$ una función convexa y diferenciable en u tal que $\nabla J(u) = 0$. Entonces $J(.)$ admite un mínimo en u.

Demostración:

$J(.)$ convexa y diferenciable implica

$$J(v) \geq J(u) + (\nabla J(u), v-u) \quad \forall v \in \mathbb{R}^d$$

es decir,

$$J(v) \geq J(u) \quad \forall v \in \mathbb{R}^d$$

■

Ejercicios

1. Estudiar los puntos críticos (puntos u donde $\nabla J(u) = 0$) de las funciones

   ▪

$$J(x,y) = e^{1+x^2+y^2}$$

- $$J(x,y,z) = x^2 + y^2 + z^2 + xy$$
- $$J(x,y) = x^4 + y^4$$
- $$J(x,y) = x^2 - 2xy + 2y^2$$

2. Sea $\Omega \subset \mathbb{R}^2$ y $J : \Omega \to \mathbb{R}$ una función diferenciable en Ω y sea $DJ(a) = 0$ donde $a \in \Omega$. Supongamos además que existe $D^2J(a)$ siendo la matriz Hessiana

$$H(a) = \begin{bmatrix} r & s \\ s & t \end{bmatrix}.$$

Demostrar que si $r > 0$ y $rt - s^2 > 0$ entonces $J(.)$ tiene un mínimo local en a

Capítulo 3

Algoritmos de optimización de problemas en dimensión 1

3.1. Convexidad y quasi-convexidad

Empezamos este capítulo con recordando el concepto de función convexa y dando algunas generalizaciones del concepto de función convexa

3.1.1. Definiciones

Sea K un conjunto convexo de un espacio vectorial E y una función

1. Función convexa: $f : K \longrightarrow \mathbb{R}$ es convexa si

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y) \quad \forall x, y \in K \quad \forall \lambda \in (0,1)$$

2. Función estrictamente convexa: $f : K \longrightarrow \mathbb{R}$ es estrictamente convexa si

$$f((1-\lambda)x + \lambda y) < (1-\lambda)f(x) + \lambda f(y) \quad \forall x, y \in K \text{ con } x \neq y \quad \forall \lambda \in (0,1)$$

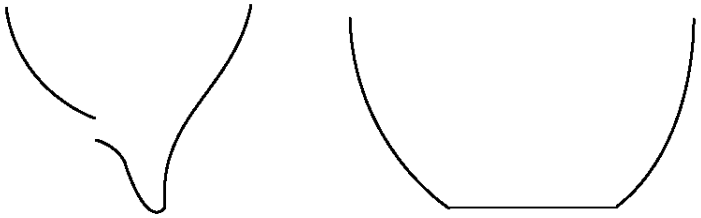

Figura 3.1: Funciones estrictamente quasi-convexas

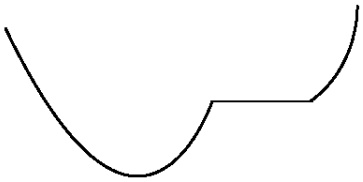

Figura 3.2: Función no estrictamente quasi-convexas

3. Función quasi-convexa: $f : K \longrightarrow \mathbb{R}$ es quasi-convexa si

$$f((1-\lambda)x + \lambda y) \leq max\{f(x), f(y)\} \quad \forall x, y \in K \quad \forall \lambda \in (0,1)$$

4. Función fuertemente quasi-convexa: $f : K \longrightarrow \mathbb{R}$ es fuertemente quasi-convexa si

$$f((1-\lambda)x + \lambda y) < max\{f(x), f(y)\} \quad \forall x, y \in K \; x \neq y \quad \forall \lambda \in (0,1)$$

5. Función estrictamente quasi-convexa: $f : K \longrightarrow \mathbb{R}$ es estrictamente quasi-convexa si $\forall x, y \in K$ con $f(x) \neq f(y)$

$$f((1-\lambda)x + \lambda y) < max\{f(x), f(y)\} \quad \forall x, y \in K \quad \forall \lambda \in (0,1)$$

3.1.2. Ejercicios

1. Demostrar que toda función convexa es quasi-convexa.
2. Demostrar que toda función convexa es estrictamente quasi-convexa.
3. Demostrar que toda función estrictamente convexa es fuertemente quasi-convexa.
4. Demostrar que toda función fuertemente quasi-convexa es estrictamente quasi-convexa.
5. Demostrar que toda función fuertemente quasi-convexa es quasi-convexa.

6. Dar un ejemplo de función estrictamente quasiconvexa que no sea quasi-convexa.

7. Demostrar que una función estrictamente quasi-convexa semicontinua inferior es quasi-convexa.

8. Demostrar que si una función fuertemente quasi-convexa tiene un mínimo relativo en $\bar{x} \in K$ es un mínimo global y además es único.

9. Demostrar que si una función estrictamente quasi-convexa tiene un mínimo relativo en $\bar{x} \in K$ es un mínimo global.

10. Demostrar que f es quasi-convexa si y solo si el conjunto

$$S_\alpha = \{x \in K;\ f(x) \leq \alpha\}$$

es convexo cualquiera que sea $\alpha \in \mathbb{R}$

11. Demostrar que f es convexa si y solo si el conjunto

$$Epif = \{(x, \lambda) \in K \times \mathbb{R};\ f(x) \leq \lambda\}$$

llamado epígrafe de f es convexo.

3.2. Algoritmos de optimización de funciones de una sola variable real

En esta sección consideramos el problema siguiente: Dada un intervalo de la recta real $[a, b]$ y una función

$$f : [a, b] \to \mathbb{R}$$

hallar $\bar{x} \in [a, b]$ tal que

$$f(\bar{x}) = \inf_{x \in [a,b]} f(x) \tag{3.1}$$

Si f es diferenciable la búsqueda de mínimos en el interior de $[a, b]$ se puede realizar buscando los mínimos locales en el interior del intervalo, resolviendo la ecuación $f'(x) = 0$ utilizando por ejemplo el método de la bisección, punto fijo o Newton (si es dos veces diferenciable). Si no hay mínimos locales en el interior entonces el mínimo se alcanzará en uno de los extremos del intervalo.

En este capítulo estudiaremos métodos más generales que no necesitan la diferenciabilidad de la función f. Consideremos el problema (3.1) donde f es una función estrictamente quasi-convexa en el intervalo $[a, b]$ de modo que si f tiene un mínimo relativo en $[a, b]$ este mínimo es global y único. El intervalo $[a, b]$ donde se localiza un mínimo le llamaremos intervalo de incertidumbre. En el proceso de búsqueda del mínimo excluiremos partes del intervalo que no contienen este mínimo, entonces el intervalo de incertidumbre se reduce. En general, $[a, b]$ se llama intervalo de incertidumbre si el mínimo $\bar{x}$ está en $[a, b]$ aunque el valor exacto de este mínimo no se conozca.

Empezaremos con una propiedad de las funciones estrictamente quasi-convexas que juega un papel importante en los algoritmos que estudiaremos a continuación.

Teorema 3.1 Sea $f : [a, b] \to \mathbb{R}$ estrictamente quasiconvexa en $[a, b]$. Sean λ, $\mu \in [a, b]$ tales que $\lambda < \mu$. Si $f(\lambda) > f(\mu)$ entonces $f(z) \geq f(\mu) \quad \forall z \in (a, \lambda]$ y el mínimo está en el intervalo $[\lambda, b]$.

Análogamente si $f(\lambda) < f(\mu)$ entonces $f(z) \geq f(\lambda) \quad \forall z \in [\mu, b)$ y el mínimo está en el intervalo $[a, \mu]$.

Demostración: Supongamos $\lambda < \mu$ y $f(\lambda) > f(\mu)$. Sea $z \in [a, \lambda)$.

Supongamos por reducción al absurdo que $f(z) < f(\mu)$.

Como λ se puede escribir como combinación lineal convexa de z y de μ, resulta

$$f(\lambda) < \text{máx}\{f(z),\, f(\mu)\} = f(\mu)$$

en contra de la hipótesis $f(z) < f(\mu)$.

Por tanto el mínimo estará en el intervalo $[\lambda, b]$ que será el nuevo intervalo de incertidumbre.

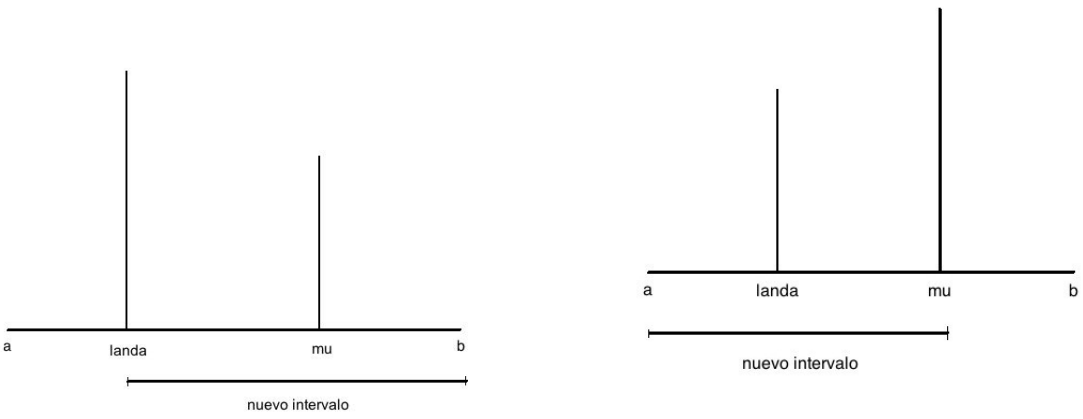

Figura 3.3: Elección del nuevo intervalo

3.2.1. Método de la Dicotomía

Sea $[a_0, b_0] = [a, b]$, $\varepsilon > 0$ y l longitud del intervalo final.

Algortimo de la Dicotomía: Descripción del paso general

Sea $[a_k, b_k]$,

- Si $b_k - a_k < l$ STOP

- En caso contrario

$$\lambda_k = \frac{a_k + b_k}{2} - \varepsilon$$
$$\mu_k = \frac{a_k + b_k}{2} + \varepsilon$$

  - Si $f(\lambda_k) < f(\mu_k)$ el nuevo intervalo es

$$a_{k+1} = a_k$$
$$b_{k+1} = \mu_k$$

  - Si $f(\lambda_k) > f(\mu_k)$ el nuevo intervalo es

$$a_{k+1} = \lambda_k$$
$$b_{k+1} = b_k$$

  - Si $f(\lambda_k) = f(\mu_k)$ STOP

3.2.2. Ejercicios

1. Demostrar que en el método de la Dicotomía la longitud del intervalo de incertidumbre en el paso k-ésimo es

$$b_k - a_k = \frac{1}{2^k}(b-a) + 2\varepsilon(1 - \frac{1}{2^k}) \tag{3.2}$$

3.2.3. Método de la Sección Áurea

En el método de la dicotomía se requiere realizar dos evaluaciones de la función en cada iteración. Vamos a ver dos métodos que solo requieren realizar una evaluación de la función en cada iteración.

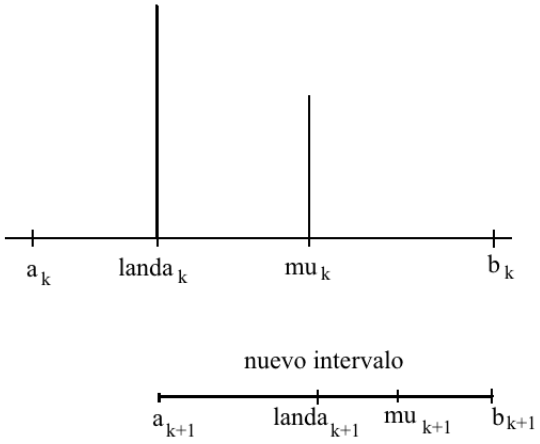

Figura 3.4: Elección del nuevo intervalo. Caso 1

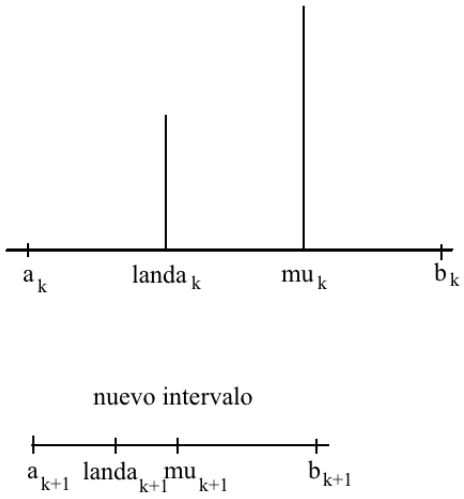

Figura 3.5: Elección del nuevo intervalo. Caso 2

Seleccionaremos λ_k y μ_k de modo que se cumplan dos criterios.

1. Primer criterio: Se seleccionan λ_k y μ_k de modo que la longitud del nuevo intervalo $[a_{k+1}, b_{k+1}]$ no

dependa del resultado de la iteración k, es decir, tanto si estamos en el caso 1 figura (3.4)

$$f(\lambda_k) > f(\mu_k) \quad \Rightarrow \text{nuevo intervalo} \quad [\lambda_k, b_k]$$

o en el caso 2, figura (3.5)

$$f(\lambda_k) < f(\mu_k) \quad \Rightarrow \text{nuevo intervalo} \quad [a_k, \mu_k]$$

tengamos

$$b_k - \lambda_k = \mu_k - a_k$$

Poniendo $\lambda_k = a_k + (1-\alpha)(b_k - a_k)$, resulta

$$\begin{aligned} \mu_k &= a_k + b_k - \lambda_k = a_k + b_k - a_k - (1-\alpha)(b_k - a_k) \\ &= b_k - b_k + a_k + \alpha b_k - \alpha a_k \\ &= a_k + \alpha(b_k - a_k) \end{aligned}$$

de donde en el caso 1

$$b_{k+1} - a_{k+1} = b_k - \lambda_k = b_k - a_k - (1-\alpha)(b_k - a_k) = \alpha(b_k - a_k)$$

y análogamente en el caso 2.

2. Segundo criterio: λ_{k+1} y μ_{k+1} se seleccionan con el propósito que la nueva iteración bien λ_{k+1} coincida con μ_k (caso 1) o μ_{k+1} coincida con λ_k (caso 2). Esta condición nos determina el valor de α, en efecto, consideremos por ejemplo el caso 1. Si $f(\lambda_k) > f(\mu_k)$ el nuevo intervalo es

$$\begin{aligned} a_{k+1} &= \lambda_k \\ b_{k+1} &= b_k \end{aligned}$$

y queremos que $\lambda_{k+1} = \mu_k$, tendremos

$$\begin{aligned} \mu_k &= \lambda_{k+1} = a_{k+1} + (1-\alpha)(b_{k+1} - a_{k+1}) \\ &= \lambda_k + (1-\alpha)(b_k - \lambda_k) \end{aligned}$$

sustituyendo la expresión de μ_k y de λ_k en la anterior resulta

$$\begin{aligned} a_k + \alpha(b_k - a_k) &= a_k + (1-\alpha)(b_k - a_k) + (1-\alpha)\big(b_k - a_k - (1-\alpha)(b_k - a_k)\big) \\ &= b_k - \alpha^2 b_k + \alpha^2 a_k \end{aligned}$$

de donde

$$(1 - \alpha - \alpha^2)a_k = (1 - \alpha - \alpha^2)b_k$$

y α es la solución positiva de

$$\alpha^2 + \alpha - 1 = 0$$

es decir

$$\alpha = \frac{-1 + \sqrt{5}}{2}$$

.

Algoritmo de la Sección Áurea

- Lee los datos del problema: Extremos del intervalo inicial, precisión requerida, número máximo de iteraciones:

  $$a,\ b,\ \delta,\ imax$$

  Definicion de la función $f(.)$

- Inicializaciones

  $$\alpha = \frac{-1+\sqrt{5}}{2}$$

  $$e = b - a$$

  Si $e < \delta$ STOP

- Paso Inicial

  $$\begin{aligned} \lambda &= a + (1-\alpha) * e \\ \mu &= a + \alpha * e \\ fln &= f(\lambda) \\ fmu &= f(\mu) \end{aligned}$$

- Para $i = 1$ hasta $i = imax$
  - Si $fln > fmu$

    $$\begin{aligned} a &= \lambda \\ \lambda &= \mu \\ \mu &= a + \alpha(b-a) \\ fln &= fmu \\ fmu &= f(\mu) \end{aligned}$$

  - Si $fln > fmu$

    $$\begin{aligned} b &= \mu \\ \mu &= \lambda \\ \lambda &= a + (1-\alpha)(b-a) \\ fmu &= fln \\ fln &= f(\lambda) \end{aligned}$$

  - Si $fln == fmu$ STOP
  - $e = b - a$
  - Si $e \leq \delta$ STOP

3.2.4. Método de Fibonacci

En el método de Fibonacci seguiremos cumpliendo el segundo criterio del la sección anterior, aunque el valor de α que determina el valor de los puntos λ_k y μ_k en el intervalo $[a_k, b_k]$ va a depender de la iteración k, es decir, dado un intervalo de incertidumbre $[a_k, b_k]$ calculamos dos puntos intermedios, λ_k y μ_k eligiendo $0 < \alpha_k < 1$ y

$$\lambda_k = a_k + (1 - \alpha_k)(b_k - a_k)$$

$$\mu_k = a_k + \alpha_k(b_k - a_k)$$

donde $\alpha_k = \frac{F_{N-k}}{F_{N-k+1}}$ siendo $(F_n)_{n=1,2,\ldots,N}$ los N+1 primeros términos de la sucesión de Fibonacci definida por

- $$F_0 = F_1 = 1$$
- $$F_{n+1} = F_n + F_{n-1}$$

podemos escribir,

$$\lambda_k = a_k + \frac{F_{N-k-1}}{F_{N-k+1}}(b_k - a_k)$$

$$\mu_k = a_k + \frac{F_{N-k}}{F_{N-k+1}}(b_k - a_k)$$

en efecto,

$$\frac{F_{N-k-1}}{F_{N-k+1}} + \frac{F_{N-k}}{F_{N-k+1}} = 1$$

la reducción de la longitud del intervalo de incertidumbre será: En el caso $f(\lambda_k) > f(\mu_k)$, el nuevo intervalo es $[\lambda_k, b_k]$ cuya longitud es,

$$\begin{aligned} b_{k+1} - a_{k+1} &= b_k - \lambda_k \\ &= b_k - a_k - \frac{F_{N-k-1}}{F_{N-k+1}}(b_k - a_k) \\ &= \frac{F_{N-k}}{F_{N-k+1}}(b_k - a_k) \end{aligned}$$

y análogamente en el caso $f(\lambda_k) \leq f(\mu_k)$, el nuevo intervalo es $[a_k, \mu_k]$ cuya longitud es,

$$\begin{aligned} b_{k+1} - a_{k+1} &= \mu_k - a_k \\ &= a_k + \frac{F_{N-k}}{F_{N-k+1}}(b_k - a_k) - a_k \\ &= \frac{F_{N-k}}{F_{N-k+1}}(b_k - a_k) \end{aligned}$$

Después de $N-1$ iteraciones la longitud el intervalo de incertidumbre será

$$\frac{F_{N-1}}{F_N}\frac{F_{N-2}}{F_{N-1}}\cdots\frac{F_{N-(N-1)}}{F_{N-(N-2)}}=\frac{1}{F_N}$$

Algoritmo de Fibonacci

- Lee los datos del problema: Extremos del intervalo inicial, precisión requerida

  $$a,\ b,\ \delta$$

  Si $b-a<\delta$ STOP Definicion de la función $f(.)$
- Calcula el número N tal que F_N el menor número de Fibonacci verificando

  $$F_N \geq \frac{b-a}{\delta}$$
- Paso inicial

  $$\alpha=\frac{F_{N-1}}{F_N}$$

  $$\begin{aligned}\lambda &= a+(1-\alpha)*(b-a)\\ \mu &= a+\alpha*(b-a)\\ fln &= f(\lambda)\\ fmu &= f(\mu)\end{aligned}$$
- Para $n=1$ hasta $n=N-1$
  - Si $fln>fmu$

    $$\begin{aligned}a &= \lambda\\ \lambda &= \mu\\ \mu &= a+\frac{F_{N-n-1}}{F_{N-n}}(b-a)\\ fln &= fmu\\ fmu &= f(\mu)\end{aligned}$$
  - Si $fln<fmu$

    $$\begin{aligned}b &= \mu\\ \mu &= \lambda\\ \lambda &= a+\frac{F_{N-n-2}}{F_{N-n}}(b-a)\\ fmu &= fln\\ fln &= f(\lambda)\end{aligned}$$

- Si $fln == fmu$ STOP
- $e = b - a$
- Si $e \leq \delta$ STOP

Capítulo 4

Algoritmos para problemas sin restricciones

4.1. Métodos de gradiente

4.1.1. Método de gradiente para la minimización sin restricciones

Sea $J : \mathbb{R}^d \longrightarrow \mathbb{R}$ diferenciable. Consideraremos el problema siguiente: Hallar $u \in \mathbb{R}^d$ tal que

$$J(u) = \inf_{v \in \mathbb{R}^d} J(v)$$

El método de gradiente es

1. $u^0 \in \mathbb{R}^d$, arbitrario
2. Conocido u^n se calcula u^{n+1} de la siguiente manera

$$u^{n+1} = u^n - \rho_n \nabla J(u^n)$$

Nota 4.1 $d^n = -\nabla u^n$ es la dirección de máximo descenso local. En efecto,

$$J(u^n + \rho_n d^n) = J(u^n) + \rho_n(\nabla J(u^n), d^n) + ...$$

$$J(u^n + \rho_n d^n) - J(u^n) \approx \rho_n(\nabla J(u^n), d^n)$$

El término $(\nabla J(u^n), d^n)$ para $||d^n|| = 1$ toma el valor máximo si $d^n = \frac{\nabla J(u^n)}{||\nabla J(u^n)||}$

Analizaremos la convergencia del anterior método en el caso de funciones elípticas, es decir,

Definición 4.1 Función elíptica

Una función $J : \mathbb{R}^d \longrightarrow \mathbb{R}$, es elíptica si es diferenciable con continuidad y existe $\alpha > 0$ tal que

$$(\nabla J(v) - \nabla J(u), v - u) \geq \alpha ||v - u||^2, \quad \forall u, v \in \mathbb{R}^d$$

Teorema 4.1 Sea $J : \mathbb{R}^d \longrightarrow \mathbb{R}$ una función derivable en $\mathbb{R}^d$ tal que

$$\exists \alpha > 0 \quad (\nabla J(v) - \nabla J(u), v - u) \geq \alpha ||v - u||^2, \quad \forall u, v \in \mathbb{R}^d$$

$$\exists \beta \quad ||\nabla J(v) - \nabla J(u)|| \leq \beta ||v - u|| \quad \forall u, v \in \mathbb{R}^d$$

Entonces el problema de optimización: Hallar $u \in \mathbb{R}^d$ tal que

$$J(u) = \inf_{v \in \mathbb{R}^d} J(v)$$

tiene solución única y existen dos números a y b tales que para todo $n \geq 0$ verifican

$$0 < a \leq \rho_n \leq b < \frac{2\alpha}{\beta^2}$$

de modo que el método de gradiente antes descrito es convergente.

Para realizar la demostración necesitamos conocer algunas propiedades de las funciones elípticas.

Lema 4.1 Una función $J : \mathbb{R}^d \to \mathbb{R}$ elíptica verifica la siguiente desigualdad,

$$J(v) - J(u) \geq (\nabla J(u), v - u) + \frac{\alpha}{2} ||v - u||^2 \quad \forall u, v \in \mathbb{R}^d$$

y por tanto es estrictamente convexa.

Demostración:

$$J(v) - J(u) = \int_0^1 (\nabla J(u + t(v - u)), v - u) dt$$

en efecto, la anterior igualdad no es más que

$$f(1) - f(0) = \int_0^1 f'(t) dt$$

para la función $f : \mathbb{R} \to \mathbb{R}$ definida por

$$f(t) = J(u + t(v - u)) \quad u, v \in \mathbb{R}^d$$

resulta pues,

$$\begin{aligned} J(v) - J(u) &= (\nabla J(u), v-u) + \int_0^1 (\nabla J(u + t(v-u) - \nabla J(u), (v-u))dt \\ &= (\nabla J(u), v-u) + \int_0^1 \frac{1}{t}(\nabla J(u + t(v-u) - \nabla J(u), t(v-u))dt \\ &\geq (\nabla J(u), v-u) + \int_0^1 \alpha t||v-u||^2 dt = (\nabla J(u), v-u) + \frac{\alpha}{2}||v-u||^2 \end{aligned}$$

■

Teorema 4.2 Si $J(.)$ es una función elíptica el problema de encontrar $u \in \mathbb{R}^d$ tal que

$$J(u) = \inf_{v \in \mathbb{R}^d} J(v)$$

tiene solución única y se verifica

$$\nabla J(u) = 0$$

Demostración:

Si $J(.)$ es elíptica verifica

$$\lim_{||v|| \to \infty} J(v) = \infty$$

en efecto,

$$J(v) - J(u) \geq (\nabla J(u), v-u) + \frac{\alpha}{2}||v-u||^2 \quad \forall u, v \in \mathbb{R}^d$$

en particular, tomando $u = 0$

$$\begin{aligned} J(v) &\geq J(0) + (\nabla J(0), v) + \frac{\alpha}{2}||v||^2 \\ &\geq J(0) - ||\nabla J(0)||.||v|| + \frac{\alpha}{2}||v||^2 \end{aligned}$$

La variante del teorema de Weierstrass nos da la existencia de solución. La unicidad se obtiene utilizando la convexidad estricta de $J(.)$. En efecto, supongamos que u_1 y u_2 son dos soluciones. Por la convexidad estricta

$$J(\frac{u_1 + u_2}{2}) < \frac{1}{2}J(u_1) + \frac{1}{2}J(u_2)$$

Si $\gamma = J(u_1) = J(u_2) = \text{ínf}(J(v)$, entonces $J(\frac{u_1+u_2}{2}) < \gamma$ y u_1 y u_2 no podrían ser soluciones.

Finalmente, la solución única u, es también un mínimo relativo, por tanto verifica $\nabla J(u) = 0$.■

Estamos en condiciones de demostrar el teorema de convergencia del algoritmo de gradiente.

Demostración del teorema de convergencia: El algoritmo de gradiente es

1. $u^0 \in \mathbb{R}^d$, arbitrario
2. Conocido u^n se calcula u^{n+1} de la siguiente manera

$$u^{n+1} = u^n - \rho_n \nabla J(u^n)$$

por otra parte, si u es solución, tenemos $\nabla J(u) = 0$ es decir,

$$u = u - \rho_n \nabla J(u)$$

de donde

$$\begin{aligned} ||u - u^{n+1}||^2 &= ||u - u^n - \rho_n(\nabla J(u) - \nabla J(u^n))||^2 \\ &= ||u - u^n||^2 - 2\rho_n(\nabla J(u) - \nabla J(u^n), u - u^n) + \rho_n^2 ||\nabla J(u) - \nabla J(u^n||^2 \\ &\leq ||u - u^n||^2 - 2\rho_n \alpha ||u - u^n||^2 + \beta^2 \rho_n^2 ||u - u^n||^2 \\ &= (1 - 2\rho_n \alpha + \beta^2 \rho_n^2) ||u - u^n||^2 \end{aligned}$$

siempre podemos elegir ρ_n de modo que

$$\tau(\rho_n) = (1 - 2\rho_n \alpha + \beta^2 \rho_n^2) < 1$$

en efecto, la función

$$\tau(\rho) : \rho \longrightarrow 1 - 2\rho\alpha + \beta^2 \rho^2$$

alcanza su valor mínimo en $\rho_{min} = \frac{\alpha}{\beta^2}$ y tenemos $0 < \tau(\rho_{min}) < 1$, por tanto siempre existen dos números a y b tales que , si elegimos $a < \rho_n < b$, y denotamos mediante $\gamma^2 = \text{máx}\{\tau(a), \tau(b)\} < 1$ resulta

$$||u - u^{n+1}|| \leq \gamma ||u - u^n||$$

de donde

$$||u - u^n|| \leq \gamma^n ||u - u^0|| \overset{n \to \infty}{\longrightarrow} 0$$

La convergencia es al menos lineal, pues

$$\lim_{n \to \infty} \frac{||u - u^{n+1}||}{||u - u^n||} \leq \gamma < 1$$

■

4.1.2. Método del gradiente con paso óptimo

El método de gradiente con paso óptimo nos da un criterio para elegir el valor de ρ_n en cada iteración. El algoritmo es el siguiente:

1. $u^0 \in \mathbb{R}^d$, arbitrario

2. Conocido u^n se calcula u^{n+1} de la siguiente manera: Se calcula $\rho_n = \rho(u^n)$ resolviendo el problema de minimización de una variable real

$$J(u^n - \rho_n \nabla J(u^n)) = \inf_{\rho \in \mathbb{R}} J(u^n - \rho \nabla J(u^n))$$

obtenido ρ_n se calcula u^{n+1} mediante

$$u^{n+1} = u^n - \rho_n \nabla J(u^n)$$

Teorema 4.3 Sea $J : \mathbb{R}^d \longrightarrow \mathbb{R}$ elíptica y diferenciable con continuidad, entonces el método de gradiente con paso óptimo es convergente.

Demostración:

Supongamos que $\nabla J(u^n) \neq 0$ (en caso contrario el método sería convergente en un número finito de pasos. Consideremos la función

$$f : \rho \in \mathbb{R} \longrightarrow J(u^n - \rho \nabla J(u^n)) \in \mathbb{R}$$

se comprueba sin gran dificultad

- $f(.)$ es estrictamente convexa.
- $f(.)$ verifica $\lim_{\rho \to \infty} f(\rho) = \infty$
- $f'(\rho) = -(\nabla J(u^n - \rho \nabla J(u^n)), \nabla J(u^n))$

por tanto el problema de hallar $\rho_n = \rho(u^n)$ tal que

$$f(\rho_n) = \inf_{\rho \in \mathbb{R}} f(\rho)$$

tiene solución única y está caracterizada por

$$f'(\rho_n) = 0$$

es decir

$$(\nabla J(u^n - \rho_n \nabla J(u^n)), \nabla J(u^n)) = 0$$

o bien

$$(\nabla J(u^{n+1})), \nabla J(u^n)) = 0$$

por tanto las direcciones de descenso consecutivas son ortogonales. A partir de aquí la demostración se realiza en cinco etapas.

1. $\lim_{n \to \infty} (J(u^n) - J(u^{n+1})) = 0$

2. $\lim_{n\to\infty} ||u^n - u^{n+1}|| = 0$
3. $||\nabla J(u^n)|| \leq ||\nabla J(u^n) - \nabla J(u^{n+1})||$
4. $\lim_{n\to\infty} ||\nabla J(u^n|| = 0$
5. $\lim_{n\to\infty} ||u^n - u|| = 0$

Demostración:

1. La sucesión $(J(u^n))_{n\geq}$ es decreciente por construcción y acotada inferiormente por $J(u)$. Por lo tanto es convergente, en consecuencia,

$$\lim_{n\to\infty} (J(u^n) - J(u^{n-1})) = 0$$

2. Tenemos, según se ha visto en el teorema anterior $(\nabla J(u^{n+1}), \nabla J(u^n)) = 0$. Como $u^{n+1} = u^n - \rho(u^n)\nabla J(u^n)$, resulta

$$(\nabla J(u^{n+1}), u^n - u^{n+1}) = 0$$

Por otra parte, la elipticidad de $J(.)$ implica

$$J(u^n) \geq J(u^{n+1}) + (\nabla J(u^{n+1}), u^n - u^{n+1}) + \frac{\alpha}{2}||u^n - u^{n+1}||^2$$

de donde

$$J(u^n) - J(u^{n+1}) \geq \frac{\alpha}{2}||u^n - u^{n+1}||^2$$

de la parte 1 deducimos

$$\lim_{n\to\infty} ||u^n - u^{n+1}|| = 0$$

3. Por la ortogonalidad de $\nabla J(u^n)$ y de $\nabla J(u^{n+1})$ resulta

$$\begin{aligned} ||\nabla J(u^n)||^2 &= (\nabla J(u^n), \nabla J(u^n) - \nabla J(u^{n+1})) \\ &\leq ||\nabla J(u^n)||.||\nabla J(u^n) - \nabla J(u^{n+1})|| \end{aligned}$$

de donde se obtiene 3) dividiendo ambos miembros de la desigualdad por $||\nabla J(u^n)||$

4. La sucesión $(J(u^n))_{n\geq 0}$ es acotada. Como $J(.)$ verifica $\lim_{||v||\to\infty} J(v) = \infty$, la sucesión $(u^n)_n$ está acotada y se puede incluir en un conjunto compacto. Por otra parte $J(.)$ es diferenciable con continuidad, es decir, la aplicación

$$DJ(.) : v \in \mathbb{R}^d \longrightarrow DJ(v) \in \mathcal{L}(\mathbb{R}^d; \mathbb{R})$$

es continua. O dicho de otra forma la aplicación

$$\nabla J(.) : v \in \mathbb{R}^d \longrightarrow \nabla J(v) \in \mathbb{R}^d$$

es continua. Por tanto sobre los conjuntos compactos es uniformemente continua. En consecuencia

$$\begin{aligned}\lim_{n\to\infty}||\nabla J(u^n)|| &\leq \lim_{n\to\infty}||\nabla J(u^n)-\nabla J(u^{n+1})||\\ &= \lim_{||u^n-u^{n+1}||\to 0}||\nabla J(u^n)-\nabla J(u^{n+1})|| = 0\end{aligned}$$

5.

$$\begin{aligned}\alpha||u^n-u||^2 &\leq (\nabla J(u^n)-\nabla J(u),u^n-u)\\ &= (\nabla J(u^n),u^n-u)\\ &\leq ||\nabla J(u^n||.||u^n-u||\end{aligned}$$

de donde finalmente

$$||u^n-u||\leq\frac{1}{\alpha}||\nabla J(u^n)||$$

y finalmente

$$\lim_{n\to\infty}||u^n-u|| = 0$$

■

Ejercicios

1. Sea $J:\mathbb{R}^d\longrightarrow\mathbb{R}$ dos veces diferenciable en $\mathbb{R}^d$. Demostrar:

   a) $J(.)$ es convexa si y solo si la diferencial segunda en todo punto, $D^2J(u)$, es semidefinida positiva, es decir,

   $$D^2J(u)(v,v)\geq 0\quad\forall v\in\mathbb{R}^d,\forall u\in\mathbb{R}^d$$

   b) Si la diferencial segunda en todo punto $u\in\mathbb{R}^d$ es definida positiva, es decir,

   $$D^2J(u)(v,v)>0\quad\forall v\in\mathbb{R}^d,\forall u\in\mathbb{R}^d$$

   entonces, $J(.)$ es estrictamente convexa.

   c) Encontrar un ejemplo sencillo que permita asegurar que el recíproco de b) no es cierto.

2. Considerar la función

   $$\begin{aligned}J:\quad \mathbb{R}^2 &\longrightarrow\mathbb{R}\\ (x,y) &\longrightarrow (x-2)^4+(x-2y)^2\end{aligned}$$

   y considerar el problema

   Hallar $u=(x,y)$ tal que

   $$J(u)=\inf_{v\in\mathbb{R}^2}J(v)$$

   a) Demostrar que el problema tiene al menos una solución.

   b) Hallar una solución y deducir que esta solución es única.

   c) Aplicar el método del gradiente con paso óptimo para aproximar la solución anterior partiendo de $(0.,3.)$

4.2. Método de relajación

Vamos a describir el algoritmo de relajación para resolver el problema:

Hallar $u \in \mathbb{R}^d$ tal que

$$J(u) = \inf_{v \in \mathbb{R}^d} J(v)$$

Descripción del algoritmo

1. Sea $u^0 \in \mathbb{R}^d$ arbitrario.
2. Obtenido $u^n = (u_1^n, ..., u_d^n) \in \mathbb{R}^d$ calculamos $u^{n+1} = (u_1^{n+1}, ..., u_d^{n+1}) \in \mathbb{R}^d$ en d etapas que son: Para $i = 1, 2, ..., d$ calculamos sucesivamente

$$u^{n+\frac{i}{d}} = (u_1^{n+1}, u_2^{n+1}, ..., u_i^{n+1}, u_{i+1}^n, ..., u_d^n) \in \mathbb{R}^d$$

solución de

$$J(u^{n+\frac{i}{d}}) = \inf_{v=(u_1^{n+1}, u_2^{n+1}, ..., v_i, u_{i+1}^n, ..., u_d^n)} J(v)$$

Observación: En cada etapa, se trata de un problema de minimización en una sola variable real. En cada etapa la condición necesaria de mínimo (y suficiente si $J(.)$ es convexa) es:

$$\frac{\partial J}{\partial x_i}(u_1^{n+1}, ..., u_{i-1}^{n+1}, u_i^{n+1}, u_{i+1}^n, ..., u_d^n) = 0$$

Estudiemos ahora la convergencia del método.

Teorema 4.4 Si $J : \mathbb{R}^d \longrightarrow \mathbb{R}$ es elíptica y diferenciable con continuidad, el método de relajación es convergente.

Demostración: Primero observemos que el algoritmo de relajación está bien definido. En efecto, en cada etapa, $i = 1, ..., d$ se resuelve un problema de optimización, para una función J_i de una sola variable que es elíptica y por tanto tiene solución única.

$$J_i(\xi) = J(u_1^{n+1}, ..., u_{i-1}^{n+1}, \xi, u_{i+1}^n, ..., u_d^n)$$

y la solución de

$$J_i(u_i^{n+1}) = \inf_{\xi \in \mathbb{R}} J_i(\xi)$$

está caracterizada por

$$\frac{\partial J}{\partial x_i}(u^{n+\frac{i}{d}}) = 0$$

La demostración ahora se hace en varias etapas:

1.

$$\lim_{n\to\infty}(J(u^n)-J(u^{n+1}))=0$$

En efecto, por la propia construcción del algoritmo tenemos

$$J(u)\leq J(u^{n+1})\leq J(u^n)\leq J(u^0)$$

es decir $(J(u^n))_n$ es una sucesión decreciente de números reales acotada inferiormente, por tanto convergente. En particular $\lim_{n\to\infty}(J(u^n)-J(u^{n+1}))=0$

2.

$$\lim_{n\to\infty}||u^n-u^{n+1}||=0$$

y en particular

$$\lim_{n\to\infty}||u^{n+\frac{i}{d}}-u^n||=0$$

La elipticidad implica

$$J(u^{n+\frac{i-1}{d}})-J(u^{n+\frac{i}{d}})\geq(\nabla J(u^{n+\frac{i}{d}}),u^{n+\frac{i-1}{d}}-u^{n+\frac{i}{d}})+\frac{\alpha}{2}||u^{n+\frac{i-1}{d}}-u^{n+\frac{i}{d}}||^2$$

como

$$\begin{aligned}(\nabla J(u^{n+\frac{i}{d}}),u^{n+\frac{i-1}{d}}-u^{n+\frac{i}{d}}) &= \sum_{j=1}\frac{\partial J}{\partial x_j}(u^{n+\frac{i}{d}})(u_j^{n+\frac{i-1}{d}}-u_j^{n+\frac{i}{d}})\\ &= \sum_{j\neq i}\frac{\partial J}{\partial x_j}(u^{n+\frac{i}{d}})(u_j^{n+\frac{i-1}{d}}-u_j^{n+\frac{i}{d}})+\frac{\partial J}{\partial x_i}(u^{n+\frac{i}{d}})(u_i^{n+\frac{i-1}{d}}-u_i^{n+\frac{i}{d}})=0\end{aligned}$$

pues para $j\neq i$ se tiene $u_j^{n+\frac{i-1}{d}}-u_j^{n+\frac{i}{d}}=0$ y por otra parte $\frac{\partial J}{\partial x_i}(u^{n+\frac{i}{d}})=0$ por la caracterización de la solución del problema de minimización en dimensión uno correspondiente. De modo que,

$$J(u^{n+\frac{i-1}{d}})-J(u^{n+\frac{i}{d}})\geq\frac{\alpha}{2}||u^{n+\frac{i-1}{d}}-u^{n+\frac{i}{d}}||^2=\frac{\alpha}{2}|u_i^n-u_i^{n+1}|^2$$

sumando para $i=1,...,d$

$$\sum_{i=1}^{d}\left(J(u^{n+\frac{i-1}{d}})-J(u^{n+\frac{i}{d}})\right)\geq\frac{\alpha}{2}\sum_{i=1}^{d}|u_i^n-u_i^{n+1}|^2$$

de donde

$$J(u^n)-J(u^{n+1})\geq\frac{\alpha}{2}||u^n-u^{n+1}||^2$$

Y finalmente aplicando el resultado de la parte a obtenemos el resultado.

3.

$$\lim_{n\to\infty}|\frac{\partial J}{\partial x_j}(u^{n+\frac{i}{d}})-\frac{\partial J}{\partial x_j}(u^n)|=0$$

en efecto, como cada sucesión $(J(u^{n+\frac{i}{d}}))_n$ es decreciente por construcción, $(u^{n+\frac{i}{d}})_n$ es acotada pues $J(.)$ verifica la propiedad $\lim_{||v||\to\infty} J(v) = \infty$ y por otra parte cada derivada parcial $\frac{\partial J}{\partial x_j}(.)$ es continua, y por lo tanto es uniformemente continua en los compactos, resulta

$$\lim_{n\to\infty} ||u^{n+\frac{i}{d}} - u^n|| = 0 \Rightarrow \lim_{n\to\infty} |\frac{\partial J}{\partial x_j}(u^{n+\frac{i}{d}}) - \frac{\partial J}{\partial x_j}(u^n)| = 0$$

4.

$$\lim_{n\to\infty} ||\nabla J(u^n)|| = 0$$

En efecto, tomando $j = i$ en el resultado del paso anterior tenemos

$$\begin{aligned} \lim_{n\to\infty} ||\nabla J(u^n)||^2 &= \lim_{n\to\infty} \sum_{i=1}^{d} (\frac{\partial J(u^n)}{\partial x_i})^2 \\ &= \lim_{n\to\infty} \sum_{i=1}^{d} (\frac{\partial J(u^n)}{\partial x_i} - \frac{\partial J(u^{n+\frac{i}{d}})}{\partial x_i})^2 = 0 \end{aligned}$$

5.

$$\lim_{n\to\infty} ||u^n - u|| = 0$$

en efecto, tenemos

$$\begin{aligned} \alpha||u^n - u||^2 &\leq (\nabla J(u^n) - \nabla J(u), u^n - u) \\ &= (\nabla J(u^n), u^n - u) \leq ||\nabla J(u^n)||.||u^n - u|| \end{aligned}$$

y finalmente

$$||u^n - u|| \leq \frac{1}{\alpha}||\nabla J(u^n)||$$

de donde tomando límites cuando $n \to \infty$

$$\lim_{n\to\infty} ||u^n - u|| \leq \frac{1}{\alpha} \lim_{n\to\infty} ||\nabla J(u^n)|| = 0$$

■

4.3. Métodos de Newton

Consideraremos en esta sección métodos para resolver ecuaciones de la forma siguiente: Sea $\Omega \in \mathbb{R}^d$ un conjunto abierto. Dada $F : \Omega \longrightarrow \mathbb{R}^d$ queremos hallar $u \in \Omega$ tal que $F(u) = 0$. En particular el método será aplicable a problemas de optimización de una función $J(.)$ diferenciable. Según hemos visto anteriormente si $J(.)$ tiene un extremo realtivo en $u \in \Omega$ entonces $\nabla J(u) = 0$. De hecho los métodos

de gradiente estudiados en las secciones anteriores son métodos de punto fijo para resolver la ecuación $\nabla J(u) = 0$. Vamos ahora a estudiar el método de Newton que no es más que una generalización a varias variables del método de Newton estudiado en el capítulo 1.

Supongamos que F es diferenciable y sea u^0 un valor en un entorno de la solución u. Pongamos $u = u^0 + h$, tendremos utilizando el desarrollo de Taylor

$$F(u^0 + h) = 0 = F(u^0) + F'(u^0).h + \varepsilon(h^2)$$

para valores de $||h||$ pequeños resulta

$$F'(u^0)h \approx -F(u^0)$$

Si ponemos $u^1 = u^0 + h$ esperamos que u^1 sea una mejor aproximación de u que u^0. De ahí el siguiente algoritmo de Newton:

Algoritmo de Newton

1. $u^0 \in \Omega$ "cercano" a u
2. Obtenido el valor de u^n calculamos u^{n+1} resolviendo

$$F'(u^n)h^n = -F(u^n)$$

$$u^{n+1} = u^n + h^n$$

Vamos a estudiar ahora la convergencia del método de Newton. Será un teorema de convergencia local. Para ello necesitaremos algunos resultados previos.

Lema 4.2 Sea $||.||$ una norma matricial subordinada y E una matriz cuadrada de orden d. Si $||E|| < 1$ entonces existe la matriz inversa de $(I + E)$ y $||(I + E)^{-1}|| \leq \frac{1}{1-||E||}$.

Demostración: Considreremos el sistema $x + Ex = 0$, es decir $Ex = -x$, entonces si $||E|| < 1$,

$$||x|| = ||Ex|| \leq ||E||.||x|| < ||x||$$

La única solución de la ecuación anterior es $x = 0$. Por tanto $I + E$ es no singular.

Por otra parte $I = (I + E) - E$, multiplicando por la derecha por $(I + E)^{-1}$, resulta

$$(I + E)^{-1} = I - E(I + E)^{-1}$$

tomando normas

$$||(I + E)^{-1}|| \leq 1 + ||E||.||(I + E)^{-1}||$$

y finalmente reordenando y agrupando términos

$$||(I + E)^{-1}|| \leq \frac{1}{1 - ||E||}$$

■

Lema 4.3 Sean A y B matrices cuadradas de orden d. Si A es no singular y $||A^{-1}(B-A)|| < 1$ entonces B es no singular y

$$||B^{-1}|| \leq \frac{||A^{-1}||}{1 - ||A^{-1}(B-A)||}$$

Demostración: Pongamos $E = A^{-1}(B-A)$.

$$I + E = I + A^{-1}(B-A) = I + A^{-1}B - I = A^{-1}B$$

Por tanto $A^{-1}B$ es no singular y $(A^{-1}B)^{-1} = B^{-1}A$ y finalmente aplicando el lema anterior

$$||B^{-1}A|| \leq \frac{1}{1 - ||A^{-1}(B-A)||}$$

de donde teniendo en cuenta $||B^{-1}|| = ||B^{-1}AA^{-1}|| \leq ||B^{-1}A||.||A^{-1}||$ resulta

$$||B^{-1}|| \leq \frac{||A^{-1}||}{1 - ||A^{-1}(B-A)||}$$

■

Lema 4.4 Sea $\Omega \in \mathbb{R}^d$ un conjunto abierto y convexo y $F : \Omega \subset \mathbb{R}^d \longrightarrow \mathbb{R}$ diferenciable con continuidad en Ω. Para todo $u \in \Omega$ y $h \in \mathbb{R}^d$ tal que $u + h \in \Omega$ se verifica

$$F(u+h) - F(u) = \int_0^1 F'(u+th).h dt$$

Demostración: consideremos la función $f : \mathbb{R} \to \mathbb{R}$ definida por $t \in [-1, 1] \to f(t) = F(u+th)$. tenemos por la regla de Barrow

$$f(1) - f(0) = \int_0^1 f'(t) dt$$

que es la expresión buscada teniendo en cuenta que

$$f'(t) = F'(u+th).h$$

■

Observación: El lema es válido para $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}^p$ aplicando lo anterior a las p componentes de F.

Lema 4.5 Sea $F : \Omega \subset \mathbb{R}^d \longrightarrow \mathbb{R}^p$ donde $\Omega \in \mathbb{R}^d$ es un conjunto abierto y convexo, una función tal que $DF(.)$ es Lipschitziana, es decir

$$||F'(v) - F'(u)|| \leq \gamma ||v - u|| \quad \forall\, u, v \in \Omega$$

entonces para todo $u+h \in \Omega$

$$||F(u+h) - F(u) - F'(u)h|| \leq \frac{\gamma}{2}||h||^2$$

Demostración: Tenemos

$$F(u+h) - F(u) - F'(u)h = \int_0^1 (F'(u+th) - F'(u))hdt$$

tomando normas y mayorando obtenemos el resultado buscado ■

Teorema 4.5 Sea $\Omega \in \mathbb{R}^d$ un conjunto abierto y convexo y $F : \Omega \subset \mathbb{R}^d \longrightarrow \mathbb{R}^d$ diferenciable con continuidad en Ω. Supongamos que existe un punto $u \in \Omega$ y tres constantes $r > 0$, $\beta > 0$ y $\gamma \geq 0$ tales que

- $F(u) = 0$
- $\mathcal{U}(u,r) = \{v \in \mathbb{R}^d;\ ||v-u|| < r\} \subset \Omega$
- $F'(u)$ es no singular y $||F'(u)^{-1}|| \leq \beta$
- $||F'(v) - F'(u)|| \leq \gamma||v-u|| \quad \forall\ v,u \in \mathcal{U}$

Entonces para $\varepsilon = \text{mín}\{r, \frac{1}{2\beta\gamma}\}$ y para todo $u^0 \in \mathcal{U}(u,\varepsilon) = \{v \in \mathbb{R}^d;\ ||v-u|| < \varepsilon\}$ la sucesión generada por el algoritmo de Newton está bien definida y converge hacia u con convergencia cuadrática, es decir

$$||u^{n+1} - u|| \leq \beta\gamma||u^n - u||^2$$

Demostración: tomemos $\varepsilon = min\{r, \frac{1}{2\beta\gamma}\}$ entonces $\forall u^0 \in \mathcal{U}(u,\varepsilon)$ $F'(u^0)$, es no singular, en efecto

$$\begin{aligned} ||F'(u)^{-1}[F'(u^0) - F'(u)]|| &\leq ||F'(u)^{-1}||.||F'(u^0) - F'(u)|| \\ &\leq \beta\gamma||u^0 - u|| \leq \beta\gamma\varepsilon \leq \frac{1}{2} \end{aligned}$$

Entonces por la relación del lema 4.3, $F'(u^0)$ es no singular y

$$||F'(u^0)^{-1}|| \leq \frac{||F'(u)^{-1}||}{1 - ||F'(u)^{-1}[F'(u^0) - F'(u)]||} \leq 2||F'(u)^{-1}|| \leq 2\beta$$

u^1 está bien definido y

$$\begin{aligned} u^1 - u &= u^0 - u - F'(u^0)^{-1}[F(u^0) - F(u)] \\ &= F'(u^0)^{-1}[F(u) - F(u^0) - F'(u^0)(u - u^0)] \end{aligned}$$

$$\begin{aligned} ||u^1 - u|| &\leq ||F(u^0)^{-1}||.||F(u) - F(u^0) - F'(u^0)(u - u^0)|| \\ &\leq \beta\gamma||u^0 - u||^2 \end{aligned}$$

Por otra parte como $||u^0 - u|| \leq \frac{1}{2\beta\gamma}$, resulta

$$||u^1 - u|| \leq \frac{1}{2}||u^0 - u||$$

lo que prueba $u^1 \in \mathcal{U}(u, \varepsilon)$ y por inducción probamos que $||u^{n+1} - u|| \leq \frac{1}{2}||u^n - u||$, el método es convergente y la convergencia es cuadrática pues,

$$||u^{n+1} - u|| \leq \beta\gamma||u^n - u||^2$$

■

Evaluación del coste del método de Newton

En cada iteración hay que

- Evaluar $F(u^n)$, d evaluaciones funcionales.
- Calcular los términos de $F'(u^n)$, es decir, d^2 ($\frac{d(d+1)}{2}$ si $F(u) = \nabla J(u)$) evaluaciones funcionales.
- Resolver un sistema lineal de d ecuaciones con d incógnitas, es decir, del orden de $\frac{d^3}{3}$ ($\frac{d^3}{6}$ si $F(u) = \nabla J(u)$)) operaciones, si lo resolvemos con un método directo.

Muchas veces no se conoce la expresión analítica de las funciones, por lo que no se conoce la expresión de las derivadas parciales. Se recurre entonces al cálculo numérico de estas derivadas mediante diferencias finitas,

$$[F'(u^n)]_{ij} \approx \frac{F_i(u^n + \lambda_n e_j) - F_i(u^n)}{\lambda_n}$$

Observación: Si se elige λ_n adecuadamente, por ejemplo, $\lambda_n \leq C||F(u^n)||$, la convergencia sigue siendo cuadrática.

Algoritmo de Newton modificado

El gran inconveniente del método de Newton descrito anteriormente es la necesidad de resolver un sistema de ecuaciones en cada iteración, lo que hace que en general sea muy costoso. De ahí la necesidad de modificar el método de Newton para soslayar estas dificultades. Un primer método es el algoritmo de Newton modificado que consiste en realizar una primera iteración de Newton y en las siguientes conservar la matriz de la iteración inicial. El algoritmo de Newton modificado es:

1. $u^0 \in \Omega$ "cercano" a u

2. $A = F'(u_0)$

3. Obtenido el valor de u^n calculamos u^{n+1} resolviendo

$$Ah^n = -F(u^n)$$

$$u^{n+1} = u^n + h^n$$

La matriz A solo se factoriza una vez. El inconveniente de este método es que la convergencia deja de ser cuadrática y pasa a ser lineal.

Ejercicios:

1. Sea $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ definida por

$$F(x, y) = \begin{bmatrix} x + y - 3 \\ x^2 + y^2 - 9 \end{bmatrix}.$$

   Resolver utilizando el método de Newton la ecuación $F(x, y) = 0$ tomando como valor inicial $(x, y)^0 = (1, 5)$.

2. Resolver utilizando el método de Newton el problema siguiente: Hallar $(\bar{x}, \bar{y}) \in \mathbb{R}^2$ tal que

$$J(\bar{x}, \bar{y}) = \inf_{(x,y)\in\mathbb{R}^2} J(x, y)$$

   donde $J(x, y) = (x - 2)^4 + (x - 2y)^2$.

3. - Considerar la función $J : \mathbb{R}^2 \to \mathbb{R}^2$ definida por

$$J(x, y) = log(x^2 + y^2 + 1)$$

     Demostrar que el problema: Hallar $u = (x, y)^t \in \mathbb{R}^2$ tal que

$$J(u) = \inf_{v\in\mathbb{R}^2} J(v)$$

     tiene solución única.
   - Comprobar que la Diferencial Segunda de $J(.)$ es definida positiva en el punto solución.
   - Calcular la solución del problema anterior utilizando el método de relajación y tomando como valor inicial $u^0 = (1, 1)$.

4. Considerar el siguiente problema en $\mathbb{R}^d$: hallar $u \in \mathbb{R}^d$ tal que

$$Au + F(u) = f$$

   donde $f \in \mathbb{R}^d$, A es una matriz de orden n definida positiva, y por tanto verifica

$$\exists\ \alpha > 0\ (Av, v) \geq \alpha||v||^2\ \forall\ v \in \mathbb{R}^d$$

y $F : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ es Lipschitziana, es decir

$$\exists\, M \geq 0 \; ||F(u) - F(v)|| \leq M||u - v|| \; \forall\, u, v \in \mathbb{R}^d$$

Demostrar que si $\frac{M}{\alpha} < 1$ el siguiente método de punto fijo converge

- $u^0 \in \mathbb{R}^d$ arbitrario
- Obtenido u^n calculamos u^{n+1} resolviendo

$$Au^{n+1} = f - F(u^n)$$

4.4. Métodos de Quasi-Newton

Método de Broyden

Los métodos de Quasi-Newton se pueden considerar una generalización del método de la secante. Vamos aquí a considerar el más sencillo de ellos que es el conocido como método de Broyden.

Recordemos el método de la secante para resolver ecuaciones de una sola variable real, $f(x) = 0$. Una iteración del método de la secante es:

$$x^{n+1} = x^n - \frac{f(x^n)}{a_n}$$

donde a_n es una aproximación de la derivada de f en x^n, $f'(x^n)$, de la forma $a_n = \frac{f(x^n) - f(x^{n+1})}{x^n - x^{n-1}}$ que se puede escribir de la forma $a_n(x^n - x^{n-1}) = f(x^n) - f(x^{n-1})$ y a la que llamaremos ecuación de la secante.

Consideremos ahora un sistema de ecuaciones como en el apartado 5.4. y efectuamos una primera iteración del algoritmo de Newton. Es decir, dado

$u^0 \in \Omega$ "cercano" a u, solución de $F(u) = 0$ Obtenemos el valor de u^1 resolviendo

$$F'(u^1)h^1 = -F(u^0)$$

$$u^1 = u^0 + h^1$$

En la siguiente iteración sustituiremos $F'(u^1)$ por una aproximación de la misma $A_1 \approx F'(u^1)$ de manera que esta aproximación verifique la ecuación de la secante, es decir,

$$A_1(u^1 - u^0) = F(u^1) - F(u^0)$$

Esta ecuación no determina la matriz A_1 de forma única pues no nos da información de como opera A_1 sobre las direcciones ortogonales a $u^1 - u^0$. Procederemos en el primer paso de la siguiente manera:

1. $u^0 \in \Omega$ "cercano" a u, $A_0 = F'(u^0)$
2. Calculamos u^1 resolviendo

$$A_0 s = -F(u^0)$$
$$u^1 = u^0 + s$$

Donde $s = u^1 - u^0$ y denotamos $y = F(u^1) - F(u^0)$. En la iteración siguiente determinamos A_1 de modo que cumpla, por una parte, la ecuación de la secante

$$A_1 s = y$$

y por otra parte añadiremos las condiciones

$$A_1 z = A_0 z \quad \forall z \quad \text{ortogonal a } s$$

Estas condiciones determinan de forma única la matriz A_1. En efecto,

$$(A_1 - A_0)z = 0 \ \forall z \text{ ortogonal a } s$$

$A_1 - A_0$ tiene que ser una matriz de rango 1, que podemos buscar de la forma $u.s^t$, de esta manera $u.s^t z = u(s^t z) = 0$. Finalmente como

$$\begin{aligned} A_1 s &= y \\ (A_1 - A_0)s &= y - A_0 s \\ us^t s &= y - A_0 s \\ u &= \frac{y - A_0 s}{s^t s} \end{aligned}$$

es decir

$$\begin{aligned} A_1 &= A_0 + us^t \\ &= A_0 + \frac{(y - A_0 s)s^t}{||s||^2} \end{aligned}$$

Algoritmo de Broyden

1. $u^0 \in \Omega$ "cercano" a u
2. $A_0 = F'(u^0)$
3. Para $n = 0, 1, ...$, obtenido u^n, calculamos u^{n+1} resolviendo
   - $A_n s^n = -F(u^n)$
   - $u^{n+1} = u^n + s^n$
   - $y^n = F(u^{n+1}) - F(u^n)$
   - $A_{n+1} = A_n + \frac{(y^n - A_n s^n)(s^n)^t}{||s^n||^2}$

Algoritmo de Broyden con adaptación de la inversa de la matriz de iteración

Para sistemas de tamaño pequeño puede ser interesante adaptar la inversa de la matriz de la iteración. Para ello se puede utilizar la formula de Sherman Morrison siguiente: Si A es no singular y x e y son vectores tales que $1 + y^t A^{-1} x \neq 0$ entonces $A + xy^t$ es no singular y se tiene

$$(A + xy^t)^{-1} = A^{-1} - \frac{A^{-1} x y^t A^{-1}}{1 + y^t A^{-1} x}$$

que se puede comprobar mediante verificación directa.

El algoritmo de Broyden con adaptación de la inversa de la matriz es

1. $u^0 \in \Omega$ "cercano" a u
2. $A_0 = F'(u^0)$, calcular A_0^{-1},
3. Para $n = 0, 1, ...$, obtenido u^n, calculamos u^{n+1} resolviendo
   - $s^n = -A_n^{-1} F(u^n)$
   - $u^{n+1} = u^n + s^n$
   - $y^n = F(u^{n+1}) - F(u^n)$
   - $A_{n+1}^{-1} = A_n^{-1} + \frac{(s^n - A_n^{-1} y^n)(s^n)^t A_n^{-1}}{(s^n)^t A_n^{-1} y^n}$

Interpretación de la matriz de iteración de Broyden a partir de un problema de optimización

Vamos a ver otra manera de deducir la matriz de adaptación de Broyden.

Teorema 4.6 Sea $A, B \in \mathbb{R}^{d \times d}$, $s, y \in \mathbb{R}^d$ con $s \neq 0$. Para todo par de normas matriciales $||.||$, $|||.|||$, tales que

$$||A.B|| \leq ||A||.|||B|||$$

y $|||\frac{vv^t}{v^t v}||| = 1$ para todo $v \in \mathbb{R}^d$, la solución del problema: Hallar $\bar{A} \in Q(y, s)$ tal que

$$||\bar{A} - A|| = \inf_{B \in Q(y,s)} ||B - A||$$

donde

$$Q(y, s) = \{B \in \mathbb{R}^{d \times d}; Bs = y\}$$

es

$$\bar{A} = A + \frac{(y - As)s^t}{s^t s}$$

Demostración:

$$\begin{aligned} ||\bar{A} - A|| &= ||\frac{(y - As)s^t}{s^t s}|| = ||\frac{(B - A)ss^t}{s^t s}|| \\ &\leq ||B - A||.|||\frac{ss^t}{s^t s}||| \leq ||B - A|| \end{aligned}$$

■

Observación: En particular podemos tomar $||.|| = |||.|||$ la norma matricial inducida por la norma euclídea habitual. En efecto, para la norma euclídea tenemos para todo $v \in \mathbb{R}^d$

$$\begin{aligned} ||\frac{vv^t}{v^t v}|| &= \frac{1}{||v||^2}||vv^t|| = \frac{1}{||v||^2}\sup_{||x||=1}||vv^t x|| \\ &= \frac{1}{||v||^2}\sup_{||x||=1}|v^t x|||v|| = \frac{1}{||v||}\sup_{||x||=1}|v^t x| \\ &= \frac{1}{||v||}||v^t\frac{v}{||v||}|| = 1 \end{aligned}$$

Método BFGS (Broyden, Fletcher, Goldfarb, Shanno)

La adaptación de Broyden no conserva la simetría de la matriz ni el carácter definido positivo. Cuando se trata de minimizar funciones, es decir $F = \nabla J$ donde $J : \Omega \in \mathbb{R}^d \to \mathbb{R}$, la matriz jacobiana de F, es decir la matriz Hessiana de J es simétrica y como veremos en el apartado siguiente si queremos que el valor de la función vaya disminuyendo en cada paso al aplicar el algoritmo de Newton, tiene que ser definida positiva. Por ello en los métodos de Quasi-Newton cuando se aplican a la minimización de una función es razonable buscar adaptaciones de la matriz que conserven la simetría y el carácter definido positivo. En el siguiente método BFGS, se adapta directamente la inversa de la matriz que sustituye a la matriz Hessiana del método de Newton conservándose la simetría y el carácter definido positivo. Verifica además la ecuación de la secante.

1. $u^0 \in \Omega$ "cercano" a u
2. $B_0 = \left(F'(u^0)\right)^{-1}$ o bien $B_0 = \beta I$
3. Para $n = 0, 1, ...$, obtenido u^n, calculamos u^{n+1} resolviendo
   - $s^n = -B_n F(u^n)$
   - $u^{n+1} = u^n + s^n$

- $y^n = F(u^{n+1}) - F(u^n)$
- $B_{n+1} = \left(I - \frac{1}{(y^n)^t s} s^n (y^n)^t\right) B_n \left(I - \frac{1}{(y^n)^t s^n} y^n (s^n)^t\right) + \frac{1}{(y^n)^t s^n} s^n (s^n)^t$

En el siguiente ejercicio se justifica esta elección.

Ejercicios:

1. Sean $y, s \in \mathbb{R}^d$ tales que $\rho = \frac{1}{y^t s} > 0$. Sea B simétrica y definida positiva

   - Demostrar que $\bar{B}$ dada por
   $$\bar{B} = (I - \rho s y^t) B (I - \rho y s^t) + \rho s s^t$$
   es simétrica y definida positiva.
   - Demostrar que $\bar{B}$ verifica la ecuación $\bar{B} y = s$
   - Sea $J : \Omega \in \mathbb{R}^d \to \mathbb{R}$ una función estrictamente convexa. En el algoritmo BFGS anterior considerar $F(v) = \nabla J(v)$. Sean $s^n = u^{n+1} - u^n$ e $y^n = F(u^{n+1}) - F(u^n)$. Demostrar que $(y^n)^t s^n > 0$ si $s^n \neq 0$.

2. - Aplicar los métodos de Newton y de Broyden para resolver
   $$x + y - 3 = 0$$
   $$x^2 + y^2 - 9 = 0$$
   utilizando como valor inicial $(x^0, y^0)^t = (1, 5)^t$
   - Comprobar que converge hacia la solución del problema anterior $(x, y)^t = (0, 3)^t$
   - Elegir otro valor inicial con el fin de obtener la otra solución del problema, es decir $(x, y)^t = (3, 0)^t$
   - Comprobar la convergencia cuadrática del método de Newton y la convergencia superlineal del método de Broyden para este ejemplo.

4.5. Método de Levenberg-Marquardt

El método de Newton es un método que resuelve la ecuación $\nabla J(u) = 0$ cuando converge, aunque no está garantizado que converja a una solucíon u que es un mínimo de $J(.)$, puede converger hacia un máximo o a un punto silla. Por otra parte el método en general solo tiene convergencia local. En esta sección estudiaremos una modificación del método de Newton en el se que pretende por una parte que la convergencia sea global y por otra que eml método converja siempre a un mínimo.

Para ello consideraremos el método de Newton como método para minimizar funciones y veremos qué condiciones debe cumplir la matriz Hessiana de $J(.)$ para que el método sea propiamente un método de descenso.

Consideraciones sobre el método general de descenso

El método general de descenso se escribe

1. $u^0 \in \mathbb{R}^d$, arbitrario
2. Conocido u^n se calcula u^{n+1} de la siguiente manera:
   - Elegimos la dirección de descenso d^n y $\rho_n \in \mathbb{R}$
   - $u^{n+1} = u^n + \rho_n d^n$

y el método de descenso con parámetro óptimo es

1. $u^0 \in \mathbb{R}^d$, arbitrario
2. Conocido u^n se calcula u^{n+1} de la siguiente manera:
   - Elegimos la dirección de descenso d^n
   - Calculamos $\rho_n \in \mathbb{R}$ tal que

$$J(u^n + \rho_n d^n) < J(u^n + \rho d^n) \quad \forall \rho \in \mathbb{R}$$

   - $u^{n+1} = u^n + \rho_n d^n$

Veamos qué condiciones debe cumplir d^n para asegurarnos que $J(u^{n+1}) < J(u^n)$. Sin perder generalidad podemos suponer que $\rho_n > 0$ y que $||d^n|| = 1$. Si $J(.)$ es diferenciable

$$J(u^{n+1}) = J(u^n + \rho_n d^n) = J(u^n) + \rho_n(\nabla J(u^n), d^n) + |\rho_n|\varepsilon(\rho_n d^n)$$

donde $\varepsilon(v) \to 0$ cuando $||v|| \to 0$. Si elegimos d^n de manera que

$$(\nabla J(u^n), d^n) < 0$$

tomando $\rho_n > 0$ suficientemente pequeño obtendremos $J(u^{n+1}) < J(u^n)$.

Comentario 4.1 : Entre los anteriores métodos:

- En el método de Gradiente $d^n = -\nabla J(u^n)/||\nabla J(u^n||$.
- En el método de relajación d^n se toma sucesivamente igual a las direcciones de los ejes coordenados $e_i = (0, ..., 1, .., 0)^t$ y $\rho_n \in \mathbb{R}$
- En el método de Newton $\rho_n d^n = -H(u^n)^{-1}\nabla J(u^n)$.

Interpretación del método de Newton bajo la óptica de minimización de una función

Sea $J : \mathbb{R}^d \to \mathbb{R}$ diferenciable la función a minimizar. Sea u^n una aproximación de la solución. Desarrollando en serie de Taylor Maclaurin

$$J(v) = J(u^n) + (\nabla J(u^n), v - u^n) + \frac{1}{2}(H(u^n)(v - u^n), v - u^n) + \varepsilon(v - u^n)||v - u^n||^2$$

con $\varepsilon(v) \to 0$ cuando $v \to 0$. La iteración de Newton se puede interpretar como la búsqueda del mínimo de la función cuadrática siguiente

$$q(v) = J(u^n) + (\nabla J(u^n), v - u^n) + \frac{1}{2}(H(u^n)(v - u^n), v - u^n)$$

El mínimo de esta función que llamamos u^{n+1} verifica

$$q(u^{n+1}) \leq q(v) \quad \forall v \in \mathbb{R}^d$$

y viene dado por $\nabla q(u^{n+1}) = 0$, es decir,

$$\nabla J(u^n) + H(u^n)(u^{n+1} - u^n) = 0$$

de donde,

$$u^{n+1} = u^n - H(u^n)^{-1}\nabla J(u^n)$$

que es el método de Newton.

El método de Newton se puede interpretar como método de descenso, tomando por ejemplo, $\rho_n = 1$ y $d^n = -H(u^n)^{-1}\nabla J(u^n)$. Para que en cada paso la función a minimizar decrezca, es decir, $J(u^{n+1}) < J(u^n)$, necesitaremos que

$$(H(u^n)^{-1}\nabla J(u^n), \nabla J(u^n) > 0$$

es decir que la matriz $H(u^n)^{-1}$ y por tanto también la matriz Hessiana $H(u^n)$ sean definidas positivas.

Como esto no siempre es así vamos a modificar la matriz Hessiana con el fin de que en cada paso obtengamos una matriz definida positiva: Para ello sustituimos $H(u^n)$ por

$$\varepsilon_n I + H(u^n)$$

y elegimos $\varepsilon_n > 0$ de manera que la matriz perturbada sea definida positiva. La iteración modificada se escribe: Conocido un valor u^n, calculamos u^{n+1} resolviendo el problema

$$\big(\varepsilon_n I + H(u^n)\big)(u^{n+1} - u^n) = -\nabla J(u^n) \tag{4.1}$$

que se conoce como el método de Levenberg-Marquardt.

El método que obtendremos es un método que se aproximará al método de gradiente si $\varepsilon_n I$ es grande comparado con $H(u^n)$ y se aproximará al método de Newton si ε_n es próximo a cero.

Dada una matriz A no singular y simétrica podemos averiguar si es definida positiva calculando la factorización de Cholesky $A = RR^t$ que se puede poner de la forma $A = LDL^t$ con $R = LD^{1/2}$ donde L es triangular inferior con los elementos de la diagonal iguales a 1. Para matrices simétricas, no singulares, aunque no necesariamente definidas positivas la factorización $A = LDL^t$ es siempre posible. Si la matriz diagonal D tiene elementos negativos, entonces la matriz no es definida positiva y la factorización de Cholesky no es posible.

Queda por determinar una estrategia adecuada para elegir en cada iteración el valor de ε_n. Dada una aproximación u^n y un valor del parámetro $\varepsilon_n > 0$, se efectúa la factorización de $\varepsilon_n I + H(u^n) = LDL^t$. Si algún valor de la matriz diagonal D es negativo (es decir, no existe factorización de Cholesky), entonces multiplicamos ε_n por un factor 4 y repetimos hasta que los elementos de la diagonal de D sean todos positivos. Resolvemos entonces (4.1) haciendo uso de que la matriz ya está factorizada y obtenemos u^{n+1}. Calculamos $J(u^{n+1})$ y determinamos R_n como el cociente entre el decrecimiento de $J(u^n) - J(u^{n+1})$ y el decrecimiento de la correspondiente aproximación cuadrática $q(u^n) - q(u^{n+1})$. Observar que cuanto más próximo esté R_n de la unidad, más fiable es la aproximación cuadrática y más pequeño se puede elegir el valor de ε_n. Bajo esta óptica, si $R_n < 0.25$, elegimos $\varepsilon_{n+1} = 4\varepsilon_n$; si $R_n > 0.75$ elegimos $\varepsilon_{n+1} = \varepsilon_n/2$; en otro caso $\varepsilon_{n+1} = \varepsilon_n$. Además si $R_n \leq 0$, no hay mejora en el valor de $J(.)$, entonces hacemos $u^{n+1} = u^n$.

Algoritmo de Levenberg-Marquardt

1. $u^0 \in \mathbb{R}^d$, $\varepsilon_0 > 0$
2. Para $n = 0, 1, 2, ...$, obtenido u^n, calculamos u^{n+1} resolviendo
   - $A_n = \varepsilon_n I + H(u^n) = LDL^t$
     Si A_n no es definida positiva $\varepsilon_n \leftarrow 4\varepsilon_n$ y volvemos al paso anterior para recalcular A_n y factorizar de nuevo.
     Si A_n es definida positiva
   - $LDL^t s^n = -\nabla J(u^n)$
   - $u^{n+1} = u^n + s^n$
   - Calculamos $J(u^{n+1})$
   - Calculamos $q(u^{n+1}) - J(u^n) = (\nabla J(u^n), s^n) + \frac{1}{2}(H(u^n)s^n, s^n)$
   - Calculamos $R_n = \frac{J(u^{n+1}) - J(u^n)}{q(u^{n+1}) - J(u^n)}$ (observar que en este paso $q(u^n) = J(u^n)$)
   - Si $R_n < 0.25$, $\varepsilon_{n+1} = 4\varepsilon_n$
     Si $R_n > 0.75$, $\varepsilon_{n+1} = \varepsilon_n/2$
     En otro caso $\varepsilon_{n+1} = \varepsilon_n$
     Además si $R_n \leq 0$, $u^{n+1} = u^n$.

Ejercicio (para clases prácticas a resolver con ordenador):

Dada la función $J : \mathbb{R}^2 \to \mathbb{R}$, definida por

$$J(x, y) = (x^2 + y - 1)^2 + (x + y^2 - 8)^2$$

1. Aplicar el método de Newton para obtener todos sus puntos críticos, es decir los puntos $u = (x, y)^t \in \mathbb{R}^2$ verificando, $\nabla J(u) = 0$.
2. Calcular todos los mínimos relativos de $J(.)$ y el mínimo absoluto utilizando el algoritmo de Levenberg-Marquardt.
3. Dibujar sobre una gráfica de isovalores el resultado de las iteraciones con uno y otro método.
4. Calcular todos los mínimos relativos de $J(.)$ utilizando el algoritmo BFGS
5. Comparar las propiedades de convergencia de los tres métodos.

Capítulo 5

Optimización de funciones cuadrátricas

5.1. Generalidades sobre las funciones cuadráticas

Una función cuadrática es una función de la forma

$$J : v \in \mathbb{R}^d \longrightarrow J(v) \in \mathbb{R}$$

donde $J(v) = \frac{1}{2}(Av, v) - (b, v)$ siendo A una matriz de d filas por d columnas simétrica y donde $b \in \mathbb{R}^d$

Una función cuadrática $J(.)$, es elíptica si y solo si la matriz asociada A es definida positiva. En efecto, calculemos la diferencial de $J(.)$ en un punto u,

$$\begin{aligned} DJ(u)(v) &= (\nabla J(u), v) \\ &= \frac{1}{2}((Au, v) + (Av, u)) - (b, v) \\ &= (Au, v) - (b, v) \end{aligned}$$

pues A es simétrica y $(Av, u) = (v, Au) = (Au, v)$. Por tanto

$$\nabla J(u) = Au - b$$

Si $J(.)$ es elíptica, existe $\alpha > 0$ tal que

$$(\nabla J(u) - \nabla J(v), u - v) \geq \alpha ||u - v||^2 \quad \forall u, v \in \mathbb{R}^d$$

como $\nabla J(u) = Au - b$ y $\nabla J(v) = Av - b$ resulta

$$(Au - Av, u - v) \geq \alpha ||u - v||^2 \quad \forall u, v \in \mathbb{R}^d$$

como u y v son cualesquiera

$$(Av, v) \geq \alpha ||v||^2$$

Recíprocamente, si A es definida positiva, existe $\alpha > 0$ tal que

$$(Av, v) \geq \alpha ||v||^2 \quad \forall v \neq 0$$

por tanto $\forall u, v \in \mathbb{R}^d \quad u \neq v$, entonces

$$(A(u - v), u - v) \geq \alpha ||u - v||^2$$

y finalmente

$$(\nabla J(u) - \nabla J(v), u - v) \geq \alpha ||u - v||^2$$

Si A es definida positiva el mínimo de $J(.)$ está caracterizado por $\nabla J(u) = 0$ es decir, $Au = b$. De modo que el problema de hallar $u \in \mathbb{R}^d$ tal que $J(u) = \text{ínf}_{v \in \mathbb{R}^d} J(v)$ equivale a resolver $Au = b$.

Vamos a considerar ahora la minimización de funciones cuadráticas con matriz asociada A definida positiva, por tanto elípticas. Sabemos que el problema tiene solución única.

5.2. Métodos de descenso

5.2.1. Método general de descenso

Sea $u^0 \in \mathbb{R}^d$ vector inicial arbitrario. Construiremos una sucesión de vectores $(u^n)_n$ a partir de u^0. El paso general para construir u^{n+1} a partir de u^n es

- Fijamos una dirección de descenso $d^n \neq 0$ en el punto u^n
- Resolvemos el problema del mínimo siguiente: Hallar $\rho_n = \rho(u^n, d^n)$ tal que

  $$J(u^n + \rho_n d^n) = \inf_{\rho \in \mathbb{R}} J(u^n + \rho d^n)$$

  que tiene solución única pues $J(.)$ es elíptica.
- $$u^{n+1} = u^n + \rho_n d^n$$

En el caso de funciones cuadráticas el cálculo de $\rho_n = \rho(u^n, d^n)$ es sencillo, en efecto derivando la función $\rho \to J(u^n + \rho d^n)$ e igualando a 0

$$(\nabla J(u^n + \rho_n d^n), d^n) = 0$$

$$(A(u^n + \rho_n d^n) - b, d^n) = 0$$

despejando ρ_n

$$\rho_n = \frac{(b - Au^n, d^n)}{(Ad^n, d^n)} = \frac{(r^n, d^n)}{(Ad^n, d^n)}$$

donde hemos introducido el residuo correspondiente $r^n = b - Au^n$.

Algoritmo general de descenso

1. $u^0 \in \mathbb{R}^d$ arbitrario.
2. $r^n = b - Au^n$
3. $\rho_n = \frac{(r^n, d^n)}{(Ad^n, d^n)}$
4. $u^{n+1} = u^n + \rho_n d^n$

o bien observando que $r^{n+1} = r^n - \rho_n Ad^n$

1. $u^0 \in \mathbb{R}^d$ arbitrario; $r^0 = b - Au^0$.
2. $\rho_n = \frac{(r^n, d^n)}{(Ad^n, d^n)}$
3. $u^{n+1} = u^n + \rho_n d^n$
4. $r^{n+1} = r^n - \rho_n Ad^n$

Para distintas elecciones d^n, obtenemos distintos métodos.

Ejercicio: Verificar que si en el método general de descenso elegimos como direcciones de descenso d^n las direcciones de los ejes, es decir, $e_i = (0, ..., 0, 1, 0, ..., 0)$ obtenemos el método de Gauss-Seidel para resolver $Au = b$.

5.2.2. Propiedades de convergencia de los métodos de descenso

Vamos a estudiar ahora las Propiedades de los métodos de descenso.

Propiedad 5.1 Cualquiera que sea la dirección de descenso d^n elegida, para ρ_n óptimo se tiene para todo $n \geq 0$

$$(d^n, r^{n+1}) = 0$$

es decir, la dirección de descenso y el nuevo gradiente de la función son ortogonales.

Demostración: Hemos comprobado ya que $r^{n+1} = r^n - \rho_n A d^n$. De donde,

$$\begin{aligned}(d^n, r^{n+1}) &= (d^n, r^n - \rho_n A d^n) = (d^n, r^n) - \rho_n (d^n, A d^n) \\ &= (d^n, r^n) - \frac{(r^n, d^n)}{(A d^n, d^n)} (A d^n, d^n) = 0\end{aligned}$$

Propiedad 5.2 El problema de minimizar una función $J : v \to \frac{1}{2}(Av, v) - (b, v)$ con A simétrica y definida positiva, es equivalente a minimizar

$$E(v) = (A(v-u), v-u) = ||v-u||_A^2$$

donde u es la solución buscada, es decir, verificando $Au = b$.

Nota 5.1 Hemos introducido la notación $||v||_A = (Av, v)^{1/2}$ para la norma asociada al producto escalar $u, v \longrightarrow (Au, v)$

Nota 5.2 $E(.)$ es la función error asociada al valor v, más precisamente, es el cuadrado de la norma asociada a la matriz A del error $e = v - u$.

Demostración:

$$\begin{aligned}E(v) &= (A(v-u), v-u) = (Av, v) - 2(Au, v) + (Au, u) \\ &= (Av, v) - 2(b, v) + (Au, u) \\ &= 2J(v) + (Au, u)\end{aligned}$$

Como (Au, u) es constante, es decir, independiente de v, $E(.)$ y $2J(.)$ y por lo tanto $J(.)$ alcanzan el mínimo en el mismo punto u. ■

Demos ahora una interpretación geométrica de los métodos de descenso: Consideremos el caso de funciones cuadráticas en $\mathbb{R}^2$. Las ecuación $E(v) = cte$ es una elipse. Para diferentes valores $E(v) = E(u^n)$ obtenemos una familia de elipses concéntricas centradas en el mínimo de u de la función y que representan las curvas de nivel. El vector d^n es tangente a la elipse $E(v) = E(u^{n+1})$. Como r^{n+1} es ortogonal a d^n, r^{n+1} es ortogonal a la tangente de la curva de nivel.

Vamos a estudiar ahora cuáles son las posibles elecciones de d^n que permitan asegurar la convergencia del método de descenso.

Lema 5.1 Para $d^n \neq 0$ y $r^n \neq 0$

$$E(u^{n+1}) = E(u^n)(1 - \gamma_n)$$

donde

$$\gamma_n = \frac{(r^n, d^n)^2}{(A d^n, d^n)(A^{-1} r^n, r^n)}$$

Demostración: Sustituimos el valor de ρ_n en la expresión de $E(u^{n+1})$:

$$\begin{aligned} E(u^{n+1}) &= E(u^n + \rho_n d^n) = (A(u^n - u + \rho_n d^n), u^n - u + \rho_n d^n) \\ &= (A(u^n - u), u^n - u) + 2\rho_n (A(u^n - u), d^n) + \rho_n^2 (Ad^n, d^n) \\ &= E(u^n) - 2\rho_n (r^n, d^n) + \rho_n^2 (Ad^n, d^n) \\ &= E(u^n) - 2\frac{(r^n, d^n)^2}{(Ad^n, d^n)} + \frac{(r^n, d^n)^2}{(Ad^n, d^n)^2}(Ad^n, d^n) \\ &= E(u^n) - \frac{(r^n, d^n)^2}{(Ad^n, d^n)} \\ &= E(u^n)(1 - \frac{1}{E(u^n)}\frac{(r^n, d^n)^2}{(Ad^n, d^n)}) \end{aligned}$$

Finalmente como

$$E(u^n) = (A(u^n - u), u^n - u) = (r^n, u - u^n) = (r^n, A^{-1}r^n)$$

obtenemos el resultado. ■

Observación: El número γ_n es siempre positivo pues A y por lo tanto también A^{-1} son matrices simétricas y definidas positivas.

Lema 5.2 Mayoración de γ_n: Cualquiera que sea $d^n \neq 0$, y siendo ρ_n el valor óptimo local tenemos la siguiente relación válida para $n \geq 0$

$$\gamma_n = \frac{(r^n, d^n)^2}{(Ad^n, d^n)(A^{-1}r^n, r^n)} \geq \frac{1}{\kappa(A)}(\frac{r^n}{||r^n||}, \frac{d^n}{||d^n||})^2$$

donde $\kappa(A)$ es el número de condicionamiento de la matriz A.

Demostración: La matriz A siendo simétrica y definida positiva, tiene todos sus valores propios reales y positivos.

$$0 < \lambda_{min} = \lambda_1 \leq \lambda_2 \leq .... \leq \lambda_d = \lambda_{max}$$

Sabemos $\frac{(Ad^n, d^n)}{||d^n||^2} \leq \lambda_{max}$ es decir $(Ad^n, d^n) \leq \lambda_{max}||d^n||^2$.

Los valores propios de A^{-1} son

$$0 < \frac{1}{\lambda_{max}} \leq, ..., \leq \frac{1}{\lambda_{min}}$$

de donde $\frac{(A^{-1}r^n, r^n)}{||r^n||^2} \leq \frac{1}{\lambda_{min}}$, es decir $(A^{-1}r^n, r^n) \leq \frac{1}{\lambda_{min}}||r^n||^2$.

Y finalmente

$$\frac{(Ad^n, d^n)(A^{-1}r^n, r^n)}{||d^n||^2||r^n||^2} \leq \frac{\lambda_{max}}{\lambda_{min}} = \kappa(A)$$

de donde

$$\gamma_n \geq \frac{1}{\kappa(A)} \frac{(r^n, d^n)^2}{||r^n||^2 ||d^n||^2}$$

■

El anterior lema nos va a permitir elegir las direcciones de descenso:

Teorema 5.1 Sea ρ_n el óptimo local; Si para toda dirección d^n se verifica para todo $n \geq 0$

$$(\frac{r^n}{||r^n||}, \frac{d^n}{||d^n||})^2 \geq \mu > 0$$

donde μ es independiente de n, entonces la sucesión $(u^n)_n$ generada por el algoritmo de descenso es convergente hacia la solución que minimiza $E(v)$, y por lo tanto $J(v)$.

Demostración: El lema anterior permite escribir

$$E(u^{n+1}) = E(u^n)(1 - \gamma_n) \leq E(u^n)(1 - \frac{\mu}{\kappa(A)})$$

de donde, aplicando recursivamente la relación anterior

$$E(u^n) \leq (1 - \frac{\mu}{\kappa(A)})^n E(u^0)$$

Por otra parte, de la desigualdad de Cauchy-Schwarz

$$0 < \mu \leq (\frac{r^n}{||r^n||}, \frac{d^n}{||d^n||})^2 \leq 1$$

y como $\kappa(A) \geq 1$, resulta $0 \leq 1 - \frac{\mu}{\kappa(A)} < 1$, de donde

$$\lim_{n\to\infty} ||u^n - u||_A^2 = \lim_{n\to\infty} E(u^n) = 0$$

y también

$$0 < \lambda_{min} \leq \frac{(A(u^n - u), u^n - u)}{||u^n - u||^2}$$

$$||u^n - u||^2 \leq \frac{1}{\lambda_{min}} (A(u^n - u), u^n - u) = \frac{1}{\lambda_{min}} E(u^n) \overset{n\to\infty}{\to} 0$$

■

Observación: En el marco de los métodos de descenso con $\rho = \rho^n$ óptimo local, el anterior teorema permite dar una condición suficiente en la elección de d^n para asegurar la convergencia: Para todo $n \geq 0$ d^n debe ser no ortogonal a r^n.

5.3. Método de gradiente con paso óptimo

5.3.1. Descipción del método de gradiente con paso óptimo

Entre las direcciones posibles d^n que aseguran la convergencia del método de descenso con elección óptima del parámetro ρ_n, una elección obvia es $d^n = r^n$, para la que el parámetro μ verifica

$$\mu = (\frac{r^n}{||r^n||}, \frac{d^n}{||d^n||})^2 = ||\frac{r^n}{||r^n||}||^2 = 1$$

El método de gradiente consiste en elegir el gradiente como dirección de descenso, concretamente $d^n = -\nabla J(u^n) = -(Au^n - b) = r^n$. El algoritmo de gradiente con paso óptimo se escribe

1. $u^0 \in \mathbb{R}^d$ arbitrario.
2. Una vez obtenido u^n se calcula u^{n+1} de la siguiente manera:
   - $r^n = b - Au^n$
   - $\rho_n = \frac{(r^n, r^n)}{(Ar^n, r^n)}$
   - $u^{n+1} = u^n + \rho_n r^n$

para evitar el producto de una matriz por un vector en el cálculo de r^n observemos

$$Au^{n+1} = Au^n + \rho_n Ar^n$$

de donde

$$r^{n+1} = r^n - \rho_n Ar^n$$

y el algoritmo queda de la forma

1. $u^0 \in \mathbb{R}^d$ arbitrario.
   - $r^0 = b - Au^0$
   - $\rho_0 = \frac{(r^0, r^0)}{(Ar^0, r^0)}$
   - $u^1 = u^0 + \rho_0 r^0$
   - $r^1 = r^0 - \rho_0 Ar^0$
2. Una vez obtenido u^n se calcula u^{n+1} de la siguiente manera:
   - $\rho_n = \frac{(r^n, r^n)}{(Ar^n, r^n)}$
   - $u^{n+1} = u^n + \rho_n r^n$

- $r^{n+1} = r^n - \rho_n A r^n$

La expresión del error es en este caso

$$E(u^{n+1}) = E(u^n)(1 - \frac{||r^n||^4}{(Ar^n, r^n)(A^{-1}r^n, r^n)})$$

5.3.2. Convergencia del método de gradiente con paso óptimo

Vamos a estudiar con más profundidad la convergencia del método de gradiente con paso óptimo. Para ello necesitaremos un lema previo.

Lema 5.3 Desigualdad de Kantorovich

Sea A una matriz simétrica y definida positiva. Para todo $x \neq 0$, tenemos:

$$1 \leq \frac{(Ax, x)(A^{-1}x, x)}{||x||^4} \leq \frac{(\lambda_{min} + \lambda_{max})^2}{4\lambda_{min}\lambda_{max}}$$

donde λ_{min} y λ_{max} son los valores propios mínimo y máximo de A respectivamente.

Demostración: Si A e simétrica definida positiva (resp. A^{-1}) es diagonalizable y tiene todos sus valores propios positivos y se puede elegir una base ortonormal de vectores propios. Sean

$$0 < \lambda_{min} = \lambda_1 \leq, ..., \leq \lambda_d = \lambda_{max}$$

los valores propios de A. Respectivamente, los valores propios de A^{-1} son

$$0 < \frac{1}{\lambda_{max}} = \frac{1}{\lambda_d} \leq, ..., \leq \frac{1}{\lambda_1} = \frac{1}{\lambda_{min}}$$

y sea $(v_i)_{i=1}^d$ la base ortonormal de vectores propios. Para $x \in \mathbb{R}^d$, $x = \Sigma_{i=1}^d x_i v_i$; $||x||^2 = \Sigma_{i=1}^d |x_i|^2$ y también $(Ax, x) = \Sigma_{i=1}^d \lambda_i |x_i|^2$ y $(A^{-1}x, x) = \Sigma_{i=1}^d \frac{1}{\lambda_i}|x_i|^2$. De modo que

$$\frac{(Ax, x)(A^{-1}x, x)}{||x||^4} = \frac{\Sigma_{i=1}^d \lambda_i |x_i|^2 \Sigma_{i=1}^d \frac{1}{\lambda_i}|x_i|^2}{\Sigma_{j=1}^d x_j^2 \Sigma_{j=1}^d x_j^2}$$

denotando $\alpha_i = \frac{x_i^2}{\Sigma_{j=1}^d x_j^2}$ para $i = 1, ..., d$ resulta $\alpha_i \geq 0$, $\Sigma_{i=1}^d \alpha_i = 1$ y

$$\frac{(Ax, x)(A^{-1}x, x)}{||x||^4} = \Sigma_{i=1}^d \alpha_i \lambda_i \Sigma_{i=1}^d \alpha_i \frac{1}{\lambda_i}$$

Consideremos en el plano $\mathbb{R}^2$ los puntos $M_i = (\lambda_i, \frac{1}{\lambda_i})$ El punto $M = \Sigma_{i=1}^d \alpha_i M_i = (\Sigma_{i=1}^d \alpha_i \lambda_i, \Sigma_{i=1}^d \alpha_i \frac{1}{\lambda_i})$ combinación lineal convexa de los puntos M_i estará contenido en la zona rayada de la figura. Llamando

$\overline{\lambda} = \Sigma_{i=1}^{d} \alpha_i \lambda_i$, resulta que la ordenada del punto M, $\Sigma_{i=1}^{d} \alpha_i \frac{1}{\lambda_i}$ es inferior a la ordenada del punto $\overline{M}$ situado sobre la recta $\overline{M_1 M_d}$. Por otra parte la ordenada del punto M es superior a $\frac{1}{\overline{\lambda}}$. Mediante un sencillo cálculo,

$$\overline{M} = (\overline{\lambda}, \frac{\lambda_1 + \lambda_d - \overline{\lambda}}{\lambda_1 \lambda_d})$$

Es decir,

$$\begin{aligned} 1 = \overline{\lambda}\,(1/\overline{\lambda}) &\leq \Sigma_{i=1}^{d} \alpha_i \lambda_i \Sigma_{i=1}^{d} \alpha_i \frac{1}{\lambda_i} \leq \overline{\lambda} \frac{\lambda_1 + \lambda_d - \overline{\lambda}}{\lambda_1 \lambda_d} \\ &\leq \max_{\lambda_1 \leq \lambda \leq \lambda_d} (\lambda \frac{\lambda_1 + \lambda_d - \lambda}{\lambda_1 \lambda_d}) = \frac{(\lambda_1 + \lambda_d)^2}{4 \lambda_1 \lambda_d} \end{aligned}$$

La función $f(\lambda) = \lambda \frac{\lambda_1 + \lambda_d - \lambda}{\lambda_1 \lambda_d}$ alcanza su máximo para $\lambda = \frac{\lambda_1 + \lambda_d}{2}$ y ese valor máximo vale $\frac{(\lambda_1 + \lambda_d)^2}{4 \lambda_1 \lambda_d}$ ■

Teorema 5.2 El método del gradiente con paso local óptimo es convergente. El factor de reducción del error en cada paso es menor o igual que $\frac{\kappa(A)-1}{\kappa(A)+1}$.

Demostración: Aplicando la desigualdad de Kantorovich para el residuo en la n-ésima iteración r^n del método de gradiente tendremos

$$\frac{||r^n||^4}{(Ar^n, r^n)(A^{-1}r^n, r^n)} \geq \frac{4\lambda_{min}\lambda_{max}}{(\lambda_{min} + \lambda_{max})^2} = \frac{4\kappa(A)}{(1 + \kappa(A))^2}$$

entonces

$$E(u^{n+1}) \leq E(u^n)(1 - \frac{4\kappa(A)}{(1 + \kappa(A))^2}) = E(u^n)(\frac{\kappa(A) - 1}{\kappa(A) + 1})^2$$

de donde finalmente

$$E(u^{n+1}) \leq E(u^0)(\frac{\kappa(A) - 1}{\kappa(A) + 1})^{2n}$$

o lo que es lo mismo

$$||u^n - u||_A \leq ||u^0 - u||_A (\frac{\kappa(A) - 1}{\kappa(A) + 1})^n$$

. ■

Observación: Si $\kappa(A)$ es próximo a 1, el método converge rápidamente. Cuando $\kappa(A) = 1$ todos los valores propios son iguales. Tomemos $A = \lambda I$ y $E(v) = (A(v-u), v-u) = \lambda(v-u, v-u) = \lambda||v-u||^2$. Es decir la ecuación $E(v) = cte$ es la ecuación de una superficie esférica de centro u (una circunferencia si $d = 2$) y el método converge en una sola iteración.

Por el contrario si $\kappa(A)$ es grande, los valores propios λ_{min} y λ_{max} son muy diferentes; Los elipsoides $E(v) = cte$ son entonces muy aplastados. La convergencia es lenta. Para que

$$\frac{E(u^n)}{E(u^0)} \leq \varepsilon$$

es suficiente que

$$(\frac{\kappa(A)-1}{\kappa(A)+1})^{2n} \leq \varepsilon$$

es decir

$$2n\ ln(\frac{\kappa(A)+1}{\kappa(A)-1}) \approx ln\frac{1}{\varepsilon}$$

Como para valores de $\kappa(A)$ mucho mayores que 1,

$$ln(\frac{\kappa(A)+1}{\kappa(A)-1}) = ln(1+\frac{2}{\kappa(A)-1}) \approx \frac{2}{\kappa(A)-1} \approx \frac{2}{\kappa(A)}$$

resulta

$$n \approx \frac{\kappa(A)}{4} ln\frac{1}{\varepsilon}$$

El número de iteraciones es proporcional a $\kappa(A)$.

Ejercicio: Método de gradiente con paso constante

Puesto que el método de gradiente con paso óptimo , debido al efecto zig-zag y al coste del cálculo de ρ_n puede no resultar óptimo desde un punto de vista global, podemos pensar en un método de gradiente con parámetro constante, donde u^{n+1} se calcula a partir de u^n mediante

1. $r^n = b - Au^n$
2. $u^{n+1} = u^n + \alpha r^n$

y donde α es independiente de n.

- Demostrar que el error en la iteración n-ésima $e^n = u^n - u$ se expresa $e^n = (I - \alpha A)^n e^0$ donde I la matriz identidad y que la condición necesaria y suficiente de convergencia es que el radio espectral de $I - \alpha A$, verifique $\rho(I - \alpha A) < 1$, es decir, que α y los valores propios de A, $\lambda_i,\ i = 1, ..., d$ verifiquen

$$|1 - \alpha\lambda_i| < 1,\ i = 1, ..., d$$

- Supongamos que A sea simétrica y definida positiva y sean $0 \leq \lambda_1 < ... \leq \lambda_d$ los valores propios de A. Demostrar que la condición de convergencia es

$$0 < \alpha < \frac{2}{\lambda_d}$$

- Demostrar que el valor óptimo de α es

$$\alpha_{opt} = \frac{2}{\lambda_1 + \lambda_d}$$

  y el correspondiente radio espectral para este valor óptimo es

$$\rho(I - \alpha_{opt} A) = \frac{\kappa(A)-1}{\kappa(A)+1}$$

5.4. Método de Gradiente Conjugado

5.4.1. Introducción

En esta sección investigaremos nuevas direcciones de descenso d^n para ser utilizadas con el paso local óptimo, $\rho_n = \frac{(d^n,r^n)}{(Ad^n,d^n)}$. Como hemos demostrado anteriormente $(d^{n-1}, r^n) = 0$.

Vamos a buscar ahora la nueva dirección de descenso d^n en el plano formado por las dos direcciones ortogonales r^n y d^{n-1}. Pongamos

$$d^n = r^n + \beta_n d^{n-1}$$

y calculemos el parámetro β_n de modo que la reducción del error sea lo más grande posible en cada paso. Tenemos

$$E(u^{n+1}) = E(u^n)(1 - \frac{(r^n, d^n)^2}{(Ad^n, d^n)(A^{-1}r^n, r^n)})$$

Elegiremos β_n de manera que

$$\frac{(r^n, d^n)^2}{(Ad^n, d^n)(A^{-1}r^n, r^n)}$$

sea máximo. Como

$$(r^n, d^n) = (r^n, r^n + \beta_n d^{n-1}) = ||r^n||^2 + \beta_n (r^n, d^{n-1}) = ||r^n||^2$$

elegiremos $d^0 = r^0$ de modo que esta relación se verifique también para $n = 0$. Tendremos pues $(r^n, d^n) = ||r^n||^2$ para todo $n \geq 0$. La determinación del máximo de

$$\frac{(r^n, d^n)^2}{(Ad^n, d^n)(A^{-1}r^n, r^n)} = \frac{||r^n||^4}{(Ad^n, d^n)(A^{-1}r^n, r^n)}$$

se reduce a minimizar (Ad^n, d^n). Desarrollando este término

$$\begin{aligned} (Ad^n, d^n) &= (A(r^n + \beta_n d^{n-1}), r^n + \beta_n d^{n-1}) \\ &= \beta_n^2 (Ad^{n-1}, d^{n-1}) + 2\beta_n (Ad^{n-1}, r^n) + (Ar^n, r^n) \end{aligned}$$

Para que este trinomio sea mínimo, hay que elegir β_n de modo que

$$\beta_n (Ad^{n-1}, d^{n-1}) + (Ad^{n-1}, r^n) = 0$$

de donde deducimos

$$\beta_n = -\frac{(Ad^{n-1}, r^n)}{(Ad^{n-1}, d^{n-1})}$$

y también

$$(Ad^{n-1}, r^n + \beta_n d^{n-1}) = 0$$

es decir,

$$(Ad^{n-1}, d^n) = 0$$

Definición 5.1 Cuando dos vectores u y v verifican la relación $(Au, v) = 0$ se dice que son A-conjugados. Cuando A es simétrica y definida positiva la aplicación $u, v \to (Au, v)$ es un producto escalar. La relación para dos vectores de A-conjugación no es más que la ortogonalidad con respecto al producto escalar $(A., .)$.

Veamos algunas propiedades que relacionan los residuos sucesivos y el valor de β_n.

Lema 5.4 Se tienen las siguientes relaciones, válidas si $r^n \neq 0$ para $i = 0, ..., n$:

$$(r^{n+1}, r^n) = 0 \quad n \geq 0$$

además

$$\beta_0 = 0, \quad \beta_n = \frac{||r^n||^2}{||r^{n-1}||^2} \quad \forall n \geq 1$$

Demostración:

$\beta_0 = 0$ pues se toma como primera dirección de descenso $d^0 = r^0$.

$$\begin{aligned}(r^{n+1}, r^n) &= (r^n - \rho_n A d^n, r^n) = ||r^n||^2 - \rho_n (Ad^n, r^n) \\ &= ||r^n||^2 - \rho_n (Ad^n, d^n - \beta_n d^{n-1}) \\ &= ||r^n||^2 - \rho_n (Ad^n, d^n) + \rho_n \beta_n (Ad^n, d^{n-1}) = 0\end{aligned}$$

teniendo en cuenta que

$$(Ad^n, d^{n-1}) = 0$$

y que

$$\rho_n = \frac{(r^n, d^n)}{(Ad^n, d^n)} = \frac{||r^n||^2}{(Ad^n, d^n)}$$

Por otra parte $Ad^{n-1} = \frac{1}{\rho_{n-1}}(r^{n-1} - r^n)$ para $n \geq 1$ de donde

$$(Ad^{n-1}, r^n) = \frac{1}{\rho_{n-1}}((r^{n-1} - r^n), r^n) = -\frac{1}{\rho_{n-1}}||r^n||^2$$

y también

$$(Ad^{n-1}, d^{n-1}) = \frac{1}{\rho_{n-1}}((r^{n-1} - r^n), d^{n-1}) = \frac{1}{\rho_{n-1}}(r^{n-1}, d^{n-1}) = \frac{1}{\rho_{n-1}}||r^{n-1}||^2$$

de donde finalmente

$$\beta_n = -\frac{(r^n, Ad^{n-1})}{(d^{n-1}, Ad^{n-1})} = \frac{||r^n||^2}{||r^{n-1}||^2}$$

■

5.4.2. Algoritmo de Gradiente Conjugado

El algoritmo de Gradiente Conjugado es el siguiente:

1. u^0 arbitrario y $d^0 = r^0 = b - Au^0$.
2. Para $n = 0, 1, ...$

   - $$\rho_n = \frac{||r^n||^2}{(Ad^n, d^n)}$$
   - $$u^{n+1} = u^n + \rho_n d^n$$
     $$r^{n+1} = r^n - \rho_n A d^n$$
   - $$\beta_{n+1} = \frac{||r^{n+1}||^2}{||r^n||^2}$$
   - $$d^{n+1} = r^{n+1} + \beta_{n+1} d^n$$

Ejercicio:

Sea c el número medio de coeficientes no nulos por línea de la matriz A de orden N. Calcular el número de operaciones de una iteración del algoritmo de Gradiente Conjugado. (Resp: $(c + 5)N + 2$ multiplicaciones y $(c + 4)N - 2$ sumas.)

5.4.3. Propiedades del algoritmo de Gradiente Conjugado

Teorema 5.3 En el método de Gradiente Conjugado, tomando $d^0 = r^0 = b - Au^0$, se verifica, para todo $n \geq 1$ siempre que $r^k \neq 0$ y para todo $k \leq n - 1$, las siguientes propiedades

$$(r^n, d^k) = 0 \quad \forall\, k \leq n - 1$$

$$(d^n, Ad^k) = (Ad^n, d^k) = 0 \quad \forall\, k \leq n - 1$$

$$(r^n, r^k) = 0 \quad \forall\, k \leq n - 1$$

Demostración: Utilizaremos el principio de inducción.

1. Para n=1 se verifica $(r^1, d^0) = (r^1, r^0) = 0$ por la propia definición del método y al propiedades señaladas anteriormente.

   Por otra parte

   $$(d^1, Ad^0) = (Ad^1, d^0) = 0$$

   es la relación de conjugación para dos direcciones de descenso consecutivas.

2. Supongamos que se verifican las relaciones del teorema para el valor n y veamos que en ese caso se verifican también para $n+1$ y $k \leq n$.

   Para la primera relación tenemos en el caso $k = n$, tenemos $(r^{n+1}, d^n) = 0$ que se cumple en todos los métodos de descenso con paso óptimo.

   Por otra parte para $k = 1, 2, ..., n-1$ tendremos

   $$(r^{n+1}, d^k) = (r^n - \rho_n Ad^n, d^k) = (r^n, d^k) - \rho_n (Ad^n, d^k)$$

   siendo estos dos últimos términos nulos por la hipótesis de recurrencia.

   Para la segunda relación en el caso $k = n$, tenemos $(Ad^{n+1}, d^n) = 0$ es la relación de conjugación.

   Para $k = 1, 2, ..., n-1$ tenemos

   $$(d^{n+1}, Ad^k) = (r^{n+1} + \beta_{n+1} d^n, Ad^k) = (r^{n+1}, Ad^k) + \beta_{n+1} (d^n, Ad^k)$$

   donde en el último término del segundo miembro $(d^n, Ad^k) = 0$ por la hipótesis de inducción. Además $r^{k+1} = r^k - \rho_k Ad^k$ de donde

   $$Ad^k = \frac{1}{\rho_k}(r^k - r^{k+1})$$

   resultando

   $$(d^{n+1}, Ad^k) = \frac{1}{\rho_k}(r^{n+1}, r^k - r^{k+1}) = \frac{1}{\rho_k}(r^{n+1}, d^k - \beta_k d^{k-1} - d^{k+1} + \beta_{k+1} d^k)$$

   siendo todos los términos nulos por la primera relación.

   Para la tercera relación en el caso $k = n$ ya ha sido demostrado en el lema anterior.

   Para $k = 1, 2, ..., n-1$ tenemos

   $$(r^{n+1}, r^k) = (r^{n+1}, d^k - \beta_k d^{k-1}) = (r^{n+1}, d^k) - \beta_k (r^{n+1}, d^{k-1}) = 0$$

   siendo los dos últimos términos nulos por verificarse la primera relación.

■

Corolario 5.1 El algoritmo de Gradiente Conjugado para la resolución de un sistema con matriz simétrica y definida positiva de orden N converge en al menos N iteraciones.

Demostración: O bien $r^n = 0$ y tenemos la convergencia en $n \leq N-1$ iteraciones o por el contrario r^N es ortogonal a $d^0, d^1, ..., d^{N-1}$ que son N vectores linealmente independientes (por ser A-ortogonales) del espacio $\mathbb{R}^N$. Necesariamente $r^N = 0$.■

Observación: El método de gradiente conjugado, introducido en 1952 por Hestenes y Stiefel es pues teóricamente un método directo pues se obtiene salvo errores de redondeo la solución exacta con un número finito de iteraciones. Sin embargo en la práctica debido a errores de redondeo la relaciones de conjugación no se tienen exactamente y el método se considera como un método iterativo. A continuación estudiaremos como depende el factor de convergencia con el condicionamiento de la matriz A. Consideraremos luego técnicas de precondicionamiento que tienen como finalidad mejorar la convergencia. El objetivo es siempre lograr la convergencia con un número de iteraciones considerablemente menor que el número de ecuaciones.

Definición 5.2 Espacios de Krylov: Llamamos espacio de Krylov de dimensión n, $\mathcal{K}_n$, al espacio generado por los vectores $r^0, Ar^0, ..., A^{n-1}r^0$. Es decir

$$\mathcal{K}_n = [r^0, Ar^0, ..., A^{n-1}r^0]$$

Teorema 5.4 En el método de gradiente conjugado, eligiendo $d^0 = r^0 = b - Au^0$ y siempre que $r^0 \neq 0$ se tiene

$$[r^0, Ar^0, ..., A^n r^0] = [r^0, r^1, ..., r^n]$$

$$[r^0, Ar^0, ..., A^n r^0] = [d^0, d^1, ..., d^n]$$

Demostración: Para $n = 1$ como $d^0 = r^0$ y $d^1 = r^1 + \beta_1 d^0$ y por lo tanto $r^1 = d^1 - \beta_1 d^0$ podemos escribir

$$[r^0, r^1] = [d^0, d^1]$$

Por otra parte $r^1 = r^0 - \rho_0 A d^0$ con $\rho_0 = \frac{||r^0||^2}{(Ar^0, r^0)} \neq 0$ así pues $Ar^0 = Ad^0 = \frac{r^0 - r^1}{\rho_0}$ de donde

$$[r^0, Ad^0] = [r^0, Ar^0] = [r^0, r^1]$$

Las relaciones son ciertas para $n = 1$. Supongamos ahora que las relaciones son ciertas para n y veamos que en ese caso también lo son para $n + 1$:

Tendremos por la hipótesis de inducción $d^n \in [r^0, Ar^0, ..., A^n r^0]$ y por otra parte $Ad^n \in A[r^0, Ar^0, ..., A^n r^0] = [Ar^0, A^2 r^0, ..., A^{n+1} r^0]$.

Como $r^{n+1} = r^n - \rho_n A d^n$ tenemos

$$r^{n+1} \in [r^0, Ar^0, ..., A^{n+1} r^0]$$

Recíprocamente demostraremos que $A^{n+1} r^0 \in [r^0, r^1, ..., r^{n+1}]$. En efecto, según se ha visto r^{n+1} es una combinación lineal de términos de la forma $A^k r^0$ para $0 \leq k \leq n+1$, es decir,

$$r^{n+1} = \Sigma_{k=0}^{n+1} \gamma_k A^k r^0 = \Sigma_{k=0}^{n} \gamma_k A^k r^0 + \gamma_{n+1} A^{n+1} r^0$$

Veamos que podemos despejar el término $A^{n+1}r^0$. En efecto, $r^{n+1} \notin [r^0, Ar^0, ..., A^n r^0] = [d^0, d^1, ..., d^n]$ pues r^{n+1} es ortogonal a $d^0, d^1, ..., d^n$ y también a $r^0, r^1, ..., r^n$ por el teorema anterior.

Podemos pues escribir

$$A^{n+1}r^0 = \frac{1}{\gamma_{n+1}}(r^{n+1} - \Sigma_{k=0}^{n}\gamma_k A^k r^0)$$

Gracias a la hipótesis de inducción

$$\Sigma_{k=0}^{n}\gamma_i A^k r^0 \in [r^0, ..., r^n]$$

de donde

$$A^{n+1}r^0 \in [r^0, ..., r^{n+1}]$$

resumiendo

$$[r^0, ..., r^{n+1}] = [r^0, Ar^0, ..., A^{n+1}r^0]$$

Análogamente se demuestra

$$[d^0, ..., d^{n+1}] = [r^0, Ar^0, ..., A^{n+1}r^0]$$

■

Teorema 5.5 El valor u^n obtenido en la $n-$ésima iteración del algoritmo de gradiente conjugado verifica

$$E(u^n) \leq E(v) \quad \forall v \in u^0 + \mathcal{K}_n$$

Demostración: Como $u^n = u^0 + \Sigma_{i=0}^{n-1}\rho_i d^i \in u^0 + \mathcal{K}_n$ para expresar que $E(u^n)$ es el mínimo de $E(v)$ sobre $u^0 + \mathcal{K}_n$, es necesario y suficiente que

$$E(u^n) \leq E(u^0 + w) \quad \forall w \in \mathcal{K}_n$$

es decir

$$(\nabla E(u^n), w) = 0 \quad \forall w \in \mathcal{K}_n$$

o sea

$$2(r^n, w) = 0 \quad \forall w \in \mathcal{K}_n$$

pero esto es cierto pues $(r^n, r^i) = 0$ para todo $i \leq n-1$ y según el teorema anterior $\mathcal{K}_n = [r^0, ..., r^{n-1}]$ ■

Teorema 5.6 El valor u^n obtenido en la $n-$ésima iteración del algoritmo de gradiente conjugado verifica

$$E(u^n) = \min_{P_{n-1} \in \mathcal{P}_{n-1}} (A(I - AP_{n-1}(A))e^0, (I - AP_{n-1}(A))e_0)$$

donde $\mathcal{P}_{n-1}$ es el espacio de polinomios de grado inferior o igual a $n-1$ y $e^0 = u^0 - u$.

Demostración: Todo $v = u^0 + \mathcal{K}_n$ de escribe $v = u^0 + P_{n-1}(A)r^0$ donde P_{n-1} es un polinomio de grado menor o igual que $n-1$.

Tenemos

$$v - u = e^0 + P_{n-1}(A)r^0 = e^0 - P_{n-1}(A)Ae^0 = (I - AP_{n-1}(A))e^0$$

Por la definición de la función $E(.)$ podemos escribir

$$E(v) = (A(v-u), v-u) = (A(I - AP_{n-1}(A))e^0, (I - AP_{n-1}(A))e^0)$$

Como $E(u^n) = \text{mín}_{v \in u^0 + \mathcal{K}_n} E(v)$ tendremos

$$E(u^n) = \min_{P_{n-1} \in \mathcal{P}_{n-1}} (A(I - AP_{n-1}(A))e^0, (I - AP_{n-1}(A))e^0)$$

■

Corolario 5.2 Se tiene la relación siguiente

$$E(u^n) \leq (\max_{1 \leq i \leq N} (1 - \lambda_i P_{n-1}(\lambda_i))^2))E(u^0)$$

para todo polinomio P_{n-1} de grado menor o igual que $n-1$ y donde λ_i para $i = 1, .., N$ son los valores propios de A.

Demostración: Siendo A una matriz simétrica definida positiva admite una base ortonormal de vectores propios $(v_1, ..., v_N)$ correspondientes a los valores propios $\lambda_1, ..., \lambda_N$.

En esta base $e^0 = u^0 - u$ se escribe $e^0 = \Sigma_{i=1}^N a_i v_i$ y

$$E(u^0) = (Ae^0, e^0) = \Sigma_{i=0}^N a_i^2 \lambda_i$$

además

$$(I - AP_{n-1}(A))e^0 = \Sigma_{i=0}^N a_i (1 - \lambda_i P_{n-1}(\lambda_i))v_i$$

De donde para todo polinomio de grado menor o igual que $n-1$, tenemos:

$$\begin{aligned} E(u^n) &\leq (A\Sigma_{i=1}^N a_i(1 - \lambda_i P_{n-1}(\lambda_i))v_i, \Sigma_{i=1}^N a_i(1 - \lambda_i P_{n-1}(\lambda_i))v_i) \\ &= \Sigma_{i=1}^N (1 - \lambda_i P_{n-1}(\lambda_i))^2 a_i^2 \lambda_i \\ &\leq [\max_i (1 - \lambda_i P_{n-1}(\lambda_i))^2][\Sigma_{i=1}^N a_i^2 \lambda_i] \\ &= [\max_i (1 - \lambda_i P_{n-1}(\lambda_i))^2]E(u^0) \end{aligned}$$

■

Corolario 5.3 El valor u^n obtenido en la n-ésima iteración del método de Gradiente Conjugado verifica

$$E(u^n) \leq 4(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1})^{2n} E(u^0)$$

Demostración: Mayoraremos $\max_i(1-\lambda_i P_{n-1}(\lambda_i))^2$ por $\max_{\lambda_1\leq\lambda\leq\lambda_N}(1-\lambda P_{n-1}(\lambda))^2$

$1-\lambda P_{n-1}(\lambda)$ es un polinomio de grado menor o igual que n que toma el valor 1 en $\lambda=0$.

Podemos entonces elegir

$$1-\lambda P_{n-1}(\lambda)=\frac{T_n(\frac{\lambda_N+\lambda_1-2\lambda}{\lambda_N-\lambda_1})}{T_n(\frac{\lambda_N+\lambda_1}{\lambda_N-\lambda_1})}$$

donde T_n es el polinomio de Chebyshev de grado n.

Se obtiene entonces la mayoración

$$\begin{aligned} E(u^n) &\leq \frac{1}{T_n^2(\frac{\lambda_N+\lambda_1}{\lambda_N-\lambda_1})}E(u^0) \\ &\leq 4(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1})^{2n}E(u^0) \end{aligned}$$

con $\kappa(A)=\frac{\lambda_N}{\lambda_1}$ ■

Ejercicios:

Sea A es simétrica y definida positiva de orden N. En los siguientes ejercicios consideramos la aplicación del método de gradiente conjugado para resolver un sistema de ecuaciones

$$Au=b$$

donde A es simétrica y definida positiva.

1. Dado $\epsilon>0$ estimar el número de iteraciones que hay que realizar con el método de gradiente conjugado para obtener la solución con un error

   $$E(u^n)=(A(u^n-u),u^n-u)\leq\epsilon(A(u^0-u),u^0-u)$$

   en función del número de condicionamiento de la matriz A.

2. Supongamos que el residuo inicial r^0 es un vector propio de la matriz A. Demostrar que el método de gradiente conjugado converge en una sola iteración.

3. Supongamos que la matriz A tiene todos los valores propios iguales. Demostrar que el método de gradiente conjugado converge en una sola iteración.

4. Sea $d^0=r^0=\Sigma_{i=1}^m\alpha_i v_i$ donde $\{v_i\}_{i=1}^m$ son m vectores propios ($m<N$). Demostrar que el método de gradiente conjugado converge en m iteraciones o menos.

5. Sea A simétrica y definida positiva con solo $m<N$ valores propios distintos. Demostrar que el método de gradiente conjugado converge en m iteraciones o menos.

6. Supongamos que los N valores propios de A están distribuidos de manera que los $N-m$ primeros valores propios están contenidos en un intervalo $[a,b]$, es decir,

$$0 < \lambda_1 \leq, ..., \leq \lambda_{N-m} \leq \lambda_{N-m+1} \leq, ..., \leq \lambda_N$$

$$\lambda_1, ..., \lambda_{N-m} \in [a,b]$$

Demostrar que en la iteración n-ésima

$$E(u^n) \leq 4 \left(\frac{\sqrt{\frac{b}{a}} - 1}{\sqrt{\frac{b}{a}} + 1} \right)^{2(n-m)} E(u^0)$$

7. Estimar el número de iteraciones n de Gradiente Conjugado necesarias para obtener una solución aproximada u^n verificando $E(u^n) \leq \varepsilon E(u^0)$ en el caso en que los valores propios de la matriz A esten distribuidos de la forma siguiente

$$\lambda_1, \lambda_2, ..., \lambda_{N-1} \in [a,b]$$

donde $0 < a < b \ll 10^6$ y $\lambda_N = 10^6$. Por ejemplo $a = 0.1$ y $b = 10^3$.

5.5. Precondicionamiento

5.5.1. Introducción

Como hemos visto, la rapidez de convergencia en los métodos de gradiente y de gradiente conjugado depende del número de condicionamiento de la matriz $\kappa(A)$ de la matriz A asociada al sistema de ecuaciones.

Cuánto más cercano a 1 sea este número más rápida es la convergencia.

La técnica de precondicionamiento consiste en remplazar la ecuación

$$Au = b$$

por una equivalente

$$C^{-1}Au = C^{-1}b$$

C^{-1} se elegir de modo que $\kappa(C^{-1}A)$ sea mucho más pequeño que $\kappa(A)$. En teoría la mejor elección es $C^{-1} = A^{-1}$ pues entonces $\kappa(C^{-1}A) = 1$. En la práctica habrá que encontrar C^{-1} lo más próximo posible a A^{-1} sin que el cálculo de C^{-1} sea demasiado costoso.

5.5.2. Algoritmo de gradiente conjugado precondicionado

En general no se aplica directamente el algoritmo al sistema $C^{-1}Au = C^{-1}b$ pues aunque C^{-1} sea una matriz simétrica no necesariamente lo será $C^{-1}A$.

Por ello se escribe el problema de la siguiente manera:

Si C^{-1} es simétrica y definida positiva se puede definir su raiz cuadrada $C^{-1/2}$ simétrica y definida positiva, es decir, $(C^{-1/2})^2 = C^{-1}$.

En lugar de considerar el sistema

$$C^{-1}Au = C^{-1}b$$

consideraremos , multiplicando por $C^{1/2}$ ambos lados,

$$C^{-1/2}AC^{-1/2}C^{1/2}u = C^{-1/2}b$$

e introduciendo una nueva variable $\tilde{u} = C^{1/2}u$, el sistema se escribe

$$\tilde{A}\tilde{u} = C^{-1/2}b$$

donde $\tilde{A} = C^{-1/2}AC^{-1/2}$. Aplicaremos el método de gradiente conjugado a este sistema de ecuaciones. Escribamos el algoritmo poniendo

- $\tilde{A} = C^{-1/2}AC^{-1/2}$
- $\tilde{u} = C^{1/2}u$, $\tilde{u}^n = C^{1/2}u^n$
- $\tilde{r}^n = C^{-1/2}b - \tilde{A}\tilde{u}^n = C^{-1/2}b - C^{-1/2}AC^{-1/2}C^{1/2}u^n = C^{-1/2}(b - Au^n) = C^{-1/2}r^n$
- $\tilde{d}^n = C^{1/2}d^n$

El algoritmo se escribe:

1. $\rho_n = \frac{||\tilde{r}^n||^2}{(\tilde{A}\tilde{d}^n, \tilde{d}^n)}$
2. $\tilde{u}^{n+1} = \tilde{u}^n + \rho_n\tilde{d}^n$
3. $\tilde{r}^{n+1} - \rho_n\tilde{A}\tilde{d}^n$
4. $\beta_{n+1} = \frac{||\tilde{r}^{n+1}||^2}{||\tilde{r}^n||^2}$
5. $\tilde{d}^{n+1} = \tilde{r}^{n+1} + \beta_{n+1}\tilde{d}^n$

Expresado en función de las variables originales será:

1. $\rho_n = \frac{(C^{-1}r^n, r^n)}{(Ad^n, d^n)}$
2. $u^{n+1} = u^n + \rho_n d^n$
3. $r^{n+1} = r^n - \rho_n A d^n$
4. $\beta_{n+1} = \frac{(C^{-1}r^{n+1}, r^{n+1})}{(C^{-1}r^n, r^n)}$
5. $d^{n+1} = C^{-1}r^{n+1} + \beta_{n+1} d^n$

La inversa de la matriz de precondicionamiento C no se calcula explícitamente. Pra ello se introduce una variable z^n y en lugar de calcular directamente $z^n = C^{-1}r^n$, se resuelve el sistema $Cz^n = r^n$.

Teniendo en cuenta esta última observación el algoritmo de gradiente conjugado precondicionado en su forma práctica será:

C , matriz de precondicionamiento.

1. u^0, $r^0 = b - Au^0$, $Cd^0 = r^0$, $z^0 = d^0$
2. Para $n = 0, 1, ...$
   - $\rho_n = \frac{(r^n, z^n)}{(Ad^n, d^n)}$
   - $u^{n+1} = u^n + \rho_n d^n$
   - $r^{n+1} = r^n - \rho_n A d^n$
   - $Cz^{n+1} = r^{n+1}$
   - $\beta_{n+1} = \frac{(r^{n+1}, z^{n+1})}{(r^n, z^n)}$
   - $d^{n+1} = z^{n+1} + \beta_{n+1} d^n$

Observemos que en cada iteración hay que resolver un sistema de ecuaciones asociado a la matriz C. En la práctica se elige C de manera que la resolución de este sistema sea mucho ms fácil que la resolución del sistema original. Este es el caso en que la matriz C es diagonal, o bien se dispone de la factorización de $C = RR^t$ en una matriz triangular superior y su correspondiente transpuesta.

Ejemplos de matrices de precondicionamiento son los siguientes:

- Precondicionador diagonal: Se elige como matriz de precondicionamiento la diagonal de la matriz A

- Precondicionadores basados en los métodos iterativos lineales (Jacobi, SSOR)
- Precondicionadores basados en la factorización incompleta de Cholesky: Se evita el llenado de la matriz manteniendo la estructura de huecos de la matriz original total o parcialmente.
- Precondicionadores multimalla: Generalmente están ligados al origen del problema, normalmente un problema de Ecuaciones en Derivadas Parciales, aunque también existen precondicionadores multimalla algebraicos.

Examinemos más detenidamente los precondicionadores asociados a los métodos iterativos lineales. Recordemos que un método iterativo lineal se basa en una descomposición de la matriz A del sistema a resolver de la forma $A = M - R$. La ecuación

$$Au = b$$

se escribe de la forma

$$Mu = Ru + b$$

lo que da lugar al método iterativo

$$Mu^{n+1} = Ru^n + b$$

o bien

$$u^{n+1} = M^{-1}Ru^n + M^{-1}b = Bu^n + c$$

donde $B = M^{-1}R$ y $c = M^{-1}b$. La condición de convergencia es $||B|| < 1$ para alguna norma matricial subordinada a una norma vectorial o de forma equivalente el radio espectral de B, $\rho(B) < 1$.

Observemos $A = M - R = M(I - M^{-1}R) = M(I - B)$. De modo que $A^{-1} = (I - B)^{-1}M^{-1}$. Como $||B|| < 1$, $(I - B)^{-1}$ existe y

$$(I - B)^{-1} = \sum_{n=0}^{\infty} B^n$$

Podemos tomar como matriz de precondicionamiento la siguiente aproximación de A^{-1}

$$C^{-1} = (\sum_{n=0}^{m-1} B^n)M^{-1}$$

Vamos a demostrar que resolver (paso 4 del algoritmo de gradiente conjugado)

$$Cz = r$$

es equivalente a realizar m iteraciones del método iterativo lineal para resolver $Az = r$ tomando como valor inicial $z^0 = 0$. En efecto,

- $z^0 = 0$

- La primera iteración es
$$Mz^1 = Rz^0 + r = r$$
es decir
$$z^1 = M^{-1}r$$
- La segunda iteración es
$$Mz^2 = Rz^1 + r = RM^{-1}r + r$$
o bien
$$z^2 = M^{-1}Rz^1 + M^{-1}r = M^{-1}RM^{-1}r + M^{-1}r = (I - M^{-1}R)M^{-1}r = (I + B)M^{-1}r$$
- En la iteración m-ésima
$$z^m = (I + B + ... + B^{m-1})M^{-1}r$$

Si nos limitamos a una sola iteración $C = M$, en el caso del método de Jacobi $M = D$ donde D es la parte diagonal de A y obtenemos el precondicionador diagonal.

Consideremos el método de Gauss-Seidel. En el método de Gauss-Seidel la descomposición $A = M - R$ es $M = D - F$ y $R = -E$ donde D es la parte diagonal, $-E$ es la parte trinagular inferior excluida la diagonal y $-F$ es la parte triangular superior excluida la diagonal. Si A es simétrica $F = E^t$.El método de Gauss-Seidel no es apropiado como precondicionador pues M no es simétrica. Por ello es conveniente considerar el método de Gauss-Seidel simétrico que consiste en realizar dos pasos alternativamente de la siguiente forma

- $$(D - E)u^{n+1/2} = Fu^n + b$$
- $$(D - F)u^{n+1} = Eu^{n+1/2} + b$$

de modo que

$$u^{n+1} = (D - F)^{-1}E(D - E^{-1}Fu^n + (D - F)^{-1}E(D - E)^{-1}b + (D - F)^{-1}b$$

Vamos a identificar la matriz M y la matriz R de la descomposición $A = M - R$ para este método. Necesitaremos el siguiente

Lema 5.5 Para un a matriz H tal que (I-H) sea no singular tenemos

$$H(I - H)^{-1} = (I - H)^{-1}H$$

Demostración: En efecto tenemos

$$H(I - H)^{-1} = (H^{-1})^{-1}(I - H)^{-1} = ((I - H)H^{-1})^{-1} = (H^{-1} - I)^{-1}$$

y también

$$(I-H)^{-1}H = (I-H)^{-1}(H^{-1})^{-1} = (H^{-1}(I-H))^{-1} = (H^{-1}-I)^{-1}$$

■

Propiedad 5.3 Sea la decomposición de A considerada anteriormente $A = D - F - E$. Tenemos

$$(D-F)^{-1}E(D-E)^{-1}F = (D-F)^{-1}D(D-E)^{-1}ED^{-1}F$$

En consecuencia el método iteractivo de Gauss-Seidel simétrico es de la forma

$$u^{n+1} = Bu^n + c$$

donde $B = M^{-1}R$ y $c = M^{-1}b$, con $M = (D-E)D^{-1}(D-F)$ y $R = ED^{-1}F$ o bien puesto que A es simétrica $M = (D-E)D^{-1}(D-E)^t$ y $R = ED^{-1}E^t$

Demostración:

En primer lugar

$$\begin{aligned} E(D-E)^{-1} &= ED^{-1}D(D-E)^{-1} = ED^{-1}(D^{-1)^{-1}}(D-E)^{-1} \\ &= ED^{-1}((D-E)D^{-1})^{-1} = ED^{-1}(I-ED^{-1})^{-1} = (I-ED^{-1})^{-1}ED^{-1} \\ &= DD^{-1}(I-ED^{-1})^{-1}ED^{-1} = D((I-ED^{-1})D)^{-1}ED^{-1} = D(D-E)^{-1}ED^{-1} \end{aligned}$$

donde hemos aplicado el lema anterior. Finalmente,

$$(D-F)^{-1}E(D-E)^{-1}F = (D-F)^{-1}D(D-E)^{-1}ED^{-1}F$$

■

Comentario 5.1 Observemos que también

$$c = (D-F)^{-1}E(D-E)^{-1}b + (D-F)^{-1}b = (D-F)^{-1}D(D-E)^{-1}b$$

En efecto,

$$\begin{aligned} E+D-E &= D \\ E(D-E)^{-1} + (D-E)(D-E)^{-1} &= D(D-E)^{-1} \\ E(D-E)^{-1} + I &= D(D-E)^{-1} \\ (D-F)^{-1}(E(D-E)^{-1}+I) &= (D-F)^{-1}D(D-E)^{-1} = M^{-1} \end{aligned}$$

■

5.6. Anexo: Polinomios de Chebyshev

Definición 5.3 Se llama polinomio de Chebyshev de grado n, al polinomio T_n definido por la relación de recurrencia

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_n(x) &= 2xT_{n-1}(x) - T_{n-2}(x) \quad \text{para} \quad n \geq 2 \end{aligned}$$

Propiedad 5.4 T_n viene dado por

$$\begin{aligned} T_n(x) &= \cos(n\ arc\cos x) \quad \text{para} \quad |x| \leq 1 \\ T_n(x) &= \cosh(n\ arg\cosh x) \quad \text{para} \quad x > 1 \\ T_n(x) &= (-1)^n T_n(-x) \quad \text{para} \quad x < -1 \end{aligned}$$

Demostración: Para $|x| \leq 1$, pongamos $x = \cos\theta$. Tenemos

$$T_n(x) = \cos n\theta = 2\cos\theta\cos(n-1)\theta - \cos(n-2)\theta$$

que se comprueba fácilmente utilizando el desarrollo del coseno de suma de dos ángulos.

Para $x > 1$ se comprueba análogamente utilizando las funciones hiperbólicas. ■

Ejemplos

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 \end{aligned}$$

Observamos que el coeficiente del término de grado n es 2^{n-1}, de modo que el polinomio $T_n(x)/2^{n-1}$ tiene el coeficiente del término de mayor grado igual a 1.

Propiedad 5.5 Para $n \geq 0$

$$T_n(x) = \frac{1}{2}\left((x + \sqrt{x^2-1})^n + (x - \sqrt{x^2-1})^n\right)$$

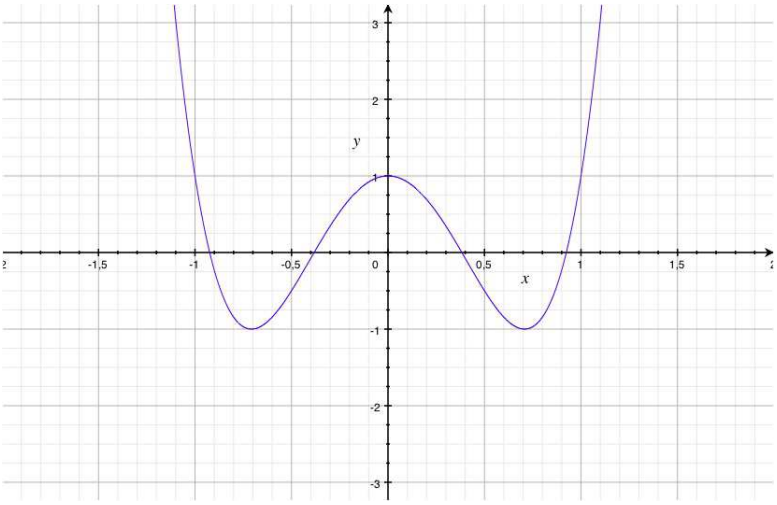

Figura 5.1: Polinomio de Chebyshev de grado 4

Demostración: Para $|x| \leq 1$, ponemos $x = \cos\theta$ y utilizamos

$$\cos n\theta = \frac{e^{in\theta} + e^{-in\theta}}{2}$$

junto con

$$e^{in\theta} = (\cos\theta + i\sin\theta)^n = (x - \sqrt{x^2-1})^n$$
$$e^{-in\theta} = (\cos\theta - i\sin\theta)^n = (x + \sqrt{x^2-1})^n$$

Para $x > 1$ ponemos $x = \cosh\theta$ y utilizamos

$$\cosh n\theta = \frac{e^{n\theta} + e^{-n\theta}}{2}$$

■

Propiedad 5.6 Para $b > a > 0$ tenemos

$$T_n(\frac{b+a}{b-a}) \geq \frac{1}{2}\left(\frac{\sqrt{\frac{b}{a}}+1}{\sqrt{\frac{b}{a}}-1}\right)^n$$

Demostración: Aplicamos la propiedad anterior y observamos

$$T_n(\frac{b+a}{b-a}) \geq \frac{1}{2}\left(\frac{b+a}{b-a}+\sqrt{\frac{(b+a)^2-(b-a)^2}{(b-a)^2}}\right)^n = \frac{1}{2}\left(\frac{\sqrt{\frac{b}{a}}+1}{\sqrt{\frac{b}{a}}-1}\right)^n$$

■

Propiedad 5.7 T_n tiene n raíces en $[-1,1]$, $x_k = \cos\frac{2k-1}{n}\frac{\pi}{2} \quad k = 1,...,n$.

Demostración:

Para que

$$T_n(x_k) = \cos n\theta_k = 0$$

es decir, como $x_k = \cos\theta_k$

$$n\theta_k = (2k-1)\frac{\pi}{2} \quad k = 1,...,n$$

y finalmente

$$x_k = \cos\frac{2k-1}{n}\frac{\pi}{2}$$

■

Propiedad 5.8 T_n tiene $n+1$ extremos relativos en $[-1,1]$, $x_k^{'} = \cos\frac{k\pi}{n} \quad k = 0,1,...,n$ para los cuales $T_n(x_k^{'}) = (-1)^k$

Demostración:

Para que $T_n(x_k^{'}) = (-1)^k$

$$x_k^{'} = \cos\frac{k\pi}{n} \quad k = 0,1,...,n$$

■

Teorema 5.7 Propiedad de optimalidad 1

Sea P_n el conjunto de polinomios de grado n cuyo coeficiente de x^n es 1, entonces el polinomio $\frac{T_n}{2^{n-1}}$ verifica

$$\max_{-1\leq x\leq 1}\frac{|T_n(x)|}{2^{n-1}} \leq \max_{-1\leq x\leq 1}|p(x)| \quad \forall p \in P_n$$

Demostración:

$\frac{T_n}{2^{n-1}}$ es un elemento de P_n que toma sus valores extremos $\frac{(-1)^k}{2^{n-1}}$, $n+1$ veces en los puntos $x_k^{'} = \cos\frac{k\pi}{n} \quad k = 0,1,...,n$.

Por reducción al absurdo supongamos que existe $p \in P_n$ tal que

$$\max_{-1\leq x\leq 1} |p(x)| < \frac{1}{2^{n-1}}$$

Sea $r = \frac{T_n}{2^{n-1}} - p$ que es un polinomio de grado menor o igual que $n-1$.

Entonces $r(x_k^{'}) = \frac{T_n(x_k^{'})}{2^{n-1}} - p(x_k^{'})$ tiene el mismo signo que $(-1)^k$ ya que $|p(x_k^{'})| < \frac{1}{2^{n-1}}$. r cambia de signo n veces en $[-1,1]$ y tiene por tanto al menos n ceros, y por lo tanto $r = 0$, al ser un polinomio de grado menor o igual que $n-1$. ■

Teorema 5.8 Propiedad de optimalidad 2

Sea F_n el conjunto de polinomios de grado menor o igual que n tal que $p(\alpha) = 1$ para $|\alpha| > 1$, entonces el polinomio $\frac{T_n}{T_n(\alpha)}$ verifica

$$\max_{-1\leq x\leq 1} |\frac{T_n(x)}{T_n(\alpha)}| \leq \max_{-1\leq x\leq 1} |p(x)| \quad \forall p \in F_n$$

Demostración:

$T_n(\alpha) \neq 0$ entonces $\frac{T_n}{T_n(\alpha)} \in F_n$ y

$$\max_{-1\leq x\leq 1} |\frac{T_n(x)}{T_n(\alpha)}| = \frac{1}{|T_n(\alpha)|}$$

Por reducción al absurdo supongamos que existe $p \in F_n$ tal que

$$\max_{-1\leq x\leq 1} |p(x)| < \frac{1}{|T_n(\alpha)|}$$

Entonces el polinomio $r = \frac{T_n}{T_n(\alpha)} - p$ es un polinomio de grado menor o igual que n que se anula para $x = \alpha$.

Además $T_n(\alpha)r(x_k^{'}) = T_n(x_k^{'}) - T_n(\alpha)p(x_k^{'})$ tiene el mismo signo que $(-1)^k$ para $k = 0, 1, ..., n$. Es decir, tiene al menos n raices en $[-1,1]$, por tanto r tiene al menos $n+1$ raices y es de grado menor o igual que n, en consecuencia $r = 0$. ■

Corolario 5.4 Sea G_n el conjunto de polinomios de grado menor o igual que n tal que $p(\alpha) = 1$ con $\alpha \notin [a,b]$, $0 < a < b$. Entonces el polinomio

$$q(x) = \frac{T_n(\frac{b+a-2x}{b-a})}{T_n(\frac{b+a-2\alpha}{b-a})}$$

verifica

$$\max_{a\leq x\leq b} |q(x)| \leq \max_{a\leq x\leq b} |p(x)| \quad \forall p \in G_n$$

y

$$\max_{a\leq x\leq b} |q(x)| = \frac{1}{|T_n(\frac{b+a-2\alpha}{b-a})|}$$

Demostración:

Hacemos el cambio de variable

$$\xi = \frac{b+a-2x}{b-a}$$

y aplicamos el teorema anterior. ■

Corolario 5.5 En particular para $\alpha = 0$, sea G_n^0 el conjunto de polinomios de grado menor o igual que n tal que $p(0) = 1$ y sea $0 < a < b$. Entonces el polinomio

$$q(x) = \frac{T_n(\frac{b+a-2x}{b-a})}{T_n(\frac{b+a}{b-a})}$$

verifica

$$\max_{a\leq x\leq b} |q(x)| \leq \max_{a\leq x\leq b} |p(x)| \quad \forall p \in G_n^0$$

■

Capítulo 6

Optimización de funciones cuadrátricas con restricciones

6.1. Planteamiento de un problema de optimización cuadrática con restricciones lineales

- $A \in \mathbb{R}^{d\times d}$ simétrica, definida positiva, es decir existe $\alpha > 0$ tal que $(Av, v) \geq \alpha||v||^2$ para todo $v \in \mathbb{R}^d$.
- Sea $B : \mathbb{R}^d \to \mathbb{R}^p$, con $p < d$ y designaremos también mediante $B \in \mathbb{R}^{p\times d}$ su correspondiente representación matricial en la base canónica. Nos
- $b \in \mathbb{R}^d$.
- $K = \{v \in \mathbb{R}^d; \quad Bv = 0\}$
- $J : \mathbb{R}^d \to \mathbb{R}$ definida por:

$$J(v) = \frac{1}{2}(Av, v) - (b, v)$$

Problema Primal, P

Consideramos el problema (P) siguiente: Hallar $u \in K$ tal que

$$J(u) = \inf_{v\in K} J(v)$$

Problema de punto silla asociado, L

Introducimos la lagrangiana

- $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$
- $\mathcal{L}(v, \mu) = J(v) + (\mu, Bv)$

y consideramos el problema asociado (L) siguiente: Hallar $(u, \lambda) \in \mathbb{R}^d \times \mathbb{R}^p$ tal que $\forall v \in \mathbb{R}^d$ y $\forall \mu \in \mathbb{R}^p$ verifica

$$\mathcal{L}(u, \mu) \leq \mathcal{L}(u, \lambda) \leq \mathcal{L}(v, \lambda)$$

Propiedad 6.1 Si $(u, \lambda) \in \mathbb{R}^d \times \mathbb{R}^p$ es solución de (L) entonces $u \in K$ y es solución de (P).

Demostración:

En efecto, si $(u, \lambda) \in \mathbb{R}^d \times \mathbb{R}^p$ verifica

$$\mathcal{L}(u, \mu) \leq \mathcal{L}(u, \lambda) \quad \forall \mu \in \mathbb{R}^p$$

entonces

$$(\mu - \lambda, Bu) \leq 0 \quad \forall \mu \in \mathbb{R}^p$$

tomando en el lugar de μ, $\mu + \lambda$ resulta

$$(\mu, Bu) \leq 0 \quad \forall \mu \in \mathbb{R}^p$$

y en esta última tomando en el luga r de μ, $-\mu$, resulta

$$(\mu, Bu) \geq 0 \quad \forall \mu \in \mathbb{R}^p$$

de donde

$$(\mu, Bu) = 0 \quad \forall \mu \in \mathbb{R}^p$$

es decir $Bu = 0$.

Por otra parte

$$J(u) + (\lambda, Bu) \leq J(v) + (\lambda, Bv) \quad \forall v \in \mathbb{R}^d$$

en particular tomando $v \in K$ y como $Bu = 0$

$$J(u) \leq J(v) \quad \forall v \in K$$

■

Propiedad 6.2 Si $(u, \lambda) \in \mathbb{R}^d \times \mathbb{R}^p$ es solución de (L), entonces (u, λ) está caracterizada por

$$Au + B^t\lambda = b \tag{6.1}$$
$$Bu = 0 \tag{6.2}$$

Demostración:

En efecto (u, λ) solución de (L) implica $Bu = 0$.

Por otra parte la función

$$\mathcal{J} : v \to J(v) + (\lambda, Bv) = J(v) + (B^t\lambda, v)$$

es convexa y diferenciable, el mínimo $u \in \mathbb{R}^d$ está caracterizado por

$$\nabla\mathcal{J}(u) = \nabla J(u) + B^t\lambda = Au - b + B^t\lambda = 0$$

Recíprocamente, si $Bu = 0$, en particular $(\mu - \lambda, Bu) \leq 0 \quad \forall\mu \in \mathbb{R}^p$ que es la primera desigualdad de (L). Por otra parte, la ecuación (6.1) se escribe $\nabla\mathcal{J}(u) = 0$ que es equivalente a la segunda desigualdad de (L) pues $\mathcal{J}$ es convexa.

■

Teorema 6.1 : Condición necesaria y suficiente de existencia y unicidad de solución del problema (L).

Sea A simétrica y definida positiva y B suprayectiva entonces entonces existe una única solución $(u, \lambda) \in \mathbb{R}^d \times \mathbb{R}^p$ del problema (L).

Demostración:

Tenemos $(ImB)^\perp = KerB^t$ de donde $\mathbb{R}^p = ImB \oplus KerB^t$. Entonces si B es suprayectiva resulta B^t es inyectiva, es decir,

$$||B^t\mu|| > 0 \quad \forall\mu \in \mathbb{R}^p,\ \mu \neq 0 \tag{6.3}$$

Hemos visto que el problema (L) es equivalente a resolver el sistema de ecuaciones (6.1)-(6.2). Eliminando la incógnita u del sistema obtenemos que λ es solución del problema

$$BA^{-1}B^t\lambda = BA^{-1}b \tag{6.4}$$

Este problema tiene solución única pues A es definida positiva, por tanto su inversa existe y es definida positiva, por tanto existe $\gamma > 0$ tal que $(A^{-1}v, v) \geq \gamma||v||^2$. De donde la matriz $BA^{-1}B^t$ es definida positiva pues para todo $\mu \in \mathbb{R}^p,\ \mu \neq 0$ tendremos

$$(BA^{-1}B^t\mu, \mu) = (A^{-1}B^t\mu, B^t\mu) \geq \gamma||B^t\mu||^2 > 0$$

Existe un único λ solución de (6.4). Finalmente obtenido λ, u es solución de

$$Au = b - B^t\lambda$$

que tiene solución única.

■

6.2. Algoritmos

6.2.1. Método de penalización

El método de penalización consiste en resolver un problema aproximado al (6.1)-(6.2). El problema penalizado es el siguiente:

$$Au^\varepsilon + B^t\lambda^\varepsilon = b \tag{6.5}$$

$$-\varepsilon\lambda^\varepsilon + Bu^\varepsilon = 0 \tag{6.6}$$

El problema penalizado (6.5)-(6.6) se resuelve en la práctica eliminando la variable λ^ε. Es decir, resolvemos

$$(A + \frac{1}{\varepsilon}B^tB)u^\varepsilon = b \tag{6.7}$$

Vamos a estudiar la convergencia del método.

Teorema 6.2 Bajo las hipótesis del teorema 6.1, tenemos que la solución de (6.5)-(6.6) verifica

$$||u - u^\varepsilon|| \leq C_1\varepsilon$$

$$||\lambda - \lambda^\varepsilon|| \leq C_2\varepsilon$$

donde (u, λ) es la solución de (6.1)-(6.2), $C_1 = \frac{||A||.||\lambda||}{\alpha\beta}$ y $C_2 = \frac{||A||^2.||\lambda||}{\alpha\beta^2}$

Demostración:

Restando las ecuaciones (6.1)-(6.5) y (6.2)-(6.6) obtenemos

$$A(u - u^\varepsilon) + B^t(\lambda - \lambda^\varepsilon) = 0 \tag{6.8}$$

$$\varepsilon\lambda^\varepsilon + B(u - u^\varepsilon) = 0 \tag{6.9}$$

de donde, utilizando que B^t es inyectiva

$$||\lambda - \lambda^\varepsilon|| \leq \frac{1}{\beta}||B^t(\lambda - \lambda^\varepsilon|| \leq \frac{||A||}{\beta}||u - u^\varepsilon||$$

por otra parte, como A es definida positiva y teniendo en cuenta (6.8) y (6.9) resulta

$$\begin{aligned}
\alpha||u-u^\varepsilon||^2 &\leq \big(A(u-u^\varepsilon), u-u^\varepsilon\big) = \big(\lambda^\varepsilon-\lambda, B(u-u^\varepsilon)\big)\\
\alpha||u-u^\varepsilon||^2 &\leq \varepsilon(\lambda^\varepsilon-\lambda, -\lambda^\varepsilon) = \varepsilon(\lambda^\varepsilon-\lambda, \lambda-\lambda^\varepsilon-\lambda)\\
\alpha||u-u^\varepsilon||^2 &\leq \varepsilon(\lambda^\varepsilon-\lambda, \lambda-\lambda^\varepsilon)+\varepsilon(\lambda, \lambda-\lambda^\varepsilon) \leq \varepsilon(\lambda, \lambda-\lambda^\varepsilon)\\
\alpha||u-u^\varepsilon||^2 &\leq \varepsilon||\lambda||.||\lambda-\lambda^\varepsilon|| \leq \varepsilon\frac{||A||}{\beta}||\lambda||.||u-u^\varepsilon||
\end{aligned}$$

finalmente

$$||u-u^\varepsilon|| \leq \frac{||A||.||\lambda||}{\alpha\beta}\varepsilon$$

y

$$||\lambda-\lambda^\varepsilon|| \leq \frac{||A||^2.||\lambda||}{\alpha\beta^2}\varepsilon$$

■

6.2.2. Algoritmo de Uzawa

Vamos a estudiar un primer algoritmo para resolver el problema (6.1)-(6.2). El algoritmo se puede interpretar como un algoritmo de Gradiente aplicado al problema dual (6.4).

Descripción del algoritmo de Uzawa

1. $\rho > 0$ elegido convenientemente.
2. $\lambda^0 \in \mathbb{R}^p$ arbitrario.
3. Para n=0,1,...
   - Obtenido λ^n , calculamos u^n solución de
   $$Au^n = b - B^t\lambda^n$$
   - Calculamos λ^{n+1} mediante
   $$\lambda^{n+1} = \lambda^n + \rho Bu^n$$

Convergencia del Algoritmo de Uzawa

Teorema 6.3 Eligiendo $0 < \rho < \frac{2\alpha}{||B||^2}$ en el algoritmo de Uzawa se tiene

$$\lim_{n\to\infty} ||u^n - u|| = 0$$

Si además B es suprayectiva entonces

$$\lim_{n\to\infty} ||\lambda^n - \lambda|| = 0$$

Demostración:

Designamos los errores en la n-ésima iteración mediante

$$\bar{u}^n = u^n - u$$
$$\bar{\lambda}^n = \lambda^n - \lambda$$

Tendremos, restando las correspondientes ecuaciones

$$A\bar{u}^n + B^t\bar{\lambda}^n = 0 \tag{6.10}$$
$$\bar{\lambda}^{n+1} = \bar{\lambda}^n + \rho B\bar{u}^n \tag{6.11}$$

En la ecuación (6.11) calculando el cuadrado de la norma

$$||\bar{\lambda}^{n+1}||^2 = ||\bar{\lambda}^n||^2 + 2\rho(\bar{\lambda}^n, B\bar{u}^n) + \rho^2||B\bar{u}^n||^2$$

reordenando

$$||\bar{\lambda}^n||^2 - ||\bar{\lambda}^{n+1}||^2 = -2\rho(\bar{\lambda}^n, B\bar{u}^n) - \rho^2||B\bar{u}^n||^2 \tag{6.12}$$

Por otra parte, multiplicando (6.10) escalarmente por $\bar{u}^n$

$$(A\bar{u}^n, \bar{u}^n) + (B^t\bar{\lambda}^n, \bar{u}^n) = 0$$

o bien

$$(A\bar{u}^n, \bar{u}^n) + (\bar{\lambda}^n, B\bar{u}^n) = 0$$

Como A es definida positiva, existe $\alpha > 0$ tal que $(Av, v) \geq \alpha||v||^2$ para todo $v \in \mathbb{R}^d$,

$$-(\bar{\lambda}^n, B\bar{u}^n) = (A\bar{u}^n, \bar{u}^n) \geq \alpha||\bar{u}^n||^2$$

de donde sustituyendo en (6.17)

$$||\bar{\lambda}^n||^2 - ||\bar{\lambda}^{n+1}||^2 \geq 2\rho\alpha||\bar{u}^n||^2 - \rho^2||B\bar{u}^n||^2$$

por otra parte

$$||B\bar{u}^n||^2 \leq ||B||^2||\bar{u}^n||^2$$

o bien

$$-||B\bar{u}^n||^2 \geq -||B||^2||\bar{u}^n||^2$$

de donde

$$||\bar{\lambda}^n||^2 - ||\bar{\lambda}^{n+1}||^2 \geq \rho(2\alpha - \rho||B||^2)||\bar{u}^n||^2$$

Eligiendo $0 < \rho < \frac{2\alpha}{||B||^2}$ resulta que la sucesión de números reales $(||\bar{\lambda}^n||^2)_{n=1}^{\infty}$ es decreciente y acotada inferiormente, por tanto es convergente y en particular de Cauchy. Finalmente

$$0 = \lim_{n\to\infty}(||\bar{\lambda}^n||^2 - ||\bar{\lambda}^{n+1}||^2) \geq \rho(2\alpha - \rho||B||^2) \lim_{n\to\infty} ||\bar{u}^n||^2 \geq 0$$

de donde

$$\lim_{n\to\infty} ||\bar{u}^n|| = 0$$

Por otra parte, si B es suprayectiva, entonces

$$||B^t\mu|| > 0 \quad \forall \mu \neq 0$$

que en espacios de dimensión finita es equivalente a la existencia de $\beta > 0$ tal que

$$||B^t\mu|| \geq \beta||\mu|| \quad \forall \mu \in \mathbb{R}^p$$

Tendremos utilizando (6.10),

$$\lim_{n\to\infty} ||\bar{\lambda}^n|| \leq \frac{1}{\beta} \lim_{n\to\infty} ||B^t\bar{\lambda}^n|| = \frac{1}{\beta} \lim_{n\to\infty} ||A\bar{u}^n|| \leq \frac{1}{\beta}||A|| \lim_{n\to\infty} ||\bar{u}^n|| = 0$$

■

6.2.3. Algoritmo de Lagrangiano Aumentado

En primer lugar vamos a sustituir el Problema (6.1)-(6.2) por un problema equivalente. En efecto si $(u, \lambda) \in \mathbb{R}^d \times R^p$ es solución de (6.1)-(6.2) y por lo tanto del problema (L) es también solución del problema

$$\begin{aligned} Au + rB^tBu + B^t\lambda &= b \qquad (6.13) \\ Bu &= 0 \qquad (6.14) \end{aligned}$$

donde $r > 0$.

Observación: : La solución $(u, \lambda) \in \mathbb{R}^d \times R^p$ es un punto silla de la función llamada Lagrangiana Aumentada dada por

$$\mathcal{L}_r(v, \mu) = J(v) + (\mu, Bv) + \frac{r}{2}||Bu||^2$$

El algoritmo de Lagrangiano Aumentado es el Algoritmo de Uzawa aplicado al sistema anterior (6.13)-(6.14), es decir,

Descripción del algoritmo de Lagrangiano Aumentado

1. Dado $r > 0$ y $\rho > 0$ elegido convenientemente.
2. $\lambda^0 \in \mathbb{R}^p$ arbitrario.
3. Para n=0,1,...
   - Obtenido λ^n , calculamos u^n solución de
   $$(A + rB^tB)u^n = b - B^t\lambda^n$$
   - Calculamos λ^{n+1} mediante
   $$\lambda^{n+1} = \lambda^n + \rho Bu^n$$

Convergencia del Algoritmo de Lagrangiano Aumentado

Teorema 6.4 Para $0 < r$ y eligiendo $0 < \rho < 2r$ en el algoritmo de Lagrangiano Aumentado se tiene

$$\lim_{n\to\infty} ||u^n - u|| = 0$$

Si además B es suprayectiva entonces

$$\lim_{n\to\infty} ||\lambda^n - \lambda|| = 0$$

Demostración:

Designamos los errores en la n-ésima iteración mediante

$$\bar{u}^n = u^n - u$$
$$\bar{\lambda}^n = \lambda^n - \lambda$$

Tendremos, restando las correspondientes ecuaciones

$$A\bar{u}^n + rB^tB\bar{u}^n + B^t\bar{\lambda}^n = 0 \tag{6.15}$$
$$\bar{\lambda}^{n+1} = \bar{\lambda}^n + \rho B\bar{u}^n \tag{6.16}$$

En la ecuación (6.16) calculando el cuadrado de la norma

$$||\bar{\lambda}^{n+1}||^2 = ||\bar{\lambda}^n||^2 + 2\rho(\bar{\lambda}^n, B\bar{u}^n) + \rho^2||B\bar{u}^n||^2$$

reordenando

$$||\bar{\lambda}^n||^2 - ||\bar{\lambda}^{n+1}||^2 = -2\rho(\bar{\lambda}^n, B\bar{u}^n) - \rho^2||B\bar{u}^n||^2 \tag{6.17}$$

Por otra parte, multiplicando (6.15) escalarmente por $\bar{u}^n$

$$(A\bar{u}^n, \bar{u}^n) + r(B\bar{u}^n, B\bar{u}^n) + (\bar{\lambda}^n, B\bar{u}^n) = 0$$

de donde teniendo en cuenta el carácter definido positivo de A

$$-(\bar{\lambda}^n, B\bar{u}^n) \geq \alpha||\bar{u}^n||^2 + r||B\bar{u}^n||^2$$

de donde

$$||\bar{\lambda}^n||^2 - ||\bar{\lambda}^{n+1}||^2 \geq 2\rho\alpha||\bar{u}^n||^2 + \rho(2r-\rho)||B\bar{u}^n||^2$$

Eligiendo $0 < \rho < 2r$

$$||\bar{\lambda}^n||^2 - ||\bar{\lambda}^{n+1}||^2 \geq 2\rho\alpha||\bar{u}^n||^2$$

Razonado como en el teorma anterior resulta

$$\lim_{n\to\infty} ||\bar{u}^n|| = 0$$

Por otra parte, si B es suprayectiva, entonces

$$||B^t\mu|| > 0 \quad \forall \mu \neq 0$$

que en espacios de dimensión finita es equivalente a la existencia de $\beta > 0$ tal que

$$||B^t\mu|| \geq \beta||\mu|| \quad \forall \mu \in \mathbb{R}^p$$

Tendremos utilizando (6.15),

$$\lim_{n\to\infty} ||\bar{\lambda}^n|| \leq \frac{1}{\beta} \lim_{n\to\infty} ||B^t\bar{\lambda}^n|| = \frac{1}{\beta} \lim_{n\to\infty} ||(A + rB^tB)\bar{u}^n|| \leq \frac{1}{\beta}||(A + rB^tB)|| \lim_{n\to\infty} ||\bar{u}^n|| = 0$$

■

Observación:

Consideremos el caso en el que B no es necesariamente suprayectiva. En este caso en el problema (6.13)-(6.14) (y también de (6.1)-(6.2)) la solución λ no es única.

Tenemos en todo caso $\mathbb{R}^p = ImB \oplus KerB^t$. Por tanto si $(u, \lambda) \in \mathbb{R}^d \times \mathbb{R}^p$ es solución de (6.13)-(6.14) $(u, \lambda + \mu)$ con $\mu \in KerB^t$ es también solución de (6.13)-(6.14).

Observemos en todo caso que en los dos algoritmos la sucesión u^n converge hacia u.

Observación:

Si consideramos el problema (6.13)-(6.14) (o también de (6.1)-(6.2)) restringido a $\mathbb{R}^d \times ImB$ entonces la solución correspondiente $(u, \lambda) \in \mathbb{R}^d \times ImB$ es única.

Estudiaremos ahora la convergencia de la sucesión λ^n en el caso general sin unicidad. Para todo elemento $\mu \in \mathbb{R}^p$ podemos considerar la descomposición ortogonal única $\mu = \mu_1 + \mu_2$ donde $\mu_1 = P_1(\mu) \in ImB$ designa la proyección ortogonal de μ sobre ImB y $\mu_2 = P_2(\mu) \in KerB^t$ designa la proyección ortogonal de μ sobre $KerB^t$. Si λ es una solución entonces $\lambda_1 = P_1(\lambda)$ es la solución de norma mínima entre todas las soluciones y es la única solución del problema restringido a $\mathbb{R}^d \times ImB$

Teorema 6.5 Sea $(u, \lambda) \in \mathbb{R}^d \times \mathbb{R}^p$ una solución de (6.13)-(6.14) (resp. de (6.1)-(6.2)). Si $0 < \rho < 2r$ (resp. $0 < \rho < \frac{2\alpha}{||B||^2}$) la sucesión λ^n definida por el algoritmo de Lagrangiano Aumentado (resp. Algoritmo de Uzawa) converge hacia $\lambda_1 + \lambda_2^0$ donde $\lambda_1 = P_1(\lambda)$ y $\lambda_2^0 = P_2(\lambda^0)$.

Demostración:

La relación

$$\lambda^{n+1} = \lambda^n + \rho B u^n$$

implica

$$P_2(\lambda^{n+1}) = P_2(\lambda^n) + \rho P_2(Bu^n) = P_2(\lambda^n) = P_2(\lambda^0) \quad \forall n$$

Tenemos

$$B^t \bar{\lambda}^n = -(A + rB^t B)\bar{u}^n$$

Bajo la condición $\lim_{n\to\infty} \bar{u}^n = 0$ resulta $\lim_{n\to\infty} B^t \bar{\lambda}^n = 0$.

Por otra parte en ImB tendremos que existe $\beta > 0$

$$||B^t \mu|| \geq \beta ||\mu|| \quad \forall \mu \in ImB$$

pues para todo $\mu \in ImB$ tal que $B^t \mu = 0$ necesariamente $\mu = 0$ (pues $ImB \cap KerB^t = 0$) y estamos en espacios de dimensión finita.

Tendremos

$$||\bar{\lambda}_1^n|| \leq \frac{1}{\beta} ||B^t \bar{\lambda}_1^n|| = \frac{1}{\beta} ||B^t \bar{\lambda}^n|| = \frac{1}{\beta} ||(A + rB^t B)\bar{u}^n||$$

$$\lim_{n\to\infty} ||\bar{\lambda}_1^n|| \leq \frac{1}{\beta} (||A|| + r||B||^2) \lim_{n\to\infty} ||\bar{u}^n|| = 0$$

Finalmente como

$$\lambda^n = \lambda_1^n + \lambda_2^n = \lambda_1^n + \lambda_2^0$$

resulta

$$\lim_{n\to\infty} \lambda^n = \lim_{n\to\infty} \lambda_1^n + \lambda_2^0 = \lambda_1 + \lambda_2^0 = P_1(\lambda) + P_2(\lambda^0)$$

■

6.2.4. Algoritmo de Gradiente Conjugado

El punto de partida es el sistema aumentado (6.13)-(6.14). Designando $A_r = A + B^tB$ y eliminando la variable u

$$u = A_r^{-1}(b - B^t\lambda)$$

sustituyendo en (6.14)

$$0 = Bu = BA_r^{-1}(b - B^t\lambda)$$

de donde λ es solución de

$$BA_r^{-1}B^t\lambda = BA_r^{-1}b \tag{6.18}$$

es decir

$$\mathcal{A}\lambda = c \tag{6.19}$$

donde $\mathcal{A} = BA_r^{-1}B^t$ y $c = BA_r^{-1}b$. Aplicaremos el algoritmo de Gradiente Conjugado al sistema (6.19).

Descripción del algortimo de Gradiente Conjugado

1. $\lambda^0 \in \mathbb{R}^p$ arbitrario.
2. $r^0 = c - \mathcal{A}\lambda^0$, $d^0 = r^0$.
3. Para n=0,1,...
   - Obtenido r^n, d^n, calculamos
   - $\rho_n = \frac{||r^n||^2}{(\mathcal{A}d^n, d^n)}$
   - $\lambda^{n+1} = \lambda^n + \rho_n d^n$
   - $r^{n+1} = r^n - \rho_n \mathcal{A}d^n$
   - $\beta_{n+1} = \frac{||r^{n+1}||^2}{||r^n||^2}$
   - $d^{n+1} = r^{n+1} + \beta_{n+1}d^n$

A continuación vamos a escribir este algoritmo en función de las variables del problema de partida. Tendremos en primer lugar para la expresión del residuo:

$$r^n = c - \mathcal{A}\lambda^n = BA_r^{-1}b - BA_r^{-1}B^t\lambda^n = BA_r^{-1}(b - B^t\lambda^n) = Bu^n$$

donde hemos tenido en cuenta que u^n es solución de

$$A_r u^n + B^t\lambda^n = b$$

Por otra parte, para calcular $\mathcal{A}d^n$ pondremos

$$\mathcal{A}d^n = BA_r^{-1}B^t d^n$$

y denotando $z^n = A_r^{-1}B^t d^n$ o bien, $\mathcal{A}d^n = Bz^n$, de modo que para calcular $\mathcal{A}d^n$ resolvemos el sistema

$$A_r z^n = B^t d^n$$

y entonces

$$\mathcal{A}d^n = Bz^n$$

El algoritmo se escribe en fución de las variables del problema original de la siguiente manera

1. $u^0 \in \mathbb{R}^d$ arbitrario.
2. $r^0 = d^0 = Bu^0$.
3. Para n=0,1,...
   - Obtenido u^n, λ^n, d^n, realizamos los siguentes cálculos
   - Resolvemos $A_r z^n = B^t d^n$
   - $\rho_n = \frac{||Bu^n||^2}{(Bz^n, d^n)} = \frac{||Bu^n||^2}{(Bz^n, Bu^n)}$
   - $\lambda^{n+1} = \lambda^n + \rho_n d^n$
   - $u^{n+1} = u^n - \rho_n z^n$
   - $\beta_{n+1} = \frac{||Bu^{n+1}||^2}{||Bu^n||^2}$
   - $d^{n+1} = Bu^{n+1} + \beta_{n+1} d^n$

donde la expresión $u^{n+1} = u^n - \rho_n z^n$ se deduce de

$$\begin{aligned} A_r u^n + B^t \lambda^n &= b \\ A_r u^{n+1} + B^t \lambda^{n+1} &= b \end{aligned}$$

restando

$$A_r(u^{n+1} - u^n) + B^t(\lambda^{n+1} - \lambda^n) = 0$$

y finalmente

$$u^{n+1} - u^n = -A_r^{-1}B^t(\lambda^{n+1} - \lambda^n) = -A_r^{-1}B^t \rho_n d^n = -\rho_n A_r^{-1}B^t d^n = -\rho_n z^n$$

Ejercicios

1. Sea $B : \mathbb{R}^d \to \mathbb{R}^p$, con $p < d$.
   a) Demostrar que $(ImB)^\perp = KerB^t$.
   b) Demostrar que $\mathbb{R}^p = ImB \oplus KerB^t$.

c) Sea $B : \mathbb{R}^d \to \mathbb{R}^p$ suprayectiva. Demostrar que B^t es inyectiva y que existe $\beta > 0$ tal que

$$||B^t\mu|| \geq \beta||\mu|| \quad \forall \mu \in \mathbb{R}^p$$

2. Sea A una matriz de orden d simétrica y definida positiva con constante de elipticidad $\alpha > 0$, es decir

$$(Av, v) \geq \alpha||v||^2 \quad \forall v \in \mathbb{R}^d$$

Sea b un vector de $\mathbb{R}$

Sea B matriz de p filas $\times$ d columnas, con $p < d$.

Sea $\mathcal{B} = \{\mu = (\mu_1, ..., \mu_p) \in \mathbb{R}^p; \quad \mu_i \geq 0 \quad \forall i = 1, ..., p\}$

Sea $J : \mathbb{R}^d \to \mathbb{R}$ definida por $J(v) = \frac{1}{2}(Av, v) - (b, v)$

Sea $\mathcal{L} : \mathbb{R}^d \times \mathcal{B} \to \mathbb{R}$ definida por $\mathcal{L}(v, \mu) = J(v) + (\mu, Bv)$

Considerar el problema de punto silla siguiente:

Hallar $(u, \lambda) \in \mathbb{R}^d \times \mathcal{B}$ tal que

$$\mathcal{L}(u, \mu) \leq \mathcal{L}(u, \lambda) \leq \mathcal{L}(v, \lambda) \quad \forall (v, \mu) \in \mathbb{R}^d \times \mathcal{B} \tag{6.20}$$

a) Demostrar que si $(u, \lambda) \in \mathbb{R}^d \times \mathcal{B}$ es un punto silla de $\mathcal{L}$ entonces (u, λ) verifica

$$(\mu - \lambda, Bu) \leq 0 \ \forall \mu \in \mathcal{B} \tag{6.21}$$

$$(\lambda, Bu) = 0 \tag{6.22}$$

$$Au + B^t\lambda = b \tag{6.23}$$

b) Sea $(u, \lambda) \in \mathbb{R}^d \times \mathcal{B}$ verificando la inecuación (2) de la pregunta anterior. Demostrar que para $\rho > 0$ se verifica

$$\lambda = \Pi_{\mathcal{B}}(\lambda + \rho Bu) \tag{6.24}$$

donde $\Pi_{\mathcal{B}}$ designa la proyección ortogonal sobre el conjunto $\mathcal{B}$

c) Demostrar que $(\mu, Bu) \leq 0 \ \forall \mu \in \mathcal{B}$ y

que Bu verifica $(Bu)_i \leq 0 \quad \forall i = 1, ..., p$

d) Demostrar que si $(u, \lambda) \in \mathbb{R}^d \times \mathcal{B}$ es solución del problema de punto silla (6.20) entonces u es solución del problema

Hallar $u \in \{v \in \mathbb{R}^d; (Bv)_i \leq 0 \quad \forall i = 1, ..., p\}$ tal que

$$J(u) \leq J(v) \quad \forall v \in \{v \in \mathbb{R}^d; (Bv)_i \leq 0 \quad \forall i = 1, ..., p\}$$

e) Considerar el siguiente algoritmo de Uzawa para resolver el problema de punto silla (6.20):

1) $\lambda^0 \in \mathcal{B}$

2) Una vez obtenido λ^n, calculamos u^n y λ^{n+1} de la siguiente manera,

- u^n solución de

$$Au^n = b - B^t\lambda^n$$

-

$$\lambda^{n+1} = \Pi_{\mathcal{B}}(\lambda^n + \rho B u^n)$$

Demostrar

$$\lim_{n\to\infty} ||u^n - u|| = 0$$

para valores adecuados de $\rho > 0$.