UN ESTUDIO SOBRE EL MÉTODO DE CAUCHY CON BÚSQUEDA LINEAL DE WOLFE SOBRE VARIEDADES DE RIEMANN

POR

MALON RAFAEL MENDOZA

UNIVERSIDAD CENTROCCIDENTAL "LISANDRO ALVARADO"

DECANATO DE CIENCIAS Y TECNOLOGIA

BARQUISIMETO

UN ESTUDIO SOBRE EL MÉTODO DE CAUCHY CON BÚSQUEDA LINEAL DE WOLFE SOBRE VARIEDADES DE RIEMANN

POR

MALON RAFAEL MENDOZA

TRABAJO DE ASCENSO PRESENTADO PARA OPTAR A LA CATEGORIA TITULAR EN EL ESCALAFON DEL PERSONAL DOCENTE Y DE INVESTIGACION.

UNIVERSIDAD CENTROCCIDENTAL “LISANDRO ALVARADO”

DECANATO DE CIENCIAS Y TECNOLOGIA

Barquisimeto, Noviembre 2013

AGRADECIMIENTO

Mi obediencia y gratitud a la misma esencia inmanente en todo lo manifestado, el Ser Supremo. En consecuencia, a los que promovieron y ayudaron con este trabajo.... Om Tat Sat.

Resumen

El objetivo planteado en este trabajo es, primero, estudiar y analizar el algoritmo de optimización de la búsqueda lineal con condición de Wolfe y mayor descenso en $\mathbb{R}^n$, desarrollado por J. Nocedal y S. Wright [2], y segundo, dar detalles de los pasos de los requerimientos que W. Ring y B. Wirth [2] realizan, para llevar esta búsqueda lineal con condición de Wolfe a variedades Riemannianas.

Índice general

Introducción

En palabras de Absil, Mahony y Sepulchre [16], la mayoría de las técnicas numéricas disponibles para la optimización y las ecuaciones no lineales asumen como base un espacio euclidiano. Sin embargo, muchos problemas de cálculo se plantean en espacios no euclidianos. Varios autores [3, 4, 5, 6, 7] han propuestos algoritmos abstractos que explotan la geometría subyacente (por ejemplo, la simétrica, la homogénea, la Riemanniana) de las variedades sobre las que se proyectan los problemas, pero la conversión de estos algoritmos geométricos abstractos en procedimientos numéricos en situaciones prácticas, es a menudo una tarea no trivial que depende fundamentalmente de una adecuada representación de la variedad.

El objetivo planteado en este trabajo es, primero, estudiar y analizar el algoritmo de optimización de la búsqueda lineal con condición de Wolfe y mayor descenso en $\mathbb{R}^n$, desarrollado por J. Nocedal y S. Wright [2], y segundo, dar detalles de los pasos de los requerimientos que W. Ring y B. Wirth [2] realizan, para llevar esta búsqueda lineal con condición de Wolfe a variedades Riemannianas.

Hay un número de problemas que puede ser expresado como una minimización de una función $f : M \to \mathbb{R}$ sobre una variedad riemanniana M. Si la variedad M puede ser inmersa en un espacio de dimensiones superiores o si se define a través de una igualdad con limitaciones, entonces existe la posibilidad de emplear las herramientas muy avanzadas de optimización con restricciones (ver [2] para una introducción). A menudo, sin embargo, esta inmersión no está a la mano, y uno tiene que recurrir a métodos de optimización diseñados para variedades de Riemann. Aunque, si se obtuviera una inmersión,uno podría esperar, que un método de optimización Riemanniano se realizaría más eficientemente ya que aprovecharía la estructura geométrica de la variedad subyacente. Para ello, varios métodos se han ideado, desde el más simple, como el de descenso del gradiente en variedades [9], y hasta más sofisticado, como lo es el de *región de confianza* [8]. Previamente, se presentaron los primeros intentos de adaptar los métodos de optimización estándar para problemas en variedades por Gabay [3], quien presentó los algoritmos de los métodos de descenso más rápido, de Newton, y cuasi-Newton; indicando sus propiedades globales y locales de convergencia (sin embargo, sin dar detalles de los análisis para el caso cuasi-Newton). Udriste [7] también indica un algoritmo para el método de máximo descenso y para el método de Newton en variedades de Riemann; prueba convergencia (lineal) del primero bajo la suposición de la búsqueda lineal exacta. Muy recientemente, Yang tomó estos métodos y analizó la convergencia del método de máximo descenso y de Newton para el control del paso de Armijo [9].

En comparación con los métodos de búsqueda lineal estándar en los espacios vectoriales, en la variedad, como se ha dicho en el párrafo anterior, todo se aproxima a sustituir el paso lineal de la dirección de búsqueda por un paso a lo largo de una curva geodésica. Sin embargo, las geodésicas pueden ser difícil de obtener. En los enfoques alternativos, las geodésicas a menudo son reemplazadas por caminos más generales, sobre la base de las llamadas retracciones (una retracción R_x es una aplicación del espacio tangente, T_xM, a la variedad M en $x \in M$; es decir, $R_x : T_xM \to M$, ver definición (4.1.1)). Usando esta retracción, por ejemplo, la función objetivo $f : M \to \mathbb{R}$ es elevada a una función objetivo, $\widehat{f}_x = f \circ R_x$ sobre T_xM. Ya que T_xM es un espacio euclidiano, es posible definir un modelo cuadrático de $\widehat{f}_x$ y adaptar los métodos clásicos en $\mathbb{R}^n$ para calcular (en general, aproximadamente) un minimizador del modelo dentro de una región factible alrededor de $0_x \in T_xM$. Este minimizador es entonces regresado de T_xM a M mediante la retracción R_x. Por lo tanto, tenemos un punto que será un candidato para la nueva iteración, el cual se acepta o se rechaza en función de la calidad de coincidencias entre el modelo cuadrático y la función f misma.

Nótese, como se verá en el capitulo 4, que teóricamente es posible escoger como retracción global en una variedad de Riemann, a la aplicación exponencial $exp_x : T_xM \to M$, $v \to exp_x\, v$, definida como $exp_x\, v = \gamma(1)$, donde γ resuelve una ecuación diferencial ordinaria de segundo orden con condiciones iniciales $\gamma(0) = x$ y $\gamma'(0) = v$ (ver 3.7). Esto corresponde a una estrategia para utilizar métodos numéricos de optimización en varieda-

des de Riemann, cuando se calcula la exponencial sobre un vector director con el fin de obtener un nuevo iterado o dirección de movimiento.

Así que, el esquema de trabajo que seguiremos es el siguiente, en el capítulo 1, tendremos las nociones básicas requeridas; en el capítulo 2, veremos el métodos de optimización de búsqueda lineal con condición de Wolfe en $\mathbb{R}^n$; en el capítulo 3, las variedades y elementos de optimización sobre las mismas y, en el capítulo 4, el métodos de optimización de búsqueda lineal con condición de Wolfe sobre variedades, siguiendo el procedimiento estándar de espacios vectoriales de dimensión finita [1]; además, se da el análisis del método de máximo descenso básico presentando convergencia. Y finalizaremos, con una aplicación del método del Armijo al cociente de Rayleigh sobre la esfera unitaria [16].

Capítulo 1

Preliminares

1.1. Nociones Básicas: En $\mathbb{R}^n$

Funciones a valores reales: Sea $f : \mathbf{X} \to \mathbb{R}$, donde $\mathbf{X} \subset \mathbb{R}^n$ es abierto.

Definición 1.1.1 *f es* diferenciable *en $\widehat{x} \in \boldsymbol{X}$ si existe un vector $\nabla f(\widehat{x})$ (el gradiente de f por $\widehat{x}$) y una función $\theta_{\widehat{x}}(y) : \boldsymbol{X} \to \mathbb{R}$ con*

$$\lim_{y \to 0} \theta_{\widehat{x}}(y) = 0,$$

tal que para cada $x \in \boldsymbol{X}$

$$f(x) = f(\widehat{x}) + \nabla f(\widehat{x})^T (x - \widehat{x}) + \|x - \widehat{x}\| \theta_{\widehat{x}}(x - \widehat{x}).$$

f es diferenciable *en $\boldsymbol{X}$ si f es diferenciable para todo $\widehat{x} \in \boldsymbol{X}$. El vector gradiente es el vector de las derivadas parciales*

$$\nabla f(\widehat{x}) = \Big(\frac{\partial f(\widehat{x})}{\partial x_1}, ..., \frac{\partial f(\widehat{x})}{\partial x_n} \Big)^T.$$

La derivada direccional *de f por $\widehat{x}$ en la dirección* d *es*

$$\lim_{\lambda \to 0} \frac{f(\widehat{x} + \lambda \mathrm{d}) - f(\widehat{x})}{\lambda} = \nabla f(\widehat{x})^T \mathrm{d}.$$

Definición 1.1.2 *La función f es* dos veces diferenciable *en $\widehat{x} \in \boldsymbol{X}$ si existe un vector $\nabla f(\widehat{x})$ y una matriz simétrica $n \times n$, $H(\widehat{x})$ (el* Hessiano *de f por $\widehat{x}$) tal que para cada $x \in \boldsymbol{X}$*

$$f(x) = f(\widehat{x}) + \nabla f(\widehat{x})^T(x - \widehat{x}) + \frac{1}{2}(x - \widehat{x})^T H(\widehat{x})(x - \widehat{x}) + \|x - \widehat{x}\|^2 \alpha(\widehat{x}, x - \widehat{x}).$$

y con

$$\lim_{y \to 0} \alpha(\widehat{x}, y) = 0.$$

f es dos veces diferenciable *en $\boldsymbol{X}$ si f es dos veces diferenciable para todo $\widehat{x} \in \boldsymbol{X}$. El* Hessiano*, que denotamos por $H(x)$ es la matriz de las segundas derivadas*

$$[H(\widehat{x})]_{ij} = \frac{\partial^2 f(\widehat{x})}{\partial x_i \partial x_j},$$

y para funciones con segundas derivadas continuas, este será siempre simétrica

$$\frac{\partial^2 f(\widehat{x})}{\partial x_i \partial x_j} = \frac{\partial^2 f(\widehat{x})}{\partial x_j \partial x_i}$$

Funciones a valores vectoriales: Sea $f : \mathbf{X} \to \mathbb{R}^m$, donde $\mathbf{X} \subset \mathbb{R}^n$ es abierto.

$$f(x) = f(x_1, ..., x_n) = \begin{pmatrix} f_1(x_1, ..., x_n) \\ f_2(x_1, ..., x_n) \\ \vdots \\ f_m(x_1, ..., x_n) \end{pmatrix}$$

donde cada una de las funciones f_i es una función a valores reales.

Definición 1.1.3 *El* Jacobiano *de la función f en $\widehat{x}$ es la matriz cuya j-ésima fila es el gradiente de f_j por $\widehat{x}$ traspuesta. Más específicamente, el* Jacobiano *de f en $\widehat{x}$ es definido como $\nabla f(\widehat{x})^T$, donde $\nabla f(\widehat{x})$ es la matriz con entrada*

$$[\nabla f(\widehat{x})]_{ij} = \frac{\partial f_j(\widehat{x})}{\partial x_i},$$

Nótese que la j-ésima columna de $\nabla f(\widehat{x})$ es el gradiente de f_j en $\widehat{x}$

1.2. Nociones Básicas en Optimización

1.2.1 Tipos de problemas en Optimización

Algunas terminologías(ver [10]): La función $f(x)$ es la función objetivo. Las restricciones $h_i(x) = 0$ se refieren a las *restricciones de igualdad*; mientras que $g_i(x) \leqslant 0$ se refiere a las *restricciones de desigualdad*. Nótese que no se utiliza restricciones de la forma $g_i(x) < 0$.

Problema de Optimización sin restricciones

$$\begin{aligned} (P) \qquad & \min_{x} \qquad f(x) \\ & sujeto\ a \quad x \in \mathbf{X}, \end{aligned}$$

donde $x = (x_1, ..., x_n)^T \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$, y $\mathbf{X}$ es un abierto (usualmente $\mathbf{X} = \mathbb{R}^n$).

Problema de Optimización con restricciones

$$\begin{aligned} (P) \quad & \min_{x} \quad f(x) \\ & sujeto\ a \quad g_i(x) \leqslant 0 \quad i = 1, ..., m \\ & \qquad\qquad h_i(x) = 0 \quad i = 1, ..., l \\ & \qquad\qquad x \in \mathbf{X} \end{aligned}$$

Colocando $g(x) = (g_1(x), ..., g_m(x))^T : \mathbb{R}^n \to \mathbb{R}^m$, $h(x) = (h_1(x), ..., h_l(x))^T : \mathbb{R}^n \to \mathbb{R}^l$. Entonces (P) se puede reescribir como

$$\begin{aligned} (P) \quad & \min_{x} \quad f(x) \\ & sujeto\ a \quad g(x) \leqslant 0 \\ & \qquad\qquad h(x) = 0 \\ & \qquad\qquad x \in \mathbf{X} \end{aligned}$$

1.2.2 Restricciones y Regiones Factibles

Un punto x es *factible* para (P) si satisface todas las restricciones; en el caso de problemas sin restricciones, $x \in \mathbf{X}$. El conjunto de todos los puntos factibles forman *la región de* puntos factibles, o *conjunto factible* que denotaremos por $\mathbf{S}$.

Se dice que la restricción de desigualdad $g_i(x) \leqslant 0$ es *activa* en el punto factible $\widehat{x}$, si $g_i(\widehat{x}) = 0$, y *no activa* si $g_i(\widehat{x}) < 0$ (todas las restricciones de desigualdad se consideraran activa en cualquier punto factible).

1.2.3 Tipos de Solución Optima

Considere el problema de optimización general

$$(P) \quad \min_{x \in \mathbf{S}} \quad o \quad \max_{x \in \mathbf{S}} \quad f(x)$$

Recordemos que una ε-vecindad de $\widehat{x}$ o la bola centrada en $\widehat{x}$ de radio ε es:

$$B(\widehat{x}, \varepsilon) = N_\varepsilon(\widehat{x}) := \{x : \|x - \widehat{x}\| \leqslant \varepsilon\}.$$

Tendremos la siguiente definición local/global, minimizador y máximizador estricto/no estricto.

Definición 1.2.1 *En el problema de optimización* (P)*,*

- $x \in \boldsymbol{S}$ *es un* minimizador global *de* (P) *si* $f(x) \leqslant f(y)$ *para todo* $y \in \boldsymbol{S}$*.*
- $x \in \boldsymbol{S}$ *es un* minimizador global estricto *de* (P) *si* $f(x) < f(y)$ *para todo* $y \in \boldsymbol{S}$*, con* $y \neq x$*.*
- $x \in \boldsymbol{S}$ *es un* minimizador local *de* (P) *si existe* $\varepsilon > 0$ *tal que* $f(x) \leqslant f(y)$ *para todo* $y \in B(x, \varepsilon) \cap \boldsymbol{S}$*.*
- $x \in \boldsymbol{S}$ *es un* minimizador local estricto *de* (P) *si existe* $\varepsilon > 0$ *tal que* $f(x) < f(y)$ *para todo* $y \in B(x, \varepsilon) \cap \boldsymbol{S}$*,* $y \neq x$ *.*
- $x \in \boldsymbol{S}$ *es un* máximizador global estricto *de* (P) *si* $f(x) > f(y)$ *para todo* $y \in \boldsymbol{S}$*, con* $y \neq x$*.*

- $x \in \boldsymbol{S}$ *es un* máximizador global *de* (P) *si* $f(x) \geqslant f(y)$ *para todo* $y \in \boldsymbol{S}$.

- $x \in \boldsymbol{S}$ *es un* máximizador local *de* (P) *si existe* $\varepsilon > 0$ *tal que* $f(x) \geqslant f(y)$ *para todo* $y \in B(x, \varepsilon) \cap \boldsymbol{S}$.

- $x \in \boldsymbol{S}$ *es un* máximizador local estricto *de* (P) *si existe* $\varepsilon > 0$ *tal que* $f(x) > f(y)$ *para todo* $y \in B(x, \varepsilon) \cap \boldsymbol{S}$, *con* $y \neq x$.

1.2.4 Existencia de Solución para Problemas de Optimización

Nos interesamos en :

- la existencia de la solución óptima,

- caracterización de la solución óptima, y

- algoritmos para calcular la solución óptima.

Para el primer caso por ejemplo, considere el siguiente problema de optimización:

- $$(P) \qquad \min_{x} \quad \frac{1+x}{2x}$$
$$sujeto\ a \quad x \geqslant 1.$$
Aquí no existe solución óptima porque la región factible no es acotada.

- $$(P) \qquad \min_{x} \quad \frac{1}{x}$$
$$sujeto\ a \quad 1 \leqslant x < 2.$$
Aquí no existe solución óptima porque la región factible no es cerrada.

- $$(P) \qquad \min_{x} \qquad f(x)$$
$$sujeto\ a \quad 1 \leqslant x < 2,$$

donde $f(x) = \begin{cases} \frac{1}{x}, & x < 2 \\ 1, & x = 2 \end{cases}$

Aquí no existe solución óptima porque la función f no es continua.

Teorema 1.2.1 *(**Teorema de Weierstrass para sucesiones**) Sea $\{x_k\}$ una sucesión infinita de puntos en el conjunto compacto $\boldsymbol{S}$ .Entonces existe una subsucesión $\{x_{k_j}\}$ de $\{x_k\}$ que converge a un punto de $\boldsymbol{S}$.*

Teorema 1.2.2 *(**Teorema de Weierstrass para funciones**) Sea f una función continua sobre el conjunto compacto no vacío $\boldsymbol{S} \subset \mathbb{R}^n$. Entonces existe un punto en $\boldsymbol{S}$ que minimiza (maximiza) a f; es decir, f alcanza un mínimo absoluto (máximo absoluto) en $\boldsymbol{S}$.*

1.2.5 Condiciones de Optimalidad para Problemas Sin Restricciones

Las condiciones para las soluciones local y global de problemas de optimización son intuitivas, y usualmente imposible de chequear directamente. En consecuencia, mostraremos más adelante como verificar fácilmente, las condiciones que son tanto necesarias para que un punto sea un minimizador local (lo que nos ayudaría a identificar los candidatos a ser minimizadores), como suficientes (que nos permitiría confirmar que el punto considerado es un minimizador local), o, algunas veces ambas.

$$(P) \qquad min \qquad f(x)$$

$$sujeto\ a \quad x \in \mathbf{X},$$

donde $x = (x_1, ..., x_n)^T \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$, y $\mathbf{X}$ es un abierto (usualmente $\mathbf{X} = \mathbb{R}^n$).

1.2.6 Condiciones de Optimalidad: Necesarias y suficientes

Condición necesaria para optimalidad local: si " $\widehat{x}$ es un minimizador local de (P), then $\widehat{x}$ debe satisfacer..." Tal condición nos permite identificar todos los candidatos para la optimización local.

Teorema 1.2.3 *Suponga que f es diferenciable en $\widehat{x}$. Si existe un vector d tal que $\nabla f(\widehat{x})^T d < 0$, entonces para todo $\lambda > 0$ suficientemente pequeño, $f(\widehat{x} + \lambda d) < f(\widehat{x})$ (d se llama la dirección de descenso si satisface la última condición).*

Prueba: Tenemos que

$$f(\widehat{x} + \lambda d) = f(\widehat{x}) + \lambda \nabla f(\widehat{x})^T d + \lambda \|d\| \alpha(\widehat{x}; \lambda d),$$

donde $\alpha(\widehat{x}; \lambda d) \to 0$ cuando $\lambda \to 0$. Acomodando tenemos

$$\frac{f(\widehat{x} + \lambda d) - f(\widehat{x})}{\lambda} = \nabla f(\widehat{x})^T d + \|d\| \alpha(\widehat{x}; \lambda d).$$

Ya que $\nabla f(\widehat{x})^T d < 0$ y $\alpha(\widehat{x}; \lambda d) \to 0$ cuando $\lambda \to 0$, $f(\widehat{x} + \lambda d) - f(\widehat{x}) < 0$ para todo $\lambda > 0$ suficientemente pequeño.$\diamond$

Corolario 1.2.1 *Suponga que f es diferenciable en $\widehat{x}$. Si $\widehat{x}$ es un minimizador local, entonces $\nabla f(\widehat{x}) = 0$ (a tal punto se le llama* estacionario*).*

Prueba: Si $\nabla f(\widehat{x}) \neq 0$, entonces $d = -\nabla f(\widehat{x})$ es una dirección de descenso, por lo cual $\widehat{x}$ no puede ser un minimizador local.$\diamond$

El corolario arriba es una condición de optimalidad necesaria de primer orden para un problema de optimización sin restricciones. Sin embargo, un punto estacionario puede ser un minimizador local, un maximizador local o ninguno de los dos. El siguiente teorema probará una condición de optimalidad necesaria de segundo orden. Primero, una definición:

Definición 1.2.2 *Una matriz M de orden $n \times n$ es llamada simétrica si $M_{ij} = M_{ji}$. Una matriz simétrica M de orden $n \times n$ es llamada*

- *definida positiva si* $x^T Mx > 0 \ \forall x \in \mathbb{R}^n, x \neq 0$
- *semidefinida positiva si* $x^T Mx \geqslant 0 \ \forall x \in \mathbb{R}^n$
- *definida negativa si* $x^T Mx < 0 \ \forall x \in \mathbb{R}^n, x \neq 0$
- *semidefinida negativa si* $x^T Mx \leqslant 0 \ \forall x \in \mathbb{R}^n$
- *indefinida si* $\exists\, x, y \in \mathbb{R}^n : x^T Mx > 0 \ \wedge \ y^T My < 0$

Decimos que M es SDP si M es simétrica y definida positiva. Similarmente, decimos que M es $SSDP$ si M es simétrica y semidefinida positiva.

Ejemplo 1

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

es definida positiva.

Ejemplo 2

$$M = \begin{pmatrix} 8 & -1 \\ -1 & 1 \end{pmatrix}$$

es definida positiva. Para ver esto, nótese que para $x \neq 0$,

$$x^T M x = 8x_1^2 - 2x_1x_2 + x_2^2 = 7x_1^2 + (x_1 - x_2)^2 > 0.$$

Ya que M es una matriz simétrica, todos sus autovalores son números reales. Se puede demostrar que M es $SSDP$ si y sólo sí todos sus autovalores son no negativos, definida positiva si todos sus autovalores son positivos, etc.

Teorema 1.2.4 *Suponga que f es dos veces diferenciable en $\widehat{x}$. Si $\widehat{x}$ es un minimizador local d entonces, $\nabla f(\widehat{x}) = 0$ y $H(\widehat{x})$ (el Hessiano en $\widehat{x}$) es semidefinido positivo.*

Prueba: De la condición necesaria de primer orden se tiene $\nabla f(\widehat{x}) = 0$. Supongamos que $H(\widehat{x})$ no es semidefinido positivo. Entonces $\exists\ d$ tal que $\nabla f(\widehat{x})d < 0$. Tenemos que

$$\begin{aligned} f(\widehat{x} + \lambda d) = &\quad f(\widehat{x}) + \lambda \nabla f(\widehat{x})^T d + \tfrac{1}{2}\lambda^2 d^T H(\widehat{x}) d + \lambda^2 \|d\|^2 \alpha(\widehat{x}; \lambda d) \\ = &\quad f(\widehat{x}) + \tfrac{1}{2}\lambda^2 d^T H(\widehat{x}) d + \lambda^2 \|d\|^2 \alpha(\widehat{x}; \lambda d) \end{aligned}$$

donde $\alpha(\widehat{x}; \lambda d) \to 0$ cuando $\lambda \to 0$. Acomodando tenemos

$$\frac{f(\widehat{x} + \lambda d) - f(\widehat{x})}{\lambda^2} = \frac{1}{2}\lambda^2 d^T H(\widehat{x}) d + \lambda^2 \|d\|^2 \alpha(\widehat{x}; \lambda d).$$

Si $d^T H(\widehat{x})d < 0$ y $\alpha(\widehat{x}; \lambda d) \to 0$ cuando $\lambda \to 0$, $f(\widehat{x} + \lambda d) - f(\widehat{x}) < 0$ para todo $\lambda > 0$ suficientemente pequeño -contradicción.$\diamond$

Ejemplo 1 *Sea*

$$f(x) = \frac{1}{2}x_1^2 + x_1x_2 + 2x_2^2 - 4x_1 - 4x_2 - x_2^3$$

Entonces

$$\nabla f(x) = (x_1 + x_2 - 4, x_1 + 4x_2 - 4 - 3x_2^2)^T,$$

y

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

$\nabla f(x) = 0$ *tiene exactamente dos soluciones:* $\widehat{x} = (4, 0)$ *y* $\widehat{x} = (3, 1)$.*Pero*

$$H(\widehat{x}) = \begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix}.$$

es indefinida, en consecuencia, el único candidato para un mínimo local es $\widehat{x} = (4, 0)$.

La condición necesaria solamente nos permiten obtener una lista de puntos candidatos a ser minimizadores. Ahora, la condición suficiente para optimalidad: "si $\widehat{x}$ satisface..., entonces $\widehat{x}$ es un minimizador local de (P)."

Teorema 1.2.5 *Suponga que* f *es dos veces diferenciable por* $\widehat{x}$ *. Si* $\nabla f(\widehat{x}) = 0$ *y* $H(\widehat{x})$ *es definido positivo, entonces* $\widehat{x}$ *es un minimizador local(estricto).*

Prueba: Tenemos que

$$f(x) = f(\widehat{x}) + \frac{1}{2}(x - \widehat{x})^T H(\widehat{x})(x - \widehat{x}) + \|(x - \widehat{x})\|^T \alpha(\widehat{x}; x - \widehat{x}).$$

Suponga que $\widehat{x}$ no es un minimizador local estricto. Entonces existe una sucesión $x_k \to \widehat{x}$ tal que $x_k \neq \widehat{x}$ y $f(x_k) \leqslant f(\widehat{x})$ para todo k.Defina $d_k = \dfrac{x_k - \widehat{x}}{\|(x - \widehat{x})\|}$. Entonces

$$f(x_k) = f(\widehat{x}) + \|(x_k - \widehat{x})\|^2 \left(\frac{1}{2} d_k^T H(\widehat{x}) d_k + \alpha(\widehat{x}; x_k - \widehat{x}) \right),$$

así

$$\frac{1}{2} d_k^T H(\widehat{x}) d_k + \alpha(\widehat{x}; x_k - \widehat{x}) = \frac{f(x_k) - f(\widehat{x})}{\|(x_k - \widehat{x})\|^2} \leqslant 0.$$

Ya que $\|d_k\| = 1$ para cualquier k, existe una subsucesión de d_k que converge hacia algún punto d tal que $\|d\| = 1$ (por el teorema 1.2.1). Sin perder generalidad suponga que $d_k \to d$. Entonces

$$\lim_{k \to \infty} \frac{1}{2} d_k^T H(\widehat{x}) d_k + \alpha(\widehat{x}; x_k - \widehat{x}) = \frac{1}{2} d^T H(\widehat{x}) d,$$

que contradice el hecho de que $H(\widehat{x})$ es definido positivo.$\diamond$

Nota:

- Si $\nabla f(x) = 0$ y $H(x)$ definido negativo, entonces $\widehat{x}$ es un maximizador local.
- Si $\nabla f(x) = 0$ y $H(x)$ semidefinido positivo, no se puede asegurar que $\widehat{x}$ es un minimizador local.

1.2.7 Convexidad y Minimización

Definición 1.2.3

- *Sean $x, y \in \mathbb{R}^n$, los puntos de la forma $\lambda x + (1 - \lambda) y$ para $\lambda \in [0, 1]$ son llamados* combinación convexa *de x con y . Más generalmente, el punto y es combinación convexa de los puntos $x_1, ..., x_k$ si $y = \sum_{i=1}^k \alpha_i x_i$ donde $\alpha_i \geqslant 0 \quad \forall i \quad y \quad \sum_{i=1}^k \alpha_i = 1$.*

- *Un subconjunto* $\boldsymbol{S} \subset \mathbb{R}^n$ *es llamado* convexo *si* $\forall x, y \in \boldsymbol{S}$ y $\lambda \in [0;1]$, $\lambda x + (1-\lambda)y \in \boldsymbol{S}$.

- *Una función* $f : \boldsymbol{S} \to \mathbb{R}^n$, *donde* $\boldsymbol{S}$ *es un conjunto convexo no vació es una función* convexa *si*

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda)f(y) \quad \forall x, y \in \boldsymbol{S}, \quad \forall \lambda \in [0;1].$$

- *Un función* f *como arriba es llamada* estrictamente convexa *si la desigualdad arriba es estricta para todo* $x \neq y$ y $\lambda \in (0,1)$.

- *Una función* $f : \boldsymbol{S} \to \mathbb{R}^n$ *es llamada* cóncava *(*estrictamente cóncava*) si es* convexa *(*estrictamente convexa*).*

Consideraremos el problema de optimización que escribiremos abajo y enunciaremos algunos teoremas cuyas demostraciones se pueden ver en [10]:

$$(CP) \qquad \min_{x} \qquad f(x)$$
$$sujeto\ a \quad x \in \mathbf{S}.$$

Teorema 1.2.6 *Suponga que* $\boldsymbol{S}$ *es un conjunto convexo no vació ,* $f : \boldsymbol{S} \to \mathbb{R}$ *es una función convexa, y* $\widehat{x}$ *es un minimizador local de* (CP). *Entonces* $\widehat{x}$ *es un minimizador global de* f *sobre* $\boldsymbol{S}$.

Prueba:(ver [10])

Nota:

- Un problema de minimizar una función convexa sobre una región factible convexa(tal como se ha considerado en el teorema) es un problema de programación convexa.
- Si f es estrictamente convexa, un minimizador local es el único minimizador global.
- Si f es (estrictamente) cóncava, un máximizador local es un (único) máximizador global.

El siguiente resultado nos ayuda a determinar cuando una función es convexa.

Teorema 1.2.7 *Suponga que $X \subset \mathbb{R}^n$ es un conjunto convexo abierto no vacío. Y que $f : X \to \mathbb{R}$ es diferenciable. Entonces f es convexa si y sólo sí ésta satisface la desigualdad del gradiente:*

$$f(y) \geqslant f(x) + \nabla f(x)^T (y - x) \quad \forall x, y \in X$$

Prueba:(ver [10])

En una dimensión, la desigualdad de gradiente tiene la forma

$$f(y) \geqslant f(x) + f'(x)(y - x) \quad \forall x, y \in X$$

El siguiente teorema proporciona otra condición suficiente y necesaria, para el caso cuando f es dos veces continuamente diferenciable.

Teorema 1.2.8 *Suponga que $X \subset \mathbb{R}^n$ es un conjunto convexo abierto no vacío, $f : X \to \mathbb{R}$ es dos veces diferenciable. Entonces f es convexa si el Hesiano de f, $H(x)$, es semidefinido positivo $\forall x \in X$.*

Prueba:(ver [10])

En una dimensión, el Hesiano es la segunda derivada de la función, la condición de semi-definido positivo se puede declara como $f'' \geqslant 0 \quad \forall x \in \mathsf{X}$.

Otro teorema relacionado al anterior (condición suficiente) es:

Teorema 1.2.9 *Suponga que $\mathsf{X} \subset \mathbb{R}^n$ es un conjunto convexo abierto no vacío, $f : \mathsf{X} \to \mathbb{R}$ es dos veces diferenciable. Entonces f es convexa si el Hesiano de f, $H(x)$, es definido positivo $\forall x \in \mathsf{X}$.*

Para problemas de optimización (sin restricciones) convexos, las condiciones de optimalidad de lo dicho en la sección anterior puede ser simplificada significativamente, proporcionando una única condición necesaria y suficiente para optimalidad global:

Teorema 1.2.10 *Suponga que $f : \mathsf{X} \to \mathbb{R}$ es convexa y diferenciable en X. Entonces $\widehat{x} \in \mathsf{X}$ es un minimizador global si y sólo si $\nabla f(\widehat{x}) = 0$.*

Prueba:

La necesidad de la condición $\nabla f(\widehat{x}) = 0$ fue establecida con independencia de la convexidad de la función.

Suponga que $\nabla f(\widehat{x}) = 0$. Entonces, por la desigualdad del gradiente

$$f(y) \geqslant f(\widehat{x}) + \nabla f(\widehat{x})^T (y - x) = f(\widehat{x})$$

para todo $y \in \mathsf{X}$, y así $\widehat{x}$ es un minimizador global.$\diamond$

Ejemplo 2 *Sea*

$$f(x) = -\ln(1 - x_1 - x_2) - \ln x_1 - \ln x_2.$$

Entonces

$$\nabla f(x) = \begin{pmatrix} \frac{1}{1-x_1-x_2} - \frac{1}{x_1} \\ \\ \frac{1}{1-x_1-x_2} - \frac{1}{x_2} \end{pmatrix},$$

y

$$H(x) = \begin{pmatrix} \left(\frac{1}{1-x_1-x_2}\right)^2 + \left(\frac{1}{x_1}\right)^2 & \left(\frac{1}{1-x_1-x_2}\right)^2 \\ \\ \left(\frac{1}{1-x_1-x_2}\right)^2 & \left(\frac{1}{1-x_1-x_2}\right)^2 + \left(\frac{1}{x_2}\right)^2 \end{pmatrix}$$

Es fácil probar que $f(x)$ es una función estrictamente convexa, y por lo tanto $H(x)$ es definido positivo en su dominio $X = \{(x_1, x_2) : x_1 > 0, x_2 > 0, x_1 + x_2 < 1\}$. En $\widehat{x} = (\frac{1}{3}, \frac{1}{3})$ tendremos que $\nabla f(\widehat{x}) = 0$, y así $\widehat{x}$ es el único minimizador local de f.

En el capitulo 2, se dará a conocer en detalle, el método de búsqueda lineal con condición de Wolfe en $\mathbb{R}^n$ y en el capítulo 4 en Variedades Riemannianas.

Capítulo 2

Método de Búsqueda Lineal con Condición de Wolfe en $\mathbb{R}^n$

2.1. Algoritmo de Optimización General

Recordemos: se quiere resolver el problema

$$(P) \qquad \min_{x} \qquad f(x)$$
$$sujeto\ a \quad x \in \mathbb{R}^n$$

Casi siempre, las soluciones de los problemas de optimización, son imposible de obtener directamente; salvo algunas excepciones. En consecuencia, en su mayor parte, se resolverán esos problemas con algoritmos iterativos. Estos algoritmos normalmente requieren que el usuario suministre un punto de partida x_0. Comenzando por x_0, un algoritmo iterativo generará una sucesión de puntos $\{x_k\}_{k=0}^{\infty}$ llamados *iterados*. Ahora bien, en la decisión de como generar el próximo iterado, x_{k+1}, el algoritmo utiliza la información de como se comporta la función f en el iterado actual, x_k, y algunas veces en los iterados pasados $x_0, ..., x_{k-1}$. En la práctica, más que construir una sucesión infinita de iterados, el algoritmo se detiene, cuando se satisface un criterio de finalización apropiado, indicando ya sea

que el problema ha sido resuelto dentro de una precisión deseada, o que no se puede hacer ningún progreso.

La discusión, en nuestro caso, caerá en la categoría de los *algoritmos de búsqueda lineal*:

Algoritmo de Optimización de Búsqueda Lineal General

Inicio Especifique un valor inicial de la solución x_0.

Iteración Para $k = 1, 2, ...$

Si x_k es el óptimo, parar.

De lo contrario,

- Determine d_k − una *dirección de búsqueda*
- Determine $\lambda_k > 0$ − *el tamaño del paso*
- Determine $x_{k+1} = x_k + \lambda_k d_k$ − un nuevo estimado de la solución

La elección de la dirección. Típicamente, se requiere que d_k sea una dirección de descenso de f por x_k, es decir,

$$f(x_k + \lambda d_k) < f(x_k) \quad \forall \lambda \in (0, \varepsilon]$$

para algún $\varepsilon > 0$. Para el caso cuando f es diferenciable, se demostró en el Teorema 1.2 que si $\nabla f(x_k) \neq 0$, cualquier d_k que cumpla con $\nabla f(x_k)^T d_k < 0$ es una dirección de descenso. Frecuentemente, la dirección es de la forma

$$d_k = -D_k \nabla f(x_k), \tag{2.1}$$

donde D_k es una matriz simétrica no singular.

Los siguientes son dos métodos básicos para escoger la matriz D_k por cada iteración; ello da como resultado dos algoritmos clásicos de optimización sin restricciones:

- *Descenso de mayor pendiente*: $D_k = I, k = 1, 2, ...$
- *Método de Newton* : $D_k = H(x_k)^{-1}$ (probado que $H(x_k) = \nabla^2 f(x_k)$ es definido positivo.)

Ahora si tenemos que D_k es una aproximación del Hessiano, estamos hablando del *método Cuasi-Newton* que se va actualizando en cada iteración mediante una fórmula de bajo-rango. Cuando d_k es definido en (2.1) y D_k es definida positiva, tendremos

$$\nabla f(x_k)^T d_k = -\nabla f(x_k)^T D_k \nabla f(x_k) < 0,$$

y en consecuencia d_k es una dirección de descenso.

La elección del Tamaño del Paso. En el cálculo de la longitud del paso, λ_k, nos encontramos ante una disyuntiva. Nos gustaría elegir λ_k para dar una reducción sustancial de f, pero al mismo tiempo, no queremos gastar demasiado tiempo en hacer la elección. Esto es, después de que d_k sea fijado, lo ideal sería que λ_k resolviera el problema de optimización unidimensional

$$\underset{\lambda_k \geqslant 0}{min} \quad f(x_k + \lambda d_k)$$

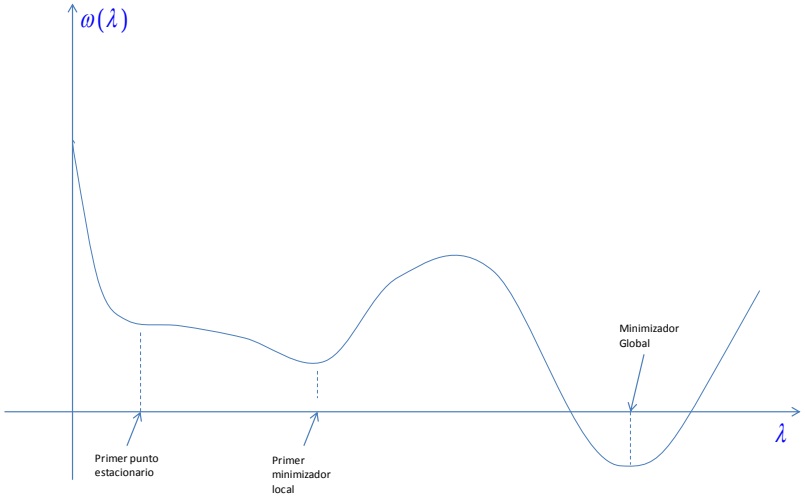

Figura 2.1: La longitud del paso ideal es el minimizador global.

Este problema de optimización suele ser también imposible de resolver exactamente. En su lugar, λ_k se calcula (a través de un procedimiento iterativo denominado búsqueda lineal), ya sea para resolver aproximadamente el anterior problema de optimización, o para asegurar una "suficiente"disminución en el valor de f. En detalle, la opción ideal sería el minimizador global de la función univariable $\omega(\cdot)$ definida por

$$\omega(\lambda) = f(x_k + \lambda d_k), \quad \lambda > 0, \tag{2.2}$$

pero en general, cuesta mucho identificar este valor (véase la Figura 2.1). Para encontrar incluso un minimizador local del ω con una precisión moderada, generalmente requiere demasiados evaluaciones de la función objetivo f y, posiblemente, del gradiente ∇f. Otras

estrategias prácticas realizan búsquedas lineales *inexactas* para identificar la longitud del paso que logran una adecuada reducción en f con un costo mínimo.

Típicamente el algoritmo de búsqueda lineal prueba una sucesión de valores candidatos para λ y, se detiene para aceptar uno de estos valores cuando se cumplen ciertas condiciones. La búsqueda lineal se realiza en dos fases: Una fase de horquillado o pruebas sucesivas para encontrar un intervalo que contiene longitudes de paso ideales, y una fase de bisección o interpolación que calcula una buena longitud de paso dentro de este intervalo.

Ahora discutiremos varias condiciones de finalización para el algoritmo de búsqueda lineal y mostrar que la longitud del paso efectiva no necesita yacer cerca del minimizador de la función real ω definida en (2.2).

Una simple condición que podríamos imponer a λ_k es que produzca una reducción en f, es decir, $f(x_k + \lambda_k d_k) < f(x_k)$. Que este requisito no es suficiente para producir convergencia hacia $\widehat{x}$ se ilustra en la Figura (2.2), en la cual el valor mínimo de la función es $\widehat{f} = -1$, pero una sucesión de *iterados* x_k para los cuales $f(x_k) = \frac{5}{k}, \quad k = 1, 2, ...$ admite un decrecimiento por cada iteración pero el valor del limite de función es cero. La dificultad es que no hay un *decrecimiento* suficiente de f en cada paso, pero este tema, lo discutiremos a continuación.

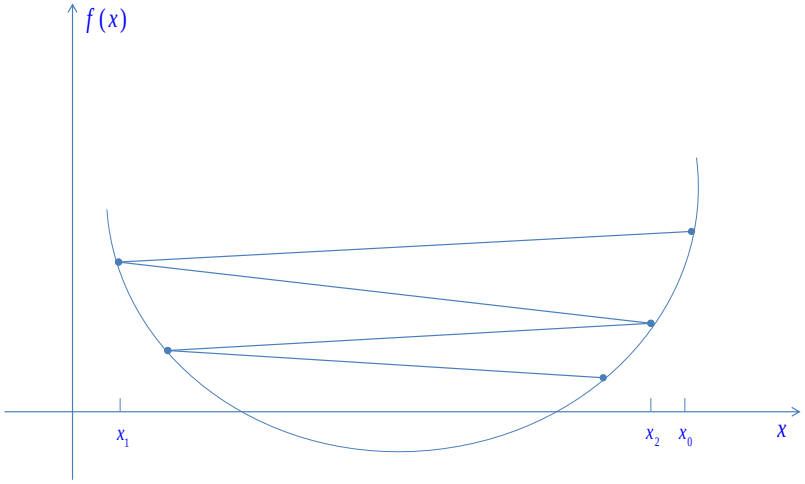

Figura 2.2: Insuficiente reducción en f.

2.2. La Condición de Wolfe

Una condición común de búsqueda lineal inexacta estipula que, λ_k debe en primer lugar dar una *disminución suficiente* a la función objetivo f, tal como se mide por la siguiente desigualdad:

$$f(x_k + \lambda d_k) < f(x_k) + c_1 \lambda \nabla f(x_k)^T d_k, \tag{2.3}$$

para alguna constante $c_1 \in (0, 1)$. En otras palabras, la reducción en f debe ser proporcional tanto a la longitud del paso λ_k y la derivada direccional $\nabla f(x_k)^T d_k$. La desigualdad (2.3) se llama la *condición de Armijo*. La regla de Armijo es uno de varios métodos de

búsqueda de lineal inexactas que garantiza un grado de precisión suficiente para asegurar la convergencia del algoritmo. La condición de disminución suficiente se ilustra en la figura 2.3. Aquí, $\omega(\lambda) = f(x_k + \lambda d_k)$ implica que $\omega'(\lambda) = \nabla f(x_k + \lambda d_k)^T d_k$ y, como d_k es una dirección de descenso, se tiene que $\omega'(0) < 0$. Entonces, la aproximación de primer orden de $\omega(\lambda)$ por $\lambda = 0$ es dada por $\omega(0) + \lambda\omega'(0)$. En consecuencia, el lado derecho de (2.3), es una función lineal, que la denotamos por $l(\lambda)$. La función $l(.)$ tiene pendiente negativa $c_1\nabla f(x_k)^T d_k$, pero a causa de que $c_1 \in (0, 1)$, su gráfica esta por arriba de la gráfica de ω para pequeños valores positivos de λ. Entonces la condición de disminución suficiente declara que λ es aceptable solo sí $\omega(\lambda) \leqslant l(\lambda)$. Los intervalos sobre los cuales esta condición se satisface se muestran en la figura 2.3. En la práctica, c_1 se escoge muy pequeño, digamos $c_1 = 10^{-4}$.

La condición de disminución suficiente no basta por sí misma para garantizar que el algoritmo haga un progreso razonable, ya que como vemos en la figura 2.3, ésta se satisface para valores suficientemente pequeños de λ. Para descartar avances cortos inaceptables introducimos un segundo requisito, que llamaremos *condición de curvatura*, la cual requiere que λ_k satisfaga la siguiente desigualdad

$$\nabla f(x_k + \lambda_k d_k)^T d_k \geqslant c_2 \nabla f(x_k)^T d_k, \tag{2.4}$$

para alguna constante $c_2 \in (c_1, 1)$, donde c_1 es la contante de (2.3). Nótese que el lado izquierdo es simplemente la derivada $\omega'(\lambda_k)$, por lo que la condición de curvatura asegura que la pendiente de ω en λ_k es mayor que c_2 veces la pendiente inicial $\omega'(0)$. Esto tiene

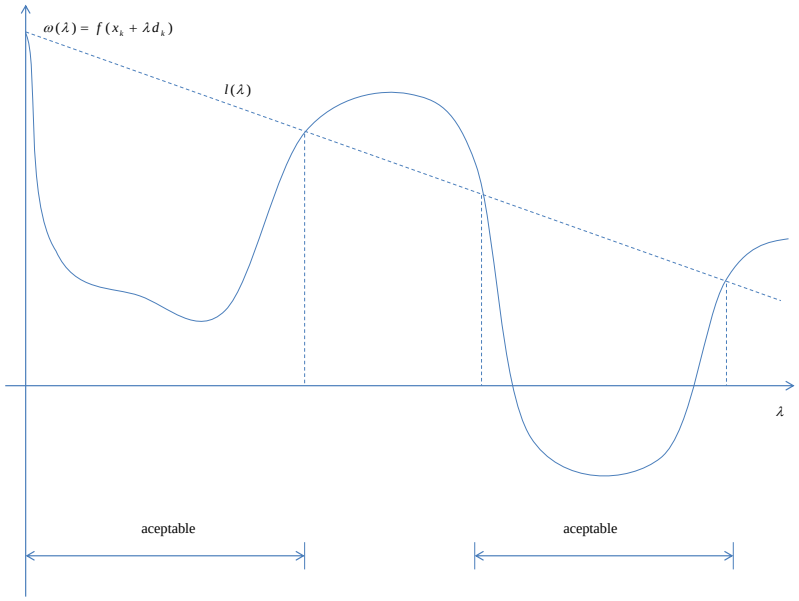

Figura 2.3: Condición de suficiente reducción.

sentido porque si la pendiente $\omega'(\lambda_k)$ es fuertemente negativa, tenemos una indicación que podemos reducir significativamente a f moviéndonos más allá de la dirección elegida. Por otro lado, si $\omega'(\lambda_k)$ es sólo ligeramente negativa, o incluso positiva, sería un signo que no podemos esperar más disminución de f en esta dirección, por lo que tendría sentido suspender la búsqueda lineal. La condición de curvatura se ilustra en la Figura 2.4. Los valores típicos de c_2 son 0,9 cuando se elige la dirección de búsqueda d_k por un método de Newton o cuasi-Newton, y $0,1$ cuando d_k se obtiene a partir de una método del gradiente conjugado no lineal.

Las condiciones de disminución suficiente y de curvatura se conocen colectivamente

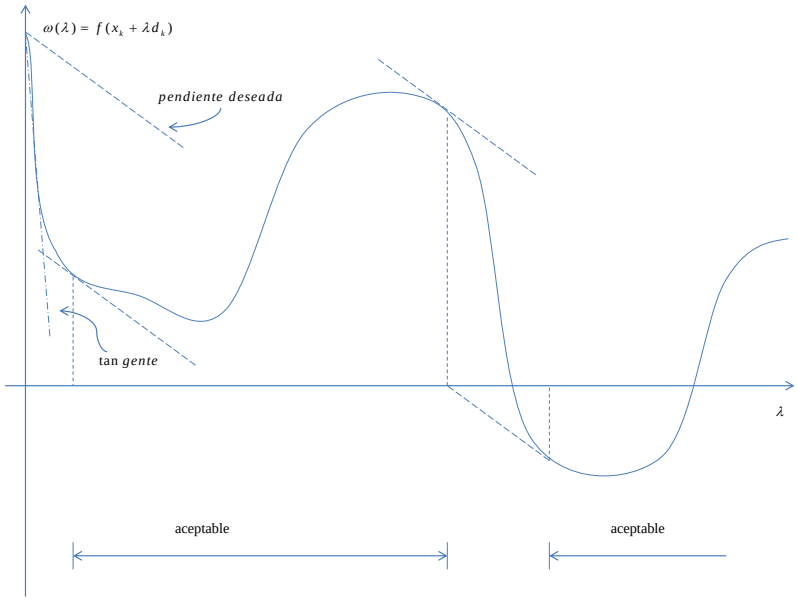

Figura 2.4: Condición de curvatura.

como *las condiciones de Wolfe*. Las ilustramos en la figura 3.5 y las reiteramos aquí para futuras referencias:

$$f(x_k + \lambda_k d_k) \leqslant f(x_k) + c_1 \lambda_k \nabla f(x_k)^T d_k, \tag{2.5a}$$

$$\nabla f(x_k + \lambda_k d_k)^T \lambda_k \geqslant c_2 \nabla f(x_k)^T d_k, \tag{2.5b}$$

con $0 < c_1 < c_2 < 1$. Una longitud de paso puede satisfacer las condiciones de Wolfe sin estar particularmente cerca de un minimizador de ω, como se muestra en la Figura 2.5. Podemos, sin embargo, modificar la condición de curvatura y forzar a λ_k a caer en al menos en una vecindad amplia de un minimizador local o de un punto estacionario de ω. También se tiene *las condiciónes fuerte de Wolfe*, que requiere que λ_k satisfaga las

siguientes desigualdades:

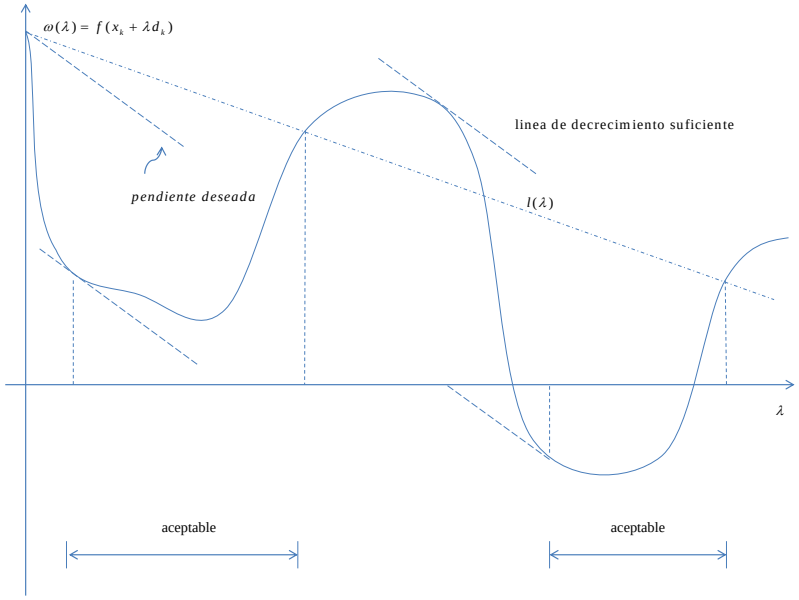

Figura 2.5: Longitudes de pasos que cumplen las condiciones de Wolfe .

$$f(x_k + \lambda_k d_k) \leqslant f(x_k) + c_1 \lambda_k \nabla f(x_k)^T d_k, \tag{2.6a}$$

$$\mid \nabla f(x_k + \lambda_k d_k)^T \lambda_k \mid \leqslant c_2 \mid \nabla f(x_k)^T d_k \mid, \tag{2.6b}$$

con $0 < c_1 < c_2 < 1$. La única diferencia con las condiciones de Wolfe es que ya no permitimos que la derivada $\omega'(\lambda_k)$ sea demasiado positiva. Por lo tanto, se excluye puntos que están lejos de los puntos estacionarios de ω.

No es difícil demostrar que existen longitudes de paso que satisfacen las condiciones

de Wolfe para cada función f que sea suave y acotada por debajo.

Lema 2.2.1 *Suponga que $f : \mathbb{R}^n \to \mathbb{R}$ es diferenciable continuamente. Sea d_k una dirección de descenso por x_k, con f acotada por abajo a lo largo del rayo $\{x_k + \lambda d_k / \lambda > 0\}$. Entonces si $0 < c_1 < c_2 < 1$, existen intervalos de longitud de paso que satisfacen las condiciones de Wolfe 2.5 y la condición fuerte de Wolfe 2.6*

Demostración 2.2.1 *Ya que $\omega(\lambda) = f(x_k + \lambda d_k)$ es acotada por abajo para toda $\lambda > 0$ y tenemos que $0 < c_1 < 1$, entonces la recta $l(\lambda) = f(x_k) + \lambda c_1 \nabla f(x_k)^T d_k$ debe interceptar el gráfico de ω por lo menos una vez. Sea $\lambda' > 0$ el más pequeño de estos valores, es decir,*

$$f(x_k + \lambda' d_k) = f(x_k) + \lambda' c_1 \nabla f(x_k)^T d_k. \tag{2.7}$$

La condición de disminución suficiente (2.5a) claramente se cumple para toda longitud de paso menor que λ'.

Por el teorema del valor medio, existe $\lambda'' \in (0, \lambda')$ tal que

$$f(x_k + \lambda' d_k) - f(x_k) = \lambda' \nabla f(x_k + \lambda'' d_k)^T d_k. \tag{2.8}$$

Por combinar (2.7) y (2.8), tenemos que

$$\nabla f(x_k + \lambda'' d_k)^T d_k = c_1 \nabla f(x_k)^T d_k \geqslant c_2 \nabla f(x_k)^T d_k, \tag{2.9}$$

ya que $c_1 < c_2$ y $\nabla f(x_k)^T d_k < 0$. Por tanto, λ'' satisface las condiciones de Wolfe 2.5, con desigualdad estricta en ambas ecuaciones (2.5a) y (2.5b) . Así, ya que f es suave, existe un intervalo que contiene a λ'' para el cual las condiciones de Wolfe se sostienen.

Más aún, ya que el término del lado izquierdo de (2.9) es negativo, las condiciones fuertes de Wolfe 2.6 se cumplen en el mismo intervalo.$\diamond$

Capítulo 3

Variedades

En este capítulo daremos a conocer el ámbito en donde extenderemos el método de búsqueda lineal con condición de Wolfe, las variedades diferenciables:

Definición 3.0.1 *Una variedad diferenciable de dimension n, es un conjunto M y una familia de aplicaciones inyectivas $\mathcal{X}_\alpha : U_\alpha \to M$, $\alpha \in I$, definidas en abiertos U_α de $\mathbb{R}^n$ en M tales que cumplen las siguientes condiciones:*

- $M = \bigcup_{\alpha \in I} \mathcal{X}_\alpha(U_\alpha)$
- *Para todo par $\mathcal{X}_\alpha$ y $\mathcal{X}_\beta$ con $\mathcal{X}_\alpha(U_\alpha) \cap \mathcal{X}_\beta(U_\beta) = W \neq \varnothing$, los conjuntos $\mathcal{X}_\alpha^{-1}(W)$ y $\mathcal{X}_\beta^{-1}(W)$ son abiertos en $\mathbb{R}^n$ y las aplicaciones $\mathcal{X}_\beta^{-1} \circ \mathcal{X}_\alpha : \mathcal{X}_\alpha^{-1}(W) \to \mathcal{X}_\beta^{-1}(W)$ son diferenciables.*

El par $(U_\alpha, \mathcal{X}_\alpha)$ con $p \in \mathcal{X}_\alpha(U_\alpha)$ es llamado una parametrización de M de p, $\mathcal{X}_\alpha(U_\alpha)$ es llamada vecindad coordenada de p . Una familia $\{(U_\alpha, \mathcal{X}_\alpha)\}$ satisfaciendo los items 1 y 2 es llamada estructura diferenciable de M.

Proposición 3.0.1 *Toda superficie regular de $\mathbb{R}^n$ de dimensión k es una variedad diferenciable de la misma dimensión.*

Prueba:ver [14]

Proposición 3.0.2 *Si M_1 y M_2 son dos variedades diferenciables de dimensión m_1 y m_2 respectivamente, entonces el producto cartesiano $M_1 \times M_2$ es una variedad de dimensión $m_1 + m_2$.*

Prueba: Ver [16]

3.1. Aplicaciones diferenciables entre variedades

Definición 3.1.1 *Sea $f : U \subset M \to \mathbb{R}$ una función definida en un subconjunto abierto U de una variedad diferenciable M. Diremos que f es diferenciable en $p \in U$, si para alguna parametrización $\mathcal{X}_\alpha : U_\alpha \subset \mathbb{R}^n \to M$, con $p \in \mathcal{X}_\alpha(U_\alpha) \subset U$, la composición $f \circ \mathcal{X}_\alpha : U_\alpha \subset \mathbb{R}^n \to \mathbb{R}$ es diferenciable en $\mathcal{X}_\alpha^{-1}(p)$. Se dice que f es diferenciable en U si es diferenciable en todo punto de U.*

Definición 3.1.2 *Una curva γ sobre una variedad diferenciable M es una aplicación $\gamma : I \to M$ donde $I = (-\varepsilon, \varepsilon)$. Diremos que γ es diferenciable en $t_0 \in I$ si para alguna parametrización $\mathcal{X}_\alpha : U_\alpha \subset \mathbb{R}^n \to M$ con con $\gamma(t_0) \in \mathcal{X}_\alpha(U_\alpha)$, la compuesta $\mathcal{X}_\alpha^{-1} \circ \gamma : I \to U_\alpha$ es diferenciable en t_0, donde $\gamma(I) \subset \mathcal{X}_\alpha(U_\alpha)$. Se dice que γ es diferenciable en I si es di-*

ferenciable en todo $t \in I$.

Definición 3.1.3 *Sean M_1 y M_2 variedades diferenciables de dimensión m y n respectivamente. Una aplicación $\varphi : M_1 \to M_2$ es diferenciable en $p \in V$, si dadas $\mathcal{X}_1 : U_1 \subset \mathbb{R}^n \to M_1$ parametrización de M_1 en p y $\mathcal{X}_2 : U_2 \subset \mathbb{R}^m \to M_2$ parametrización de M_2 en $\varphi(p)$, con $\varphi(\mathcal{X}_1(U_1)) \subset \mathcal{X}_2(U_2)$, la aplicación $\mathcal{X}_2^{-1} \circ \varphi \circ \mathcal{X}_1 : U_1 \subset \mathbb{R}^n \to \mathbb{R}^m$ es diferenciable en $\mathcal{X}^{-1}(p)$*

Definición 3.1.4 *Una aplicación $\varphi : M_1 \to M_2$ es diferenciable entre dos variedades diferenciables. Decimos que φ es difeomorfismo si φ es biyectiva y φ^{-1} es diferenciable.*

3.2. Espacio Tangente

Definición 3.2.1 *Sea M una variedad diferenciable. Consideremos una curva diferenciable $\gamma : (-\varepsilon, \varepsilon) \to M$, donde $\gamma(0) = p$ y sea D_p el conjunto de las funciones de M en $\mathbb{R}$ diferenciables en p, $D_p = \{f : M \to \mathbb{R} / f es\ diferenciable\ en\ p\}$. Se define el* ***Vector Tangente*** *a la curva γ en $t = 0$ como la función $\gamma'(0) : D_p \to \mathbb{R}$ dada por :*

$$\gamma'(0)f = \frac{d(f \circ \gamma)}{dt}\Big|_{t=0}, \ f \in D_p. \tag{3.1}$$

Definición 3.2.2 *(Espacio tangente). El espacio tangente a una variedad M en un punto p representado por T_pM, es el conjunto de todos los vectores tangentes a M en p. Así, $T_pM = \{v \in \mathbb{R}^m : v\ es\ un\ vector\ tangente\ en\ p\}$.*

Observación 3.2.1 *Si escogemos una parametrización $\mathcal{X}_\alpha : U \subset \mathbb{R}^n \to M$ con $p = \mathcal{X}_\alpha(0)$ y $q \in U$ podemos restringir la función f y la curva γ en esta parametrización por :*

$$f \circ \mathcal{X}(q) = f(q) = f(q_1, ..., q_n) \; y$$

$$\mathcal{X}^{-1} \circ \gamma(t) = (q_1(t), ..., q_n(t)),$$

restringiendo f a γ, tenemos

$$\begin{aligned} \gamma'(0)f = \frac{d}{dt}(f \circ \gamma)\Big|_{t=0} &= \frac{d}{dt} f(q_1(t), ..., q_2(t))\Big|_{t=0} \\ \sum_{i=1}^{n} q_i'(0)\Big(\frac{\partial f}{\partial q_i}(p)\Big) &= \Big(\sum_{i=1}^{n} q_i'(0)\Big(\frac{\partial}{\partial q_i}\Big)_0\Big) f. \end{aligned}$$

Así,

$$\gamma'(0) = \sum_{i=1}^{n} q_i'(0)\Big(\frac{\partial}{\partial q_i}\Big)_0 \tag{3.2}$$

es la expresión del vector tangente a f en p con relación a la parametrización $\mathcal{X}$.

Observación 3.2.2 *La expresión (3.2) muestra que el vector tangente a una curva γ en p depende de las derivadas, $q_i'(0)$, de γ en un sistema de coordenadas. De la elección de una parametrización obtenemos n vectores $\big\{\big(\frac{\partial}{\partial x_1}\big), ..., \big(\frac{\partial}{\partial x_n}\big)\big\}$ en T_pM que generan, por (3.2), los vectores en T_pM.*

Proposición 3.2.1 *El espacio tangente de una variedad diferenciable que es un subconjunto abierto de $\mathbb{R}^n$ es el propio $\mathbb{R}^n$.*

Prueba: Ver [14]

Definición 3.2.3 *Sean M_1 y M_2 dos variedades diferenciables de dimensión n y m respectivamente y sea $\varphi : M_1 \to M_2$ una aplicación diferenciable. Para cada $p \in M_1$ y cada $v \in T_pM_1$, escojamos una curva diferenciable $\alpha : (-\varepsilon, \varepsilon) \to M$, con $\alpha(0) = p$ y $\alpha'(0) = v$. Definiendo $\beta = \varphi \circ \alpha$ la aplicación:*

$$d\varphi_p : T_pM_1 \to T_{\varphi(p)}M_2,$$

dada por $d\varphi_p(v) = \beta'(0)$ es una aplicación lineal que no depende de la elección α. Esta aplicación es llamada la diferencial de φ en p.

3.3. Métricas Riemannianas

Definición 3.3.1 *Sea M una variedad diferenciable. Una métrica Riemanniana es una aplicación que asocia a cada $p \in M$ un producto interno (bilineal, simétrico y definido positivo) $\langle , \rangle_p$, dado por :*

$$\langle , \rangle_p = T_pM \times T_pM \to \mathbb{R},$$

que varía diferenciablemente en el siguiente sentido: si $\mathcal{X} : U \subset \mathbb{R}^n \to M$ es un sistema de coordenadas locales en torno de p, con $\mathcal{X}(x_1, x_2, x_3, ..., x_n) = q \in \mathcal{X}(U)$ y $\frac{\partial}{\partial x_i}(q) = d\mathcal{X}_q(0, 0, ..., 0, 1, 0, ..., 0, 0)$, entonces la función: $g_{ij} : U \to \mathbb{R}$ definida por:

$$g_{ij}(x_1, x_2, ..., x_n) = \left\langle \frac{\partial}{\partial x_i}(q), \frac{\partial}{\partial x_j}(q) \right\rangle_p ,$$

es una función diferenciable en U.

Las funciones g_{ij} son llamadas expresiones de la métrica riemanniana en el sistema coordenado $\mathcal{X}$ y la matriz $G = (g_{ij})$ es la representación de la métrica riemanniana.

Definición 3.3.2 *(Variedad riemanniana). Una variedad diferenciable para la cual se define una métrica riemanniana se denomina una variedad riemanniana.*

Definición 3.3.3 *El Producto interno de dos vectores $u, v \in T_pM$ es definido por $\langle u, v\rangle_p = g_x(u, v)$, donde g_x es la métrica riemanniana evaluada en el punto x. La norma de un vector $v \in T_pM$ es $\| v \|_p = \sqrt{\langle v, v\rangle_p}$.*

Definición 3.3.4 *Un grupo de Lie es una variedad diferenciable M con una estructura de grupo $*$, tal que la aplicación $M \times M \to M$ dada por $(x, y) \to x * y^{-1}$ es diferenciable. Para un elemento $y \in M$ la traslación a izquierda por x es el mapeo $L_x : M \to M$ definido por $L_x(y) = x * y$, el cual es un difeomorfismo.*

Definición 3.3.5 *Sea M un grupo de Lie. Una métrica riemanniana G en M es invariante por la izquierda si*

$$\langle u, v\rangle_y = \langle d(L_x)u, d(L_x)v, \rangle, \text{ para todo } x, y \in M \text{ y } u, v \in T_yM.$$

3.4. Subvariedades Encajadas (Embedding)

Un conjunto $\mathcal{X}$ puede admitir varias estructuras múltiples. Sin embargo, si el conjunto $\mathcal{X}$ es un subconjunto de una variedad $(M,\mathcal{U}_\alpha)$, entonces, esta admite a lo más una estructura de subvariedad.

3.4.1 Teoría General

Sean $(M,\mathcal{U}_\alpha)$ y $(N,\mathcal{V}_\alpha)$ variedades tal que $N \subset M$. La variedad $(N,\mathcal{V}_\alpha)$ es llamada una *subvariedad inmersa* de $(M,\mathcal{U}_\alpha)$ si la aplicación inclusión $i : N \to M, i(x) = x$ es una inmersión.

Sea $(N,\mathcal{V}_\alpha)$ una subvariedad de $(M,\mathcal{V}_\alpha)$. Ya que M y N son variedades, también son espacios topológicos con su variedad topológica. Si la variedad topológica de N coincide con su topología de subespacio inducida desde el espacio topológico M, entonces N se llama *subvariedad encajada o embedded*, o una *subvariedad regular*, o simplemente una subvariedad de la variedad M.

Proposición 3.4.1 *Sea N un subconjunto de un variedad M. Entonces N admite a lo más una estructura diferenciable que hace que sea una subvariedad encajada o embedded de M.*

Prueba: Ver [16]

A M en la proposición anterior se le llama el *espacio encajo o embedding*. Así por ejemplo, cuando el espacio incrustado $\mathbb{R}^{n+p}$ o un subconjunto abierto de $\mathbb{R}^{n+p}$, decimos que N es una *subvariedad matriz*

Proposición 3.4.2 *Un subconjunto N de una variedad M es una subvariedad encajada d-dimensional de M si y sólo si, alrededor cada punto $x \in N$, existe una carta $(\mathcal{U},\varphi)$ de M tal que*

$$N \cap \mathcal{U} = \{x \in \mathcal{U} : \varphi(d) \in \mathbb{R}^d \times \{0\}\}.$$

En este caso, la carta $\big(N \cap \mathcal{U}, \varphi\big)$*, donde* φ *es vista como una aplicación en* $\mathbb{R}^d$*, es una carta de la subvariedad encajada* N*.*

Prueba: Ver [16]

Las siguientes proposiciones proporcionan condiciones suficientes para que los subconjuntos de variedades puedan ser subvariedades encajadas o embedded, simplemente una subvariedad de la variedad M.(Para las pruebas ver [16])

Proposición 3.4.3 *(Teorema de submersión) Sea Sea* $F : M_1 \to M_2$ *una aplicación suave entre dos variedades de dimensión* d_1 *y* d_2*,* $d_1 > d_2$*, y sea* c *un punto de* M_2*. Si* c *es un valor regular (es decir, si el rango de* F *es igual a* d_2 *para cada punto de* $F^{-1}(c)$*) entonces* $F^{-1}(c)$ *es una subvariedad encajada cerrada de* M_1*, y* $dim\big(F^{-1}(c)\big) = d_1 - d_2$*.*

Proposición 3.4.4 *(Teorema de subinmersión) Sea Sea* $F : M_1 \to M_2$ *una aplicación suave entre dos variedades de dimensión* d_1 *y* d_2*, y sea* c *un punto de* $F(M_1)$*. Si* F *tiene rango constante igual a* $k < d_1$ *en una vecindad de* $\big(F^{-1}(c)\big)$*, entonces* $\big(F^{-1}(c)\big)$ *es una subvariedad encajada cerrada de* M_1 *de dimensión igual a* $d_1 - k$*.*

Las funciones en subvariedades embebidas no plantean ninguna dificultad en particular. Sea N una subvariedad encajada de una variedad M. Si f es una función suave en M, entonces $f|N$, la *restricción* de F sobre N, también es una función suave en N. Contrariamente, cualquier función suave en N se puede escribir localmente como una restricción de una función suave definida en un subconjunto abierto $\mathcal{U} \subset M$.

3.5. Campos de vectores, conexiones afines y derivada covariante

Definición 3.5.1 *(Campo de vectores en una variedad diferenciable). Un campo de vectores X en una variedad diferenciable M es una correspondencia que a cada punto $p \in M$ asocia un vector $X(p) \in T_pM$.*

Considerando una parametrización $\mathcal{X} : U \subset \mathbb{R}^n \to M$, es posible escribir:

$$X(p) = \sum_{i=1}^{n} a_i(p)(\frac{\partial}{\partial x_i})_p,$$

donde cada $a_i : M \to \mathbb{R}$ es una función en M y $\{(\frac{\partial}{\partial x_i})_p\}$ es una base asociada a $\mathcal{X}$, $1 \leqslant i \leqslant n$. Diremos que X es diferenciable si y solo si las funciones a_i son diferenciables.

Definición 3.5.2 *(Campo de vectores a lo largo de curvas). Un campo vectorial V a lo largo de una curva $\alpha : I \to M$ es una aplicación que a cada $\alpha(t) \in M$ asocia un vector tangente $V(t) \in T_{\alpha(t)}M$.*

Conexiones Afines.

Denotemos TM como el conjunto de espacios tangentes definidos en M. Sea $\mathcal{H} = \mathcal{H}(M) = \{X : M \to TM : \textit{para cada } p \in M,\ X(p) \in T_pM,\ \textit{y } X \in \mathcal{C}^\infty\}$ el conjunto de campo de vectores de clase $\mathcal{C}^\infty$ y $D = D(M) = \{f : M \to \mathbb{R} : f \in \mathcal{C}^\infty\}$ el conjunto de funciones reales de clase $\mathcal{C}^\infty$.

Definición 3.5.3 *Una conexión afín es una aplicación* $\nabla : \mathcal{H} \times \mathcal{H} \to \mathcal{H}$ *donde a cada par de campos* (X, Y) *se asocia otro campo* $\nabla_X Y$ *tal que para todo* $X, Y, Z \in \mathcal{H}$, *y* $f, g \in D$ *se cumple:*

i) $\nabla_{fX+gY} Z = f\nabla_X Z + g\nabla_Y Z$;

ii) $\nabla_X (Y + Z) = \nabla_X Y + \nabla_X Z$;

iii) $\nabla_X fY = f\nabla_X Y + X(f)Y$, *donde* $X(f) = \sum_{i=1}^{n} a_i(\cdot)\frac{\partial f(\cdot)}{\partial x_i}$

Proposición 3.5.1 *(Derivada Covariante) Sea M una variedad diferenciable con una conexión afín* ∇. *Entonces existe una única aplicación que asocia a un campo vectorial* V *a lo largo de una curva diferenciable* $\alpha : I \to M$ *otro campo vectorial* $\dfrac{DV}{dt}$ *a lo largo de* α, *denominado derivada covariante de* V *a lo largo de* α, *tal que para todo V,W campo de vectores a lo largo de* α *y* $f : I \to \mathbb{R}$ *una función diferenciable en I se cumple:*

i) $\dfrac{D}{dt}(V + W) = \dfrac{DV}{dt} + \dfrac{DW}{dt}$.

ii) $\dfrac{D}{dt}(fV) = \dfrac{df}{dt}V + f\dfrac{DV}{dt}$.

iii) *si* $V(t) = Y(\alpha(t))$, *donde* $Y \in \mathcal{H}$, *entonces* $\dfrac{DV}{dt} = \nabla_{\frac{d\alpha}{dt}} Y$.

Prueba: Ver [17]

Expresión de la conexión afín en términos de las coordenadas locales. Suponga que los campos de vectores $X, Y \in \mathcal{H}$ son representados en una cierta vecindad local $\mathcal{X} : U \subset \mathbb{R}^n \to M$ de un punto p, por:

$$X = \sum_{i=1}^{m} x_i X_i, Y = \sum_{i=1}^{m} y_i X_i,$$

donde $X_i = \dfrac{\partial}{\partial x_i}$ representan los vectores de la base del sistema de coordenadas locales. Usando las propiedades de la definición de la conexión afín:

$$\begin{aligned}\nabla_X Y &= \nabla_{\sum x_i X_i}\left[\sum_j y_j X_j\right] = \sum_i x_i\left[\nabla_{X_i}\left(\sum_j y_j X_j\right)\right] \\ &= \sum_i x_i\left[\sum_j (y_j \nabla_{X_i} X_j)\right] + \sum_i x_i\left[\sum_j \left(\frac{\partial y_j}{\partial x_i} X_j\right)\right],\end{aligned}$$

escribiendo $\nabla_{X_i} X_j$ en función de la base local:

$$\nabla_{X_i} X_j = \sum_{k=1}^{n} \Gamma_{ij}^k X_k, \tag{3.3}$$

sustituyendo en la ecuación anterior obtenemos:

$$\nabla_X Y = \sum_{k=1}^{n}\left(\sum_{i,j=1}^{n} x_i y_j \Gamma_{ij}^k + \sum_{i}^{n} x_i \frac{\partial y_k}{\partial x_i}\right) X_k.$$

Definición 3.5.4 *Los símbolos de Christoffel, o coeficientes de la conexión afín ∇ en U, son las funciones diferenciables $\Gamma_{ij}^k : U \subset M \to R$ definidas por (3.3)*

Expresión de la derivada covariante en términos de las coordenadas locales y de los símbolos de Christoffel.

Sea $\mathcal{X} : U \to M$ un sistema de coordenadas locales en torno de $p \in M$. Un resultado obtenido al demostrar la Proposición (3.5.1) es:

$$\frac{DV}{dt} = \sum_{j=1}^{n} \frac{dv^j}{dt} X_j + \sum_{i,j=1}^{n} v^j \frac{dx_i}{dt} \nabla_{X_i} X_j,$$

sutituyendo $\nabla_{X_i} X_j = \sum_{k=1}^{n} \Gamma_{ij}^k X_k$ en la ecuación anterior tenemos

$$\begin{aligned} \frac{DV}{dt} &= \sum_{j=1}^{n} \frac{dv^j}{dt} X_j + \sum_{i,j=1}^{n} v^j \frac{dx_i}{dt} \left(\sum_{k=1}^{n} \Gamma_{ij}^k X_k \right), \\ &= \sum_{j=1}^{n} \frac{dv^j}{dt} X_j + \sum_{k=1}^{n} \sum_{i,j=1}^{n} v^j \frac{dx_i}{dt} \Gamma_{ij}^k X_k \end{aligned}$$

y así,

$$\frac{DV}{dt} = \sum_{k=1}^{n} \left(\frac{dv^k}{dt} + \sum_{i,j=1}^{n} v^j \frac{dx_i}{dt} \Gamma_{ij}^k \right) X_k, \tag{3.4}$$

es la expresión de la derivada covariante en términos de coordenadas locales y de los símbolos de Christoffel.

Definición 3.5.5 *(Geodésicas) Una curva parametrizada $\alpha : I \to M$ es una geodésica si el campo tangente $\frac{d\alpha}{dt}$ verifica:*

$$\frac{D}{dt}\left(\frac{d\alpha}{dt}\right) = 0. \tag{3.5}$$

Definición 3.5.6 *(Campos paralelos) Dado M una variedad diferenciable, una conexión afín ∇ y un campo V a lo largo de una curva diferenciable $\alpha : I \to M$, V es denominado campo paralelo si*

$$\frac{DV}{dt} = 0, \ \forall t \in I. \tag{3.6}$$

Así, si α es una geodésica, entonces $\frac{d\alpha}{dt}$ es paralelo.

Proposición 3.5.2 *Sea M una variedad diferenciable con una conexión afín ∇. Sea $\alpha : I \to M$ una curva diferenciable en M y V_0 un vector tangente a M en $\alpha(t_o)$, $t_0 \in I$.*

Entonces existe un único campo de vectores paralelo V a lo largo de α, tal que $V(t_0) = V_0$ ($V(t)$ es llamado transporte paralelo de $V(t_0)$ a lo largo de α).

Prueba: Ver [17]

Ecuaciones Geodésicas

De la expresión (3.4), un campo paralelo V es determinado por las ecuaciones

$$\sum_{k=1}^{n}\left(\frac{dv^k}{dt}+\sum_{i,j=1}^{n}v^j\frac{dx_i}{dt}\Gamma_{ij}^k\right)X_k=0,$$

de forma equivalente,

$$\frac{dv^k}{dt}+\sum_{i,j=1}^{n}v^j\frac{d\alpha_i}{dt}\Gamma_{ij}^k=0,\ k=1,2,...n.$$

Cuando se trata de una geodésica $\alpha(t)=(\alpha_1(t),...,\alpha_n(t)(t))$, se tiene $v^i=\dfrac{d\alpha_i}{dt}$, entonces esta última ecuación se transforma en:

$$\frac{d}{dt}\Big(\frac{d\alpha_k}{dt}\Big)+\sum_{i,j=1}^{n}\frac{d\alpha_j}{dt}\frac{d\alpha_i}{dt}\Gamma_{ij}^k=0,\ k=1,2,...n,$$

así

$$\frac{d^2\alpha_k}{dt^2}+\sum_{i,j=1}^{n}\Gamma_{ij}^k\frac{d\alpha_j}{dt}\frac{d\alpha_i}{dt}=0,\ k=1,2,...n, \tag{3.7}$$

el cual es un sistema de n ecuaciones diferenciales de 2do. orden, que posee solución única en algún intervalo $I=[a,b]$, verificando $x(0)=\alpha(0)=p$ y $\dfrac{dx}{dt}(0)=\alpha'(0)=v$.

Conexión afín en variedades riemannianas

Definición 3.5.7 *Sea M una variedad diferenciable con una conexión afín ∇ y una métrica riemanniana $\langle,\rangle$. Se dice que ∇ es compatible con la métrica $\langle,\rangle$, si para todo par de*

campos de vectores V y W a lo largo de la curva diferenciable $\alpha : I \to M$ se tiene:

$$\frac{d}{dt}\langle V, W\rangle = \langle \frac{DV}{dt}, W\rangle + \langle V, \frac{DW}{dt}\rangle \tag{3.8}$$

Proposición 3.5.1 *Si la conexión afín ∇ es compatible con $\langle , \rangle$ y V, W son campos paralelos a lo largo de una curva diferenciable $\alpha : I \to M$ entonces, $\langle V, W\rangle$ es constante. En particular si $\alpha(t) = (\alpha_1(t), ..., \alpha_1(t))$ es una geodésica, $\langle \frac{d\alpha}{dt}, \frac{d\alpha}{dt}\rangle$ es constante.*

Prueba: Ver [17]

Definición 3.5.8 *Una conexión afín ∇ en una variedad riemanniana M es compatible con $\langle , \rangle$ si y solamente si:*

$$X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle, \ \forall\ X, Y, Z \in \mathcal{H}.$$

Definición 3.5.9 *Una conexión afín ∇ en una variedad riemanniana M es llamada simétrica si:*

$$\nabla_X Y - \nabla_Y X = [X, Y] \quad \forall\ X, Y, \in \mathcal{H},\ \textit{donde} [X, Y] = XY - YX.$$

Definición 3.5.10 *(Levi-Civita). Dada una variedad riemanniana M, existe una única conexión afín ∇ en M satisfaciendo las condiciones:*

a) ∇ *es simétrica.*

b) ∇ *es compatible con la métrica riemanniana.*

(Esta conexión es denominada conexión riemanniana).

Definición 3.5.11 *Dado un sistema de coordenadas $(U, \mathcal{X})$, las funciones conocidas como símbolos de Christoffel* $\Gamma_{ij}^k : U \subset M \to \mathbb{R}$ *que definen los coeficientes de conexión*

$$\nabla_{X_i} X_j = \sum_{k=1}^{n} \Gamma_{ij}^k X_k,$$

están relacionados con la métrica de la siguiente manera :

$$\Gamma_{ij}^m = \frac{1}{2} \sum_{k}^{n} \Big\{ \frac{\partial}{\partial x_i} g_{jk} + \frac{\partial}{\partial x_j} g_{ki} - \frac{\partial}{\partial x_k} g_{ij} \Big\} g^{km}, \tag{3.9}$$

donde $g_{ij} = \langle \frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j} \rangle$ *son elementos de la matriz* $G(x)$ *y* g^{ij} *los elementos de la inversa* $G(x)^{-1}$.

3.6. Curvatura de una variedad Riemanniana

Definición 3.6.1 *Una curvatura K de una variedad Riemanniana M es una correspondencia que asocia a cada par $X, Y \in \mathcal{H}$ una aplicación $K(X, Y) : \mathcal{H} \to \mathcal{H}$ dada por :*

$$K(X, Y)Z = \nabla_Y \nabla_X Z - \nabla_X \nabla_Y Z + \nabla_{[X,Y]} Z, \ Z \in \mathcal{H},$$

donde ∇ es una conexión de Riemanniana de M.

Observación 3.6.1 *Si consideramos un sistema de coordenadas $(U, \mathcal{X})$ en torno del punto p y $\{X_i\}$, $i = 1, 2, ..., n$ es una base de T_pM obtenemos:*

$$K(X_i, X_j)X_k = (\nabla_{X_j} \nabla_{X_i} - \nabla_{X_i} \nabla_{X_j}) X_k$$

Proposición 3.6.1 *La curvatura K de una variedad Riemanniana cumple las siguientes propiedades:*

i) *K es bilineal en* $\mathcal{H} \times \mathcal{H}$*, esto es,*

$$\begin{aligned} K(fX_1 + gX_2, Y_1) &= fK(X_1, Y_1) + gK(X_2, Y_1), \\ K(X_1, fY_1 + gY_2) &= fK(X_1, Y_1) + gK(X_1, Y_2), \end{aligned}$$

donde $f, g \in C^\infty(M)$*,* $X_1, X_2, Y_1, Y_2 \in \mathcal{H}$*.*

ii) *Para todo* $X, Y \in \mathcal{H}$ *el operador curvatura* $K(X, Y)$ *es lineal, esto es,*

$$\begin{aligned} K(X, Y)(Z + W) &= K(X, Y)Z + K(X, Y)W, \\ K(X, Y)fZ &= fK(X, Y)Z, \end{aligned}$$

donde $f \in C^\infty(M)$*,* $Z, W \in \mathcal{H}$*.*

iii) *La curvatura es antisimétrica, esto es ,*

$$K(X, Y) = -K(Y, X).$$

Prueba: Ver [17]

Proposición 3.6.2 *Sea* $(U, \mathcal{X})$ *un sistema de coordenadas en torno de* $p \in M$ *y* $\{X_i\}$ *una base de* T_pM *en este sistema de coordenadas. Entonces:*

$$K(X_i, X_j)X_k = \sum_{l=1}^{n} K^l_{ijk} X_l,$$

donde K^l_{ijk} *son dadas por :*

$$K^l_{ijk} = X_j \Gamma^l_{ik} - X_i \Gamma^l_{jk} + \sum_{s=1}^{n} \Gamma^s_{ik} \Gamma^l_{js} - \sum_{s=1}^{n} \Gamma^s_{jk} \Gamma^l_{is}.$$

Prueba: Ver [17]

Observación 3.6.2 *Si en las coordenadas* $(U, \mathcal{X})$ *escribimos* $X = \sum_{i=1}^{n} u^i X_i$, $Y = \sum_{j=1}^{n} v^j Y_j$, $Z = \sum_{k=1}^{n} w^k Z_k$, *por la linealidad de* K *tenemos:*

$$K(X,Y)Z = \sum_{i,j,k,l=1}^{n} K_{ijk}^{l} u^i v^j w^k X_l. \tag{3.10}$$

Curvatura Seccional. Íntimamente relacionado con el operador curvatura R está la curvatura seccional (o riemanniana) que definiremos a continuación.

Definición 3.6.2 *Dado un espacio vectorial* V, *indicaremos por* $|x \wedge y|$ *a la expresión*

$$\sqrt{|x|^2|y|^2 - \langle x, y\rangle^2},$$

que representa el área del paralelogramo bidimensional definido por un par de vectores $x, y \in V$

Definición 3.6.3 *Sea* $\sigma \subset T_P M$ *un subespacio bidimensional del espacio vectorial* $T_P M$ *y sean* $x, y \in \sigma$, *dos vectores linealmente independientes. Entonces,*

$$K(x,y) = \frac{(K(x,y)x, y)}{|x \wedge y|},$$

no depende de la elección de los vectores x *e* y.

Definición 3.6.4 *(Curvatura Seccional). Dado un punto* $p \in M$ *y un subespacio bidimensional* $\sigma \subset T_P M$ *. El número* $K(x,y) = K(\sigma)$, *donde* $\{x, y\}$*es una base de* σ, *es llamado Curvatura Seccional de* M.

Si $K(x,y) \leqslant 0$ para todo $x, y \in \sigma$ entonces, la curvatura seccional de la variedad riemanniana es no positiva.

Si $K(x,y) \geqslant 0$ para todo $x, y \in \sigma$ entonces, la curvatura seccional de la variedad riemanniana es no negativa.

Variedades completas

Definición 3.6.5 *Una variedad Riemanniana M es llamada (geodésicamente) completa si para todo $p \in M$, la aplicación exponencial, exp_p, están definidas para todos $v \in T_P M$.; es decir, las geodésicas que parten de p están definidas para todos los valores del parámetro $t \in \mathbb{R}$.*

Definición 3.6.6 *Dados dos puntos p y q en M, la distancia Riemanniana de p a q en la variedad, denotada por $d(p,q)$, es definida por*

$$dist(p,q) = \inf_{\substack{\gamma:[a,b]\to M \\ p=\gamma(a), q=\gamma(b)}} \int_a^b \| \gamma'(t) \| \, dt \tag{3.11}$$

Proposición 3.6.1 *Con la distancia geodésica (3.11) M es un espacio métrico.*

Teorema 3.6.1 *(Hopf-Rinow) Sea M una variedad Riemanniana y sea $p \in M$. Las siguientes afirmaciones son equivalentes:*

a) *exp_p está definida en todo $T_P M$ para algún $p \in M$*

b) *Los limitados y cerrados de M son compactos.*

c) *M es completo como espacios métrico.*

d) *M es geodésicamente completa.*

e) *Para todo $q \in M$ existe una geodésica uniendo p y q con $d(p,q) = inf_\gamma \int_a^b \parallel \gamma'(t) \parallel dt$.*

Prueba: Ver [18].

Teorema 3.6.2 *En una variedad Riemanniana completa M de dimensión finita con curvatura seccional no-negativa, tenemos:*

$$l_3^2 \leqslant l_1^2 + l_2^2 - 2l_1 l_2 \cos\theta, \tag{3.12}$$

donde l_i denota la longitud de α_i, $i = 1, 2$, $l_3 = d(\alpha_1(l_1),\ \alpha_2(l_2))$ y $\theta = \angle(\alpha_1'(0), \alpha_2'(0))$.

Prueba: Ver [18].

Definición 3.6.7 *Sea M una variedad Riemanniana y $f : M \to \mathbb{R}$, se define campo vectorial* ***grad(f)*** *(gradiente de la función f) como el único campo vectorial que cumple:*

$$df_p(X(p)) = \langle gradf(p), X(p)\rangle_x, \forall x \in \mathcal{H}. \tag{3.13}$$

Observación 3.6.3 *Sea $M \subset \mathbb{R}^n$ una variedad Riemanniana con la métrica definida por $\langle v, w\rangle_x = v^T G(x) w$ donde $G(x)$ es una matriz simétrica definida positiva. Se puede caracterizar el campo gradiente como:*

$$gradf(q) = G^{-1}(q) f'(q), \tag{3.14}$$

donde $G^{-1} = g^{ij}(q)$ *es la matriz inversa de* $G(q)$ *y* $f'(x) = (\frac{\partial f}{\partial x_1}, ..., \frac{\partial f}{\partial x_n})$ *es el vector de derivadas parciales de la función* $f \circ X$. *En efecto,*

$$df_p(X(p)) = f'(q)^T v = f'(q)^T (G^{-1}(q))^T G(q)v = (G^{-1}(q)f'(q))^T G(q)v = \langle G^{-1}(q)f'(q), v\rangle_q.$$

Definición 3.6.8 *Sea* M *una variedad Riemanniana y* $f : M \to \mathbb{R}$ *una función de clase* $C^k, k \geqslant 2$.*La Hessiana de f, denotada por* H^f *es definida como la derivada covariante del campo vectorial gradiente, esto es,*

$$H^f = \frac{D}{dt}\big(gradf\big).$$

La Hessiana en el punto p, en la dirección de $v \in T_pM$ *es:*

$$H_p^f(v, w) = \big\langle \frac{D_v}{dt}(gradf)(p), w\big\rangle_p. \tag{3.15}$$

3.7. Preliminares de optimización sobre Variedades

Definición 3.7.1 *Sea* M *una variedad Riemanniana y* $f : M \to \mathbb{R}$, *definimos por* ***Punto Critico****, los puntos donde :*

$$\langle gradf(x), v\rangle_x = 0, x \in M \text{ y } \forall v \in T_x(M). \tag{3.16}$$

Definición 3.7.2 *Sea* $f : M \to \mathbb{R},$, $x^* \in M$ *y una* $\mathbb{U}$ *vecindad abierta de* x^*, *entonces,* x^* *es un* ***Punto Mínimo Local de f****, si se cumple que* $f(x^*) \leqslant f(x)$ $\forall x \in \mathbb{U} \cap A$; *si la desigualdad se cumple para todo* $x \in M$, *entonces* x^* *es un* ***Punto Mínimo Global de f en*** A.

Definición 3.7.3 *Se dice que una función $f : M \to \mathbb{R}$ es semicontinua inferiormente en $\widehat{x} \in M$, si para toda sucesión $\{x^k\}$ de M convergente a $\widehat{x}$ se tiene que:*

$$\lim_{k\to\infty} inf f(x^k) \geqslant f(\widehat{x}).$$

Si f es semicontinua inferiormente para todo $x \in M$, entonces decimos que f es semicontinua inferiormente en M.

Teorema 3.7.1 *(Weierstrass) si $f : M \to \mathbb{R}$ es semicontinua inferiormente y M es compacto, entonces existe un punto mínimo global de f.*

Prueba: Ver [13].

Teorema 3.7.2 *(Condición necesaria de primer orden). Sea $f : M \to \mathbb{R}$ de clase C^1. Si x^* es un punto de mínimo local, entonces $gradf(x^*) = 0$.*

Demostración 3.7.1 *Sean $v \in T_{x^*}M$ y $\gamma : \mathbb{R} \to M$ una curva geodésica tal que $\gamma(0) = x^*$ y $\gamma'(0) = v$. Definamos la aplicación $h : \mathbb{R} \to \mathbb{R}$ tal que $h(t) = f(\gamma(t))$. Como x^* es punto de mínimo local de f, existe $\varepsilon > 0$ tal que $h(0) = f(x^*) \leqslant f(\gamma(0)) = h(t)$, para todo $t \in (-\varepsilon, \varepsilon)$, por lo que en $t = 0$ tenemos un punto de mínimo local de h. Por la condición necesaria de primer orden en $\mathbb{R}$ se tiene $h'(0) = \langle gradf(x^*), v\rangle = 0$. En particular para $v = gradf(x^*)$ tenemos que $gradf(x^*) = 0$.*

Teorema 3.7.3 *(Condición necesaria de segundo orden). Sea $f : M \to \mathbb{R}$ de clase C^2. Si x^* es un punto de mínimo local, entonces $\langle v, H^f_{x^*}v\rangle \geqslant 0$, $\forall\, v \in T_{x^*}M$.*

Demostración 3.7.2 *Sean $v \in T_{x^*}M$ y $\gamma : \mathbb{R} \to M$ una curva geodésica tal que $\gamma(0) = x^*$ y $\gamma'(0) = v$. Definamos la aplicación $h : \mathbb{R} \to \mathbb{R}$ tal que $h(t) = f(\gamma(t))$. De la demostración del teorema anterior, sabemos $t = 0$ hay un punto de mínimo local de h, entonces por la condición necesaria de segundo orden en $\mathbb{R}$ se tiene, $h'(0) = 0$, luego $h''(0) \geqslant 0$. Ahora bien,*

$$\begin{aligned} h'(t) &= \langle gradf(\gamma(t)), \gamma'(t)\rangle \\ h''(t) &= \left\langle \frac{D}{dt} gradf(\gamma(t)), \gamma'(t) \right\rangle + \left\langle gradf(\gamma(t)), \frac{D}{dt}\gamma'(t) \right\rangle \\ &= \langle H^f_{\gamma(t)}\gamma'(t), \gamma'(t)\rangle \\ &= \langle H^f_v v, v\rangle \geqslant 0. \end{aligned}$$

Teorema 3.7.4 *(Condición suficiente de segundo orden). Sea $f : M \to \mathbb{R}$ de clase C^2. Si $x^* \in M$ cumple:*

- *$gradf(x^*) = 0$,*
- *$H^f_{x^*}$ definida positiva,*

entonces, x^* es un punto de mínimo local estricto de f.

Prueba: Ver [13].

Definición 3.7.4 *Sea $A \subset M$ donde M es una variedad Riemanniana completa. Se dice que A es un* ***Conjunto Totalmente Convexo*** *si contiene toda geodésica γ de M que conecte a p y q, con $p, q \in A$.*

Definición 3.7.5 *Sea $A \subset M$ donde M es una variedad Riemanniana completa. Diremos que A es convexo si para todo par de puntos p y q de A existe una geodésica minimal que une p y q contenido en A.*

Definición 3.7.6 *Sea $f : M \to \mathbb{R}$; diremos que f es convexa si y solo si para toda geodésica $\gamma : \mathbb{R} \to \mathbb{M}$, la compuesta $f \circ \gamma : \mathbb{R} \to \mathbb{R}$ es convexa como una función real, es decir, $\forall a, b \in \mathbb{R}$, $\forall \lambda \in [0,1]$ se cumple:*

$$f(\gamma(1-\lambda)a + \lambda\, b) \leqslant (1-\lambda) f(\gamma(a)) + \lambda\, f(\gamma(b)). \tag{3.17}$$

Teorema 3.7.5 *Sea $f : M \to \mathbb{R}$ y M una variedad Riemanniana. Entonces f es convexa si y sólo si para todo $p \in M$ y para toda geodésica $\gamma : [o, +\infty) \to M$ tal que $\gamma(0) = p$ se cumple:*

$$f(\gamma(t)) - f(p) \geqslant t\langle gradf(p), \gamma'(0)\rangle \quad \text{con } t \in [o, +\infty). \tag{3.18}$$

Demostración 3.7.3 *Supongamos que f es convexa, entonces para toda geodésica $\gamma : \mathbb{R} \to M$; $f \circ \gamma$ es convexa, en particular para $h : [0, +\infty) \to \mathbb{R}$ dada por $h(t) = f(\gamma(t))$ con $t \in [0, +\infty)$.*

Como h es convexa como función real, entonces se cumple

$$h(t) \geqslant h(0) + th'(0). \tag{3.19}$$

Por otro lado

$$h'(t) = (f \circ \gamma)' = \langle gradf(\gamma(t)), \gamma'(t)\rangle,$$

evaluando h' *en* 0

$$\begin{aligned} h'(0) &= \langle gradf(\gamma(0)), \gamma'(0)\rangle \\ &= \langle gradf(p), \gamma'(0)\rangle \end{aligned}$$

sustituyendo la igualdad anterior en 3.19 tenemos:

$$h(t) \geqslant h(0) + t\langle gradf(p), \gamma'(0)\rangle$$

lo que implica que:

$$f(\gamma(t)) - f(p) \geqslant +t\langle gradf(p), \gamma'(0)\rangle.$$

Recíprocamente, supongamos que se cumple que:

$$f(\gamma(t)) - f(p) \geqslant t\langle gradf(p), \gamma'(0)\rangle \text{ con } t \in [0, +\infty),$$

con $\gamma : \mathbb{R} \to M$ *la geodésica tal que* $\gamma(0) = p$.

Sea $h : [0, +\infty) \to \mathbb{R}$ *dada por* $h(t) = f(\gamma(t))$ *con* $t \in [o, +\infty)$.

como $\gamma(0) = p$ *y* $h'(0) = \langle gradf(p), \gamma'(0)\rangle$, *ocurre que* $h(t) - h(0) \geqslant th'(0)$,

de esto se tiene que $h = f \circ \gamma$ *es convexa y por lo tanto* f *es convexa.*

Corolario 3.7.1 *Si* $f : M \to \mathbb{R}$ *es convexa entonces todos los puntos críticos de* f *son mínimos globales de* f.

Demostración 3.7.4 *Sea* $p \in M$ *un punto crítico de* f, *entonces* $gradf(p) = 0$. *Dado* $q \in M$ *con* $q \neq p$, *existe por teorema de Hopf-Rinow's una geodésica* $\gamma : [a, b] \to M$ *que*

conecta a p con q donde $\gamma(a) = p$ y $\gamma(b) = q$. Como f es convexa por teorema de condición de primer orden para convexidad se cumple: $f(\gamma(b)) - f(\gamma(a)) \geqslant \langle gradf(p), \gamma'(0)\rangle$ y dado que q es punto crítico $f(q) \geqslant f(p)$ $\forall q$. Por lo tanto, todo punto crítico de una función convexa es un mínimo global.

Teorema 3.7.6 *Sea M una variedad Riemanniana, una función $f : M \to \mathbb{R}$ es convexa si y solo si el Hessiano de f (H^f) es semi-definido positivo.*

Demostración 3.7.5 *Supongamos que f es convexa, sean $p \in M$ y $v \in T_pM$ y γ la única geodésica tal que $\gamma(0) = p$ y $\gamma'(0) = v$. Definamos $h : \mathbb{R} \to \mathbb{R}$ dada por $f \circ \gamma = h$ por lo que h es convexa. Así, por la teoría de convexidad, $h''(t) \geqslant 0$ $\forall t$, ahora por definición de derivada*

$$h'(t) = \langle gradf(\gamma(t)), v\rangle_{\gamma(t)},$$

luego

$$h''(0) = \langle H_p^f v, v\rangle \geqslant 0.$$

Recíprocamente, supongamos $\langle H_p^f v, v\rangle \geqslant 0$. Definiendo $h : \mathbb{R} \to \mathbb{R}$ dada por $f \circ \gamma = h$, se tiene que $h''(t) \geqslant 0$ y $h = f \circ \gamma : \mathbb{R} \to \mathbb{R}$ es convexa, finalmente por la definición de convexidad f es convexa.

Definición 3.7.7 *Sea M una variedad Riemanniana completa y $f : M \to \mathbb{R}$ una función real. f es llamada cuasi-convexa en M si para todo $x, y \in M$, $t \in [0, 1]$, se cumple:*

$$f(\gamma(t)) \leqslant max\{f(x), f(y)\},$$

para toda curva geodésica $\gamma : [0,1] \to M$*, tal que* $\gamma(0) = x$ *y* $\gamma(1) = y$*.*

Teorema 3.7.7 *Sea* $f : M \to \mathbb{R}$ *una función diferenciable y cuasi-convexa en una variedad Riemanniana completa* M *y sean* $x, y \in M$*. Si* $f(x) \leqslant f(y)$ *entonces:*

$$\langle gradf(y), \gamma'(0) \rangle \leqslant 0$$

donde γ *es la curva geodésica tal que* $\gamma(0) = y$ *y* $\gamma(1) = x$*.*

Prueba: Ver [13].

Definición 3.7.8 *Sea* $\Gamma \geqslant 0$ *una constante y* $f : M \to \mathbb{R}$*, si para cada* p, q *en* M *y el segmento geodésico* $\gamma : [0, a] \to M$ *que conecta a* p *y* q *con* $\gamma(0) = p$ *y* $\gamma'(a) = q$ *dada por:* $l(\gamma_{0,t}) =$ *longitud del segmento* $[\gamma(0), \gamma(t)] = \int_0^t \|\gamma'(t)\| dt$ *cumple lo siguiente*

$$\|gradf(\gamma(t)) - P_\gamma gradf(p)\| \leqslant \Gamma \|l(\gamma_{0,t})\|,$$

se dice que f *tiene gradiente* Γ *-lipschitziana.*

Definición 3.7.9 *Sea* $f : M \to \mathbb{R}$*, una función da la variedad* M *a los reales, se denomina conjunto de subnivel a:* $M^a = \{x \in M : f(x) \leqslant a\}$ *con* $a \in \mathbb{R}$

3.8. Funciones Convexas en Subvariedades de Riemann

3.8.1 Gradiente y Hessiano en Subvariedades

Sea (M, g) una variedad Riemanniana de dimensión $n + p$ y N una subvariedad de dimensión n cuya métrica inducida es denotada también por g . Denotamos por $\xi_1, ..., \xi_p$ el campo de vectores local sobre N que suponemos normal. Si ∇' es la conexión Riemanniana sobre M y ∇ es la conexión Riemanniana sobre N, entonces la fórmula de Gauss se sostiene

$$\nabla'_X Y = \nabla_X Y + \alpha(X, Y),$$

donde X y Y son campos de vectores sobre N, y

$$\alpha(X, Y) = \sum_{s=1}^{p} \Omega^s(X, Y)\, \xi_s$$

es la *segunda forma fundamental* sobre N.

Sea $f : M \to \mathbb{R}$ una función de clase $\mathcal{C}^2$. La restricción de f sobre N será denotada por f_N. Observe que

$$grad_M f = grad_N f_N + \sum_{s=1}^{p} a^s\, \xi_s, \tag{3.20}$$

donde

$$a^s = g\big(grad_M f, \xi_s\big) = df(\xi_s).$$

En consecuencia, si $x_o \in M$ es un punto crítico de f y si $x_o \in N$, entonces x_o es un punto crítico de f_N. Más, $x_o \in N$ es un punto crítico de f_N si y sólo si

$$grad_M\, f(x_o) = \sum_{s=1}^{p} a^s(x_o)\, \xi_s(x_o) \in T_{x_o}N^{\perp}.$$

Teorema 3.8.1 *Si $f : M \to \mathbb{R}$ es una función de clase $\mathcal{C}^2$ y f_N es su restricción sobre N, entonces*

$$Hess_M\, f = Hess_N\, f_N \, - \sum_{s=1}^{p} a^s\, \Omega^s.$$

Prueba: Ver [7].

Ya que una subvariedad N puede ser accidentalmente un subconjunto convexo totalmente de M, aceptamos aquí que una función f es convexa sobre N si $Hess\, f|_N \geqslant 0$.

Corolario 3.8.1

1. *Si $f : M \to \mathbb{R}$ es convexa y $grad_M\, f$ es tangente a N, entonces f_N es convexa.*

2. *Si $f : M \to \mathbb{R}$ es convexa y N es una subvariedad geodésica totalmente, entonces f_N es convexa.*

3. *Si $f : M \to \mathbb{R}$ es convexa y $df(\alpha)$ es semidefinida positiva, entonces f_N es convexa.*

Prueba: Ver [7].

Ejemplos.

1. La función $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = e^z$ es convexa y tendremos que $grad\, f = (0, 0, e^z)$. Sea N el cilindro circular $x^2 + y^2 = 1$ en $\mathbb{R}^3$. El campo de vectores unitarios sobre N es $\xi = (x, y, 0)$. Entonces $(\xi, grad\, f) = 0$ y en consecuencia f_N es convexa.

2. Sea $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = e^z$ y $N : z = \dfrac{x^2 + y^2}{2}$.

   Como

   $$df = e^z\, dz, \quad \xi = \frac{(-x, -y, 1)}{\sqrt{1 + x^2 + y^2}}, \quad \Omega = \frac{dx^2 + dy^2}{\sqrt{1 + x^2 + y^2}}.$$

   Se sigue que

   $$df(\xi) = \frac{e^z}{\sqrt{1 + x^2 + y^2}}$$

   y

   $$Hess_N\, f_N = Hess_{\mathbb{R}^3}\, f + df(\xi)\, \Omega = e^z \left(dz^2 + \frac{dx^2 + dy^2}{1 + x^2 + y^2} \right) \Big|_N.$$

   En consecuencia f_N es convexa.

3. La función

   $$f : \mathbb{R}^3 \to \mathbb{R}, \quad f(x, y, z) = \frac{x^2 + y^2}{2}$$

   es convexa. Considerando el paraboloide de rotación

   $$N : z = \frac{x^2 + y^2}{2},$$

   orientado por

   $$\xi = \frac{(-x, -y, 1)}{\sqrt{1 + x^2 + y^2}},$$

   par el cual la segunda forma fundamental es

   $$\Omega = \frac{dx^2 + dy^2}{\sqrt{1 + x^2 + y^2}}.$$

   Se sigue que

   $$Hess_N\, f_N = Hess_{\mathbb{R}^3}\, f + df(\xi)\, \Omega = \frac{dx^2 + dy^2}{1 + x^2 + y^2} \Big|_N,$$

y en consecuencia f_N es convexa.

Generalización. Si $f : \mathbb{R}^2 \to \mathbb{R}$ es convexa, entonces $\tilde{f} : \mathbb{R}^3 \to \mathbb{R}$, $\tilde{f}(x, y, z) = f(x, y)$ es convexa. La restricción de $\tilde{f}$ al gráfico de f es también convexa.

Capítulo 4

Método De Descenso Clásico Sobre Variedades.

En la optimización clásica en espacios vectoriales, los métodos de búsqueda lineal se utilizan ampliamente. Se basan en la actualización de la iteración por la elección de una dirección de búsqueda y luego se adiciona un múltiplo de ésta a la iteración anterior. Es decir, la búsqueda lineal en $\mathbb{R}^n$ se basa en la fórmula iterativa

$$x_{k+1} = x_k + \lambda_k d_k, \tag{4.1}$$

donde $d_k \in \mathbb{R}^n$ es la *dirección de búsqueda* y $\lambda_k \in \mathbb{R}^n$ es *la longitud del paso*. La adición de un múltiplo a la dirección de búsqueda requiere, obviamente, la estructura de un espacio vectorial y, en general, esto no es posible en variedades. El objetivo de este capítulo es desarrollar una teoría similar para problemas de optimización planteados en variedades no lineales. La extensión natural a variedades es seguir la dirección de búsqueda a lo largo de una curva; más específicamente, la generalización de (4.1) a una variedad M consiste en seleccionar d_k como un vector tangente a M en x_k, y realizar la búsqueda a lo largo de una curva en M cuyo vector tangente en t = 0 sea d_k. La elección de la curva se basa en el concepto de *retracción*, que veremos en la siguiente sección. Cabe destacar, que la

escogencia de una *retracción* eficiente computacionalmente es una decisión importante en el diseño de algoritmos numéricos de alto rendimiento en variedades no lineales.

4.1. Retracciones

Conceptualmente, el método más sencillo para la optimización de una función diferenciable es trasladar un punto de prueba $x(t)$ continuamente en la dirección de máxima pendiente,$-grad\, f(x)$, sobre el conjunto de restricciones hasta que se llegue a un punto en que la pendiente se anule. Los puntos x donde $grad\, f(x) = 0$ se llaman puntos estacionarios o puntos críticos de f. Una implementación numérica de la aproximación de descenso del gradiente de forma continua, requiere de la construcción de una curva γ de tal manera que $\gamma'(t) = -grad\, f(\gamma(t))$ para todo t. Excepto en circunstancias muy especiales, la construcción de una curva utilizando métodos numéricos es poco práctico. La analogía numérica más cercana es la clase de métodos de optimización que utilizan procedimientos de *búsqueda lineal*, es decir, algoritmos iterativos que, dado un punto x, calcula una dirección de descenso $\lambda = -grad\, f(x)$ (o alguna aproximación del gradiente) y se mueve en la dirección de λ hasta que se encuentre una disminución "razonable" de f. En $\mathbb{R}^n$, el concepto de movimiento en la dirección de un vector es sencillo. En una variedad, la noción de moverse en la dirección de un vector tangente, durante su estadía en la misma, es generalizada por la noción de la *función de retracción*.

Podemos ver a una retracción R por x, denotada por R_x (ver [8]), como una función de T_xM a M con una condición rígida que preserva el gradiente por x; ver figura 4.1.

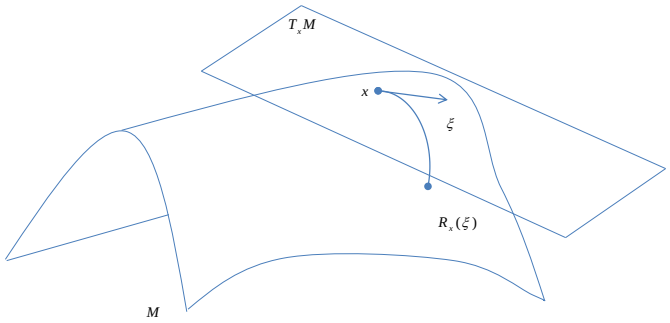

Figura 4.1: Retracción.

Definición 4.1.1 *Una* retracción *sobre una variedad M es una función R del fibrado tangente TM sobre M con las siguientes propiedades. Sea R_x la restricción de R a T_xM, así*

1. *R es continuamente diferenciable.*

2. *$R_x(0_x) = x$, donde 0_x denota el elemento cero de T_xM.*

3. *Con la identificación canónica $T_{0_x}T_xM \simeq T_xM$, R_x satisface*

$$DR_x(0_x) = id_{T_xM}, \tag{4.2}$$

donde id_{T_xM} denota la función identidad sobre T_xM.

Por lo general, suponemos que el dominio de R es todo el fibrado tangente TM. En cuanto a la condición (4.2), nótese que, ya que R_x es una función de T_xM a M que envía a 0_x a x, se sigue que $DR_x(0_x)$ es una aplicación de $T_{0_x}(T_xM)$ a T_xM. Ya que T_xM es un espacio vectorial, existe la identificación natural $T_{0_x}(T_xM) \simeq T_xM$. De aquí, por el teorema de la función inversa, R_x es un difeomorfismo local por 0_x, que además, no solamente es C^1 sino que también es biyectiva con inversa diferenciable en una vecindad V de 0_x en T_xM. Aquí nos referiremos a la condición $DR_x(0_x) = id_{T_xM}$ como la *condición de rigidez local*. Equivalentemente, para cada vector tangente ξ en T_xM, la curva $\gamma_\xi : t \to R_x(t\xi)$ satisface $\dot{\gamma_\xi}(0) = \xi$. El moverse a lo largo de esta curva γ_ξ es pensado como ir en la dirección de ξ estando señido a la variedad M.

Además de convertir los elementos de T_xM en puntos de M, un segundo propósito importante de una retracción R_x es transformar la función costo definida en una vecindad de $x \in M$ en una función costo sobre el espacio vectorial T_xM. Esto es, dada una función f a valores reales sobre una variedad M dotada con una retracción R, colocamos $\widehat{f} = f \circ R$ el cual denota el *pullback* de f a través de R. Así, para $x \in M$ se tiene

$$\widehat{f}_x = f \circ R_x, \tag{4.3}$$

que es la restricción de $\widehat{f}$ a T_xM. Nótese que $\widehat{f}_x$ es una función a valores reales sobre un espacio vectorial.También, por la condición de rigidez (4.2) tendremos (con la identificación canónica $T_x\mathcal{E} \simeq \mathcal{E}$, $\mathcal{E}$ espacio vectorial) que $D\widehat{f}_x(0_x) = Df(x)$. Ahora, si M esta dotada con una métrica de Riemannian (y T_xM con un producto interno), tendremos que

$$grad\, \widehat{f}_x(0_x) = grad\, f(x). \tag{4.4}$$

Posteriormente, retomaremos este segundo propósito de las retracciones, para conocer mejor su uso en combinación con el transporte paralelo de vectores.

Cada variedad que admite una métrica Riemannian también admite una retracción definida por la *aplicación exponencial Riemanniana*(lo que veremos más adelante). Más, cualquier otra retracción se puede pensar como una aproximación de la aplicación exponencial, aunque su dominio no necesariamente es todo TM; en tal caso M tiene que ser *completa* (ver 3.6.5). Vamos a dar un ejemplo de retracción utilizando la aplicación exponencial. Antes, vamos exponer algunos detalles teóricos que hacen falta.

Recordemos del capítulo 3 que, una *geodésica* (ver 3.5.5) sobre una variedad M dotada con una conexión afín ∇ (ver 3.5.3), es una curva con aceleración nula:

$$\frac{D^2}{dt^2}\gamma(t) = 0, \tag{4.5}$$

para todo t en el dominio de γ.

Cualquier conexión está ligada con la noción de transporte paralelo (ver 3.5.2). Dada una curva suave $\gamma : [0, 1] \to M$, el problema de valor inicial

$$\nabla_{\dot{\gamma(t)}} v = 0, v(0) = v_0, \tag{4.6}$$

define una manera de transportar el vector $v_0 \in T_{\gamma(0)}M$ hacia el vector $v(t) \in T_{\gamma(t)}M$. Para la conexión Levi-Cevita considerada aquí, esto implica que el producto interior $g_{\gamma(t)}(v(t), u(t))$ es constante para cualquier par de vectores $v(t)$ y $u(t)$, transportados paralelamente a lo largo de γ. Para $x = \gamma(0)$, $y = \gamma(t)$, denotaremos el transporte paralelo de $v_0 \in T_xM$ a $v(t) \in T_yM$ por $v(t) = T^{P_\gamma}_{x,y}v(0)$ (si existieran más de un t con $y = \gamma(t)$, la

correcta interpretación llegará a ser clara en el contexto). Como se detalla más adelante $T_{x,y}^{P_\gamma}$, podrá ser interpretado como la derivada de la aplicación $P_\gamma : T_xM \to M$. Como consideraremos M completa, entonces por el teorema de Hopf-Rinow (ver 3.6.1) cualquier par de puntos $x, y \in M$ se pueden conectar por una geodésica $\gamma : [0,1] \to M$ con $x = \gamma(0)$, $y = \gamma(1)$, donde la longitud de esta curva es

$$L[\gamma] = \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))}.$$

Y la métrica inducida por esta longitud en M es

$$dist(x,y) = \inf_{\substack{\gamma:[0,1]\to M \\ x=\gamma(0), y=\gamma(1)}} L[\gamma]$$

Ya que la geodésica γ que conecta a x con y es única, entonces $\dot{\gamma}(0) \in T_xM$ denota la aplicación logaritmo de y con respecto a x, $log_x\, y$. Esta satisface $dist(x,y) = \|log_x\, y\|_x$. La geodésica puede calcularse vía la aceleración nula,$\nabla_{\dot{\gamma}(t)}\dot{\gamma}(t) = 0$; esto es, resolviendo la ecuación (3.7) . La aplicación exponencial $exp_x : T_xM \to M,\ v \to exp_x v$, es definida como $exp_x\, v = \gamma(1)$ donde γ es solución de la ecuación ordinaria dada (3.7) con condiciones iniciales $\gamma_v(0) = x, \dot{\gamma}_v(0) = v$. Además, por la propiedad de homogeneidad de γ, tenemos la siguiente igualdad $\gamma_{kv}(t) = \gamma_v(kt)$, de la que resulta

$$exp_x(tv) = \gamma_{kv}(1) = \gamma_v(t), \tag{4.7}$$

con lo que tenemos otra posibilidad de describir una geodésica. Obviamente se tiene

$$exp_x(0) = \gamma_v(0) = x$$

Proposición 4.1.1 *Sea M una variedad completa con una conexión ∇ y $p \in M$. Entonces*

$$D\,exp_x(0_x) = id_{T_xM}. \tag{4.8}$$

Y en particular exp_x es un difeomorfismo local en 0.

Demostración:

$$D\,exp_x(0_x)(v) = \frac{d}{dt}\Big|_{t=0} exp_x(tv) = \frac{d}{dt}\Big|_{t=0} \gamma_{tv}(1) = \frac{d}{dt}\Big|_{t=0} \gamma_v(t) = v, \tag{4.9}$$

para todo $v \in M$, de donde se tiene la igualdad (4.8). Ahora, por el teorema de la función inversa, exp_x es un difeomorfismo local en $0.\diamond$

De la proposición (4.1.1), se tiene que la inversa de exp_x sobre esta vecindad es log_x; con lo que la exp_x es una parametrización (local) de M via T_xM. Por tanto, existen vecindades $\widehat{U}$ y U de $0_x \in T_xM$ y $x \in M$, respectivamente. Si $\widehat{U}$ es *estrellada* (es decir, $v \in \widehat{U}$ implica que $tv \in \widehat{U}\, para\, 0 \leqslant t \leqslant 1$), entonces U se llama *vecindad normal convexa de x*. De esto se tiene el siguiente resultado.

Proposición 4.1.2 *Sea M una variedad completa con conexión afín ∇. La aplicación exponencial sobre M inducida por ∇ es una retracción, llamada la* retracción exponencial.

Prueba: Ver [14].

Bien, a continuación el ejemplo prometido de una retracción via la exp_x : considere la geodésica $\gamma : \mathbb{R} \to M$ con $\gamma(0) = x$, parametrizada por longitud de arco, defina la retracción $P_\gamma : T_xM \to M$ como

$$P_\gamma(v) = exp_{p_\gamma(v)}\Big(T^{P_\gamma}_{x,p_\gamma(v)}[v - \pi_\gamma(v)]\Big), \tag{4.10}$$

donde $p_\gamma(v) = exp_x(\pi_\gamma(v))$ y π_γ denota la proyección ortogonal sobre el *generador* de $\dot{\gamma}(0)$. Esta retracción, corresponde a moverse a lo largo de la línea geodésica γ de acuerdo con la componente de v paralela a $\dot{\gamma}(0)$; luego, se sigue una nueva geodésica en la dirección del transporte paralelo de la componente de v ortogonal a $\dot{\gamma}(0)$ (ver figura 4.2).

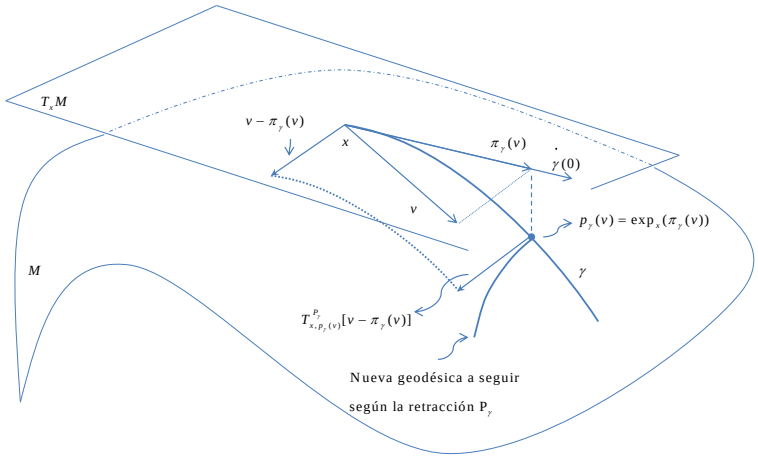

Figura 4.2: Retracción vía la exponencial.

Más adelante tendremos que considerar el transporte paralelo de un vector del espacio tangente T_xM a otro de T_yM; es decir, consideraremos el isomorfismo $T_{x,y} : T_xM \to T_yM$. Pero estamos interesados en el operador $T^{R_x}_{x,y}$ que representa la derivada $D\,R_x(v)$ de R_x por $v \in T_xM$ con $R_x(v) = y$. En ese sentido el transporte paralelo $T^{P_\gamma}_{x,y}$ a lo largo de una geodésica γ que conecta a x con y pertenece a la retracción P_γ. Por ejemplo, de nuevo por la variación de la aplicación exponencial, podemos definir el transporte vectorial evaluado

por el representativo de y en T_xM como,

$$T_{x,y}^{exp_x} = D\, exp_x(log_x\, y), \tag{4.11}$$

que lleva a $v_0 \in T_xM$ sobre el vector tangente $\dot{\gamma}(0)$ de $\gamma : t \to exp_x(log_x\, y + t\, v_0)$. También se pueden considerar para el trasporte vectorial $T_{x,y}$, los adjuntos $\left(T_{y,x}^{P_\gamma}\right)^*$, $\left(T_{y,x}^{exp_y}\right)^*$ (definidos por $g_y(v, T_{y,x}^*\, w) = g_x(T_{y,x}\, v, w)\, \forall v \in T_yM, w \in T_xM$) o las inversas $\left(T_{y,x}^{P_\gamma}\right)^{-1}$, $\left(T_{y,x}^{exp_y}\right)^{-1}$. Nótese que $T_{x,y}^{P_\gamma}$ es una isometría con $T_{x,y}^{P_\gamma} = \left(T_{y,x}^{P_{\tilde{\gamma}}}\right)^* = \left(T_{y,x}^{P_{\tilde{\gamma}}}\right)$, donde γ es la geodésica que conecta x con y y $\tilde{\gamma}(\cdot) = \gamma(-\cdot)$. Además, $log_x\, y$ se puede transportar sobre el mismo vector $T_{x,y}\, log_x\, y = \dot{\gamma}(1)$ por $T_{x,y}^{P_\gamma}$, o por $T_{x,y}^{exp_x}$, o por sus adjuntos o por sus inversas.

Bien, volviendo al segundo propósito de las retracciones, en detalle sería: Dada una función suave $f : M \to \mathbb{R}$, definimos su derivada $Df(x)$ en $x \in M$ como un elemento del espacio dual de T_xM vía $Df(x)v = v\, f$. El teorema de representación de Riez (ver [19]), implica la existencia de un $\nabla f(x) \in T_x\, M$ tal que $Df(x)v = g_x(\nabla f(x), v)$ para todo $v \in T_xM$, que denota el *gradiente* de f en x. Luego, sobre T_xM, se define

$$f_{R_x} = f \circ R_x, \tag{4.12}$$

para la retracción R_x. Ya que $D\, R_x(0) = id$ tendremos que $D\, f_{R_x}(0) = D\, f_x$. Además, $f_{R_x} = f_{R_y} \circ R_y^{-1} \circ R_x$, donde R_y^{-1} existe y así para $y = R_x(v)$, tenemos que

$$D\, f_{R_x}(v) = D\, f_{R_y}(0)\, DR_x(v) = D\, f(y)\, T_{x,y}^{R_x}, \quad \nabla\, f_{R_x}(v) = \left(T_{x,y}^{R_x}\right)^* \nabla\, f(y).$$

4.2. Búsqueda Lineal: minimización sobre Variedades

Como se dijo al comienzo de este capítulo, los métodos de búsqueda lineal, están basados en la actualización de la iteración, por la elección de una dirección de búsqueda, hay que agregar un múltiplo (el tamaño o longitud del paso) a ésta dirección, lo que obviamente, requiere de la estructura de un espacio vectorial y, en general, esto no es posible en variedades. La extensión natural a variedades es seguir la dirección de búsqueda a lo largo de una curva. Entonces, vamos a considerar el algoritmo iterativo de la siguiente forma genérica.

Algoritmo 1

Entrar: $f : M \to \mathbb{R}$, iterado inicial $x_0 \in M$, $k = 0$.

repetir

- escoja una dirección de descenso $d_k \in T_{x_k} M$
- escoja una retracción $R_{x_k} \in T_{x_k} M \to M$
- escoja una longitud de paso $\lambda_k \in \mathbb{R}$
- colocar $x_{k+1} = R_{x_k}(\lambda_k \, d_k)$
- $k \leftarrow k + 1$

hasta que x_{k+1} minimice suficientemente a f. $\diamond$

Como d_k es dirección de descenso entonces $Df(x_k)\, d_k = \big\langle (grad\, f)(x_k), d_k \big\rangle < 0$. Esta propiedad asegura que la función objetivo o costo, f, en verdad decrece en la dirección de

búsqueda.

Para la escogencia de la longitud del paso λ_k, existen varias posibilidades. En general, la escogencia de d_k tiene que satisfacer ciertos requerimiento de calidad, en nuestro caso $d_k = -Df(x_k)$. Para el objetivo de lo que estamos estudiando en este trabajo, nos concentraremos en la *condición de Wolfe*, es decir, para una dirección de descenso dada $d \in T_x M$, la escogencia de la longitud del paso λ tiene que satisfacer

$$f(R_x(\lambda\, d)) \leqslant f(x) + c_1 \lambda\, Df(x)\, d, \tag{4.13a}$$

$$Df(R_x(\lambda\, d))\, T^{R_x}_{x,R_x(\lambda\, d)}\, d \geqslant c_2 Df(x)\, d, \tag{4.13b}$$

donde $0 < c_1 < c_2 < 1$. Nótese que ambas condiciones se pueden reescribir como

$$f_{R_x}(\lambda\, d)) \leqslant f_{R_x}(0) + c_1 \lambda\, Df_{R_x}(0)\, d, \tag{4.14a}$$

$$Df_{R_x}(\lambda\, d)\, d \geqslant c_2 Df_{R_x}(0)\, d, \tag{4.14b}$$

que es la condición de Wolfe clásica para minimizar la función f_{R_x} del espacio vectorial $T_x M$ a $\mathbb{R}$. Si la segunda condición (4.13b) es reemplazada por

$$|Df(R_x(\lambda\, d))\, T^{R_x}_{x,R_x(\lambda\, d)}\, d\,| \leqslant -\, c_2 Df(x)\, d, \tag{4.15}$$

obtenemos la condición fuerte de Wolfe (generalmente se utiliza esta condición para un análisis último en un esquema cuasi-Newton). Ahora, en muchos algoritmos de optimización basta que la primera condición (4.13a) sea satisfecha, obviamente es la condición de Armijo; que en este caso es la generalizada y que trabajaremos mas adelante utilizando la

aplicación exponencial como retracción.

Similar al lema 2.2.1 para el caso $\mathbb{R}^n$, se tiene que, dada una dirección de descenso d, una longitud de paso factible $\lambda > 0$ siempre se puede encontrar.

Proposición 4.2.1 *(Longitud de paso factible). Sea $x \in M$, $d \in T_x M$ una dirección de descenso y $f_{R_x} : span\{d\} \to \mathbb{R}$ diferenciable continuamente. Entonces existe $\lambda > 0$ que satisface las condiciones (4.13) y (4.15).*

Prueba

Al reescribir las condiciones de Wolfe como en (4.14) (4.15, respectivamente), el argumento clásico que se tiene para estas condiciones en espacios vectoriales se puede aplicar.

Bien, para ver los detalles, consideraremos los siguientes ingredientes:

Sean $x_0 \in M$(completa); $\varepsilon > 0$, los conjuntos $B(x_0, \varepsilon) = \{x \in M : d(x, x_0) < \varepsilon\}$ y $\bar{B}(x_0, \varepsilon) = \{x \in M : d(x, x_0) \leqslant \varepsilon\}$, donde d es la distancia Riemanniana sobre M. $\mathcal{D}_x$ es el conjunto de vectores $v \in T_x M$ tal que la máxima geodésica $\gamma_v(t)$ está definida en $[0, 1]$. El conjunto $\mathfrak{X}(M)$ denotará los campos de vectores sobre M y $\mathfrak{F}(M)$ el de funciones suaves sobre M.

Luego, para $x \in \bar{B}(x_0, \varepsilon)$, consideraremos, $d_x \in \mathcal{D}_x$, $\gamma_{d_x}(\lambda) = R_x(\lambda\, d_x)$ la geodésica que parte de x en dirección de d_x y por $\xi_\lambda = \gamma'_{d_x}(\lambda)$ el campo vectorial de $\mathfrak{X}(M)$, tangente a lo largo de $\gamma_{d_x}(\lambda)$ en donde $\xi_0 = d_x$. Así, colocando $\omega(\lambda) = f(R_x(\lambda\, d_x)) - f(x)$ y como $\dfrac{DR_x(\lambda\, d_x)}{\lambda} = \xi_\lambda$ resulta, por el desarrollo de Taylor de primer orden de ω en 0 que

$$f(R_x(\lambda\, d_x)) - f(x) = \omega(\lambda) - \omega(0) = \lambda\, \omega'(0) + o(\lambda)$$

$$= \lambda\Big\langle \nabla f(x), d_x \Big\rangle + o(\lambda) = \lambda\langle\, grad\, f(x), d_x\rangle + o(\lambda)$$

$$= \lambda c\langle\, grad\, f(x), d_x\rangle + \lambda\,[(1-c)\,\langle\, grad\, f(x), d_x\rangle + \frac{o(\lambda)}{\lambda}], \tag{4.16}$$

y ya que $\lim\limits_{x\to+\infty} \dfrac{o(\lambda)}{\lambda} = 0,\ c < 1$ y d_k es dirección de descenso para la cual se cumple que $Df(x_k)\,d_k = \big\langle (grad\, f)(x_k), d_k\big\rangle < 0$ en el **Algoritmo 1**, entonces, la última rama de (4.16) es negativa para λ suficientemente pequeño. En consecuencia, tenemos para d_x que $f(R_x(\lambda\, d_x)) \leqslant f(x) + \lambda c\langle\, grad\, f(x), d_x\rangle$ con λ suficientemente pequeño. Es decir, la condición (4.13a) se cumple.

Por otra parte, para obtener la segunda condición (4.13b), observe que, f_{R_x} es acotada en $\mathcal{D}_x$, puesto que f lo es en el compacto $\bar{B}(x_0, \varepsilon)$ y R_x en $\mathcal{D}_x$. Esto es, ya que d_x es dirección de descenso, se tiene que f es acota por debajo en $\{x \in \bar{B}(x_0, \varepsilon) : f(R_x(\lambda\, d_x)) < f(x)\}$, con lo que f_{R_x} es acotada por abajo en el conjunto

$$\{\xi_\lambda \in \mathcal{D}_x : f_{R_x}(\xi_\lambda) < f_{R_x}(0_x)\}. \tag{4.17}$$

Ahora, para cualquier λ no negativo definimos

$$\beta(\lambda) = f_{R_x}(\lambda\, d_x) + c_1\lambda^2\langle\, grad\, f(x), d_x\rangle \tag{4.18}$$

Entonces tenemos que

$$\lim_{\lambda\to 0^+} \frac{\beta(\lambda) - \beta(0)}{\lambda} = \lim_{\lambda\to 0^+} \frac{f_{R_x}(\lambda\, d_x) - f_{R_x}(0_x)}{\lambda} = \langle\, grad\, f(x), d_x\rangle < 0.$$

Luego, existe un $\lambda' > 0$, tal que para $\lambda \in (0, \lambda']$ se tiene que

$$\frac{\beta(\lambda) - \beta(0)}{\lambda} \leqslant 0. \tag{4.19}$$

Por (4.17) tenemos que

$$\lim_{\lambda\to+\infty}\frac{\beta(\lambda)-\beta(0)}{\lambda}=+\infty,$$

puesto que este limite es

$$\lim_{\lambda\to+\infty}\Big[\frac{f_{R_x}(\lambda\, d_x)-f_{R_x}(0_x)}{\lambda}+c_1\,\lambda\langle\, grad\, f(x),d_x\rangle\Big],$$

donde el primer sumando es acotado y el segundo no.

Sea

$$\widehat{\lambda}'=inf\Big\{\lambda>0:\frac{\beta(\lambda)-\beta(0)}{\lambda}=0\Big\}.$$

Por el teorema del valor intermedio y (4.19), tenemos que $\widehat{\lambda}'$ satisface

$$\frac{\beta(\widehat{\lambda}')-\beta(0)}{\widehat{\lambda}'}=0. \tag{4.20}$$

Así, para cada $\lambda\in(0,\widehat{\lambda}']$, se tiene que

$$\frac{\beta(\lambda)-\beta(0)}{\lambda}\leqslant 0. \tag{4.21}$$

Aplicando el teorema del valor medio y (4.20) tenemos que existe $\theta'\in[0,1]$ tal que

$$\beta'(\theta'\,\widehat{\lambda}')=0.$$

Por tanto,

$$Df_{R_x}(\theta'\,\widehat{\lambda}' d_x)\,d_x+2\,c_1\,\theta'\,\widehat{\lambda}'\,\langle\, grad\, f(x),d_x\rangle=0.$$

Es decir, si $0<c_1<c_2<1$ tenemos que

$$Df_{R_x}(\theta'\,\widehat{\lambda}' d_x)\,d_x=-2\,c_1\,\theta'\,\widehat{\lambda}'\,\langle\, grad\, f(x),d_x\rangle\geqslant -2\,c_2\,\theta'\,\widehat{\lambda}'\,\langle\, grad\, f(x),d_x\rangle. \tag{4.22}$$

Al sustituir $\lambda = \theta' \widehat{\lambda}' < \widehat{\lambda}'$ y cancelando términos semejantes en (4.22), tenemos la condición de Wolfe (4.13b); más, de (4.21) y (4.22) concluimos que $\lambda = \theta' \widehat{\lambda}'$ es la longitud del paso que se requiere.$\diamond$

4.3. Convergencia para el método de búsqueda lineal en variedades

El estudio de convergencia en variedades es una generalización directa del caso $\mathbb{R}^n$. Así que, una sucesión infinita $\{x_k\}$ de puntos de una variedad M se dice que es *convergente* si existe una carta $(\mathcal{U}, \varphi)$ de M, un punto $x^* \in \mathcal{U}$, y un conjunto K > 0 tal que $x_k \in \mathcal{U}$ para todo $k \in$ K y la sucesión $\{\varphi(x_k)\}$ converge a $\varphi(x^*)$. El punto $\varphi^{-1}(\lim_{k \to \infty} \varphi(x_k))$ se llama el *limite* de la sucesión convergente $\{x_k\}$. Cada sucesión convergente de una variedad (Hausdorff) tiene uno y sólo un punto limite (ver [16]).

La convergencia de los algoritmos de búsqueda lineal, naturalmente dependen, tanto de una buena escogencia de la longitud del paso como de una buena dirección de búsqueda. Con este objetivo en mente, vamos a utilizar el teorema de Zoutendijk dado en $\mathbb{R}^n$ (ver [2]) el cual asocia una serie numérica con direcciones de búsquedas que son válidas para el método del gradiente conjugado y en general para algoritmos de descenso, adaptado a una variedad Riemanniana M (ver [1]) . Igual que en $\mathbb{R}^n$, definimos en M, el ángulo θ_k entre una dirección de búsqueda d_k y la dirección de descenso de mayor pendiente, el gradiente

negativo $-D\,f(x)$, como

$$cos\,\theta_k = \frac{-Df(x_k)\,d_k}{\|Df(x_k)\|_k\,\|d_k\|_k} \tag{4.23}$$

Teorema 4.3.1 *(Teorema de Zoutendijk). Sea $f : M \to \mathbb{R}$ acotada por abajo y de clase C^1. Suponga que λ_k en el **Algoritmo 1** satisfaces las ecuaciones (4.14). Si las funciones $f_{R_{x_k}}$ son diferenciables continuamente sobre el span d_k y existe una constante Lipschitziana uniforme $L > 0$ tal que*

$$|Df_{R_{x_k}}(t\,d_k)\,d_k - Df_{R_{x_k}}(0_{x_k})(d_k)| \leqslant L\,t \qquad d_k \in T_{x_k}M \; con \; \|d_k\|_{x_k} = 1,\; x_k \in M,\; t \geqslant 0. \tag{4.24}$$

Entonces, la serie numérica

$$\sum_{k=0}^{\infty} cos\,\theta_k\,\|Df(x_k)\|^2_{x_k} < \infty, \tag{4.25}$$

converge.

Prueba:

De la condición de Wolfe (4.14b) tenemos que

$$(Df_{R_{x_k}}(\lambda_k\,d_k)\,d_k - Df_{R_{x_k}}(0)\,d_k) \geqslant (c_2 - 1)Df_{R_{x_k}}(0)\,d_k. \tag{4.26}$$

Luego, utilizando la condición de Lipschit, resulta que

$$\lambda_k\,L\,\|d_k\|^2_{x_k} \;\geqslant\; (Df_{R_{x_k}}(\lambda_k\,d_k)\,d_k - Df_{R_{x_k}}(0)\,d_k) \geqslant (c_2 - 1)Df(x_k)\,d_k,$$

de la cual obtenemos que

$$\lambda_k \geqslant \frac{(c_2 - 1)Df(x_k)\,d_k}{L\,\|d_k\|^2_{x_k}}.$$

Ahora, del **Algoritmo 1** y de la condición de Wolfe (4.14a)se tiene que

$$f(x_{k+1}) \leqslant f(x_k) - c_1 \frac{(1-c_2)}{L} cos^2\theta_k \|Df(x_k)\|^2_{x_k}.$$

Colocando $c = c_1 \frac{(1-c_2)}{L} > 0$ y, sumando esta expresión sobre todos los índices menores que k , obtenemos

$$f(x_{k+1}) \leqslant f(x_0) - c \sum_{j=0}^{k} cos^2\theta_j \|Df(x_j)\|^2_{x_j}.$$

Ya que f es acotada por debajo, existe una constante $\mathcal{C}$ tal que $f(x) \geqslant \mathcal{C}$ para cualquier $x \in M$. Luego

$$\mathcal{C} \leqslant f(x_{k+1}) \leqslant f(x_0) - c \sum_{j=0}^{k} cos^2\theta_j \|Df(x_j)\|^2_{x_j}. \tag{4.27}$$

Así, tomando limite en (4.27) obtenemos

$$\sum_{k=0}^{\infty} cos^2\theta_k \|Df(x_k)\|^2_{x_k} \leqslant \frac{f(x_0) - \mathcal{C}}{c} < \infty,$$

lo que concluye la prueba.$\diamond$

Bien, la condición de Zoutendijk (4.25) implica que

$$cos^2\theta_k \|Df(x_k)\|^2_{x_k} \to 0. \tag{4.28}$$

Si nuestro método para la elección de la dirección de búsqueda d_k en el **Algoritmo 1** garantiza que el ángulo θ_k definido por (4.25) está acotado y alejado de 90°, existe una constante positiva δ tal que

$$cos\,\theta_k \geqslant \delta > 0 \qquad \forall\, k.$$

Pero si esto es así, tenemos de (4.28), que

$$\lim_{k \to \infty} \|Df(x_k)\| = 0. \tag{4.29}$$

En otras palabras, podemos asegurar que la norma del gradiente $\|Df(x_k)\|$ converge a cero, siempre que las direcciones de búsqueda no sean demasiado cercanas ortogonalmente al gradiente. En particular, el método de descenso de mayor pendiente (para el cual la dirección de búsqueda d_k es paralela al gradiente negativo) produce una sucesión de gradientes que converge a cero, si se utiliza una búsqueda lineal que satisface la condición de Wolfe.

Corolario 4.3.1 *(Convergencia: Descenso de mayor pendiente generalizado). Suponga que la dirección de búsqueda en el* ***Algoritmo 1*** *es solución de* $\mathcal{B}_k(d_k, v) = -Df(x_k)v$ *para todo* $v \in T_{x_k}M$*, donde las* $\mathcal{B}_k$ *son formas bilineales acotadas y coercitivas uniformemente en* $T_{x_k}M$ *(el caso* $\mathcal{B}_k(.,.) = g_{x_k}(.,.)$ *admite la dirección de descenso de mayor pendiente). Entonces, bajo las mismas condiciones del teorema (4.3.1)se tiene que* $\|Df(x_k)\|_{x_k} \to 0$.

Prueba:

Siendo las $\mathcal{B}_k$ acotadas y coercitivas uniformemente en $T_{x_k}M$,tenemos $\mathcal{B}_k(d_k, v) \to +\infty$ cuando $\|(d_k, v)\| \to +\infty$; con lo que $cos\,\theta_k = \dfrac{\mathcal{B}_k(d_k, d_k)}{\|\mathcal{B}_k(d_k, d_k)\|d_k\|_{x_k}}$ son uniformemente acotadas por encima de cero. Luego la convergencia se obtiene aplicando el teorema de Zoutendijk. $\diamond$

Por la continuidad del gradiente Df, y el corolario anterior, se tiene que, cualquier punto

limite x^* de la sucesión $\{x_k\}$ es un punto estacionario de la función f. Por lo tanto en variedades de dimensión finita, si $\{x \in M : f(x) \leqslant f(x_0)\}$ es acotado, $\{x_k\}$ se puede descomponer en subsucesiones que converjan a un punto estacionario. En variedades de dimensión infinita , donde sólo puede esperarse una convergencia débil de las subsucesiones, en general, esto no es cierto (lo cual no es sorprendente, dado que no se concede ni siquiera la existencia de puntos fijos, sin condiciones más fuertes sobre f, tales como secuencialmente débil semi-continua inferiormente). Más aún, el punto limite puede no ser único. A continuación, trataremos un caso especial de convergencia global, inspirada por la teoría clásica en $\mathbb{R}$, donde algunas restricciones se le harán a la dirección de descenso d_k y a la longitud del paso λ_k (ver [16]).

Definición 4.3.1 *(Sucesión gradiente-relacionada) Dada una función costo f sobre una variedad Riemanniana M, una sucesión $\{d_k\}$, $d_k \in T_{x_k}M$, es* gradiente-relacionada *si, para cualquier subsucesión $\{x_k\}_{k\in\mathrm{K}}$ de $\{x_k\}$ que converge a un punto no crítico de f, la correspondiente subsucesión $\{d_k\}_{k\in\mathrm{K}}$ es acotada y satisface que*

$$\limsup_{k\to\infty,\, k\in\mathrm{K}} \left\langle grad\, f(x_k), d_k \right\rangle < 0.$$

La siguiente definición, esta relacionada con el tamaño del paso λ_k en la condición de Armijo.

Definición 4.3.2 *(Punto de Armijo) Dada una función costo f sobre una variedad Riemanniana M con retracción R, un punto $x \in M$, un vector tangente $d \in T_x M$, y escalares $\widehat{\alpha} > 0$, β, $c \in (0, 1)$, el* punto de Armijo *es el numero real λ^A tal que $d^A =$*

$\lambda^A\, d = \beta^m\, \widehat{\alpha}\, d$, *donde* m *es el entero no negativo más pequeño que satisface la condición de Armijo*

$$f(R_x(\beta^m\, \widehat{\alpha}\, d)) - f(x) \leqslant c\big\langle grad\, f(x),\, \beta^m\, \widehat{\alpha}\, d\,\big\rangle_x.$$

El numero real $\lambda^A = \beta^m\, \widehat{\alpha}$ *es el tamaño del paso de la condición de Armijo.*

Nótese que, si escogemos $x_{k+1} = R_{x_k}(\lambda_k^A\, d_k)$ en el paso 4 del **Algoritmo 1**, la condición de Armijo (4.13a) se cumple; pero no es obligatoria esta escogencia. Esta condición de Armijo da mucho margen de maniobra para sacar bastante información relacionada con un problema que puede conducir a un Algoritmo eficiente. Por ejemplo, se podría escoger $x_{k+1} = R_{x_k}(\lambda_k^*\, d_k)$, donde $\lambda_k^* = arg\min_{\lambda} f(R_{x_k}(\lambda\, d_k))$ satisface (4.13a); si la búsqueda lineal se lleva eficientemente.

Teorema 4.3.2 *(Punto crítico de la función objetivo) Sea* $\{x_k\}$ *una sucesión infinita de iterados generados por el* ***Algoritmo 1****. Entonces cada punto de acumulación de* $\{x_k\}$ *es un punto crítico de la función costo* f*.*

Prueba:

Por reducción a lo absurdo, supongamos que existe una subsucesión $\{x_k\}_{k\in \mathrm{K}}$ convergiendo a un punto x^* con $grad\, f(x^*) \neq 0$. Ya que $\big\{f(x_k)\big\}$ es no creciente, se sigue que toda ella converge a $f(x^*)$. En consecuencia $f(x_{k+1}) - f(x_k)$ se va a cero. Ahora, de la construcción del algoritmo se tiene que,

$$f(x_{k+1}) - f(x_k) \leqslant c\, \sigma\, \lambda_k \big\langle grad\, f(x_k),\, d_k\,\big\rangle_x.$$

Ya que $\{d_k\}$ es gradiente-relacionada, resulta que $\{\lambda_k\}_{k\in \mathrm{K}} \to 0$. Como los λ_k son determinados por la condición de Armijo, entonces para todo k más grande que algún $\widehat{k}$, se tiene que $\lambda_k = \beta^{m_k}\,\widehat{\alpha}$, donde m_k es un entero más grande que cero. Pero esto significa que $\dfrac{\lambda_k}{\beta}\,d_k$ no satisface la condición de Armijo. Por lo que

$$f(R_{x_k}\Big(\frac{\lambda_k}{\beta}\,d_k\Big)) - f(x_k) > \sigma\frac{\lambda_k}{\beta}\big\langle grad\, f(x_k),\, d_k\,\big\rangle_{x_k} \quad \forall k \in \mathrm{K},\, k \geqslant \widehat{k}.$$

Colocando

$$\widehat{d}_k = \frac{d_k}{\|d_k\|} \quad y \quad \widehat{\lambda}_k = \frac{\lambda_k\,\|d_k\|}{\beta},$$

en la desigualdad anterior tenemos

$$\frac{f_{R_{x_k}}(\widehat{\lambda}_k\,\widehat{d}_k) - f_{R_{x_k}}(0_{x_k})}{\widehat{\lambda}_k} > \sigma\,\big\langle grad\, f(x_k),\, \widehat{d}_k\,\big\rangle_{x_k}, \quad \forall k \in \mathrm{K},\, k \geqslant \widehat{k}$$

donde $f_{R_{x_k}} = f \circ R_{x_k}$. Por el teorema del valor medio, existe $\tau \in [0, \widehat{\lambda}_k]$ tal que

$$Df_{R_{x_k}}(\tau\,\widehat{d}_k)\,\widehat{d}_k > \sigma\,\big\langle grad\, f(x_k),\, \widehat{d}_k\,\big\rangle_{x_k}, \quad \forall k \in \mathrm{K},\, k \geqslant \widehat{k}, \tag{4.30}$$

donde la diferencial se toma sobre el espacio vectorial $T_{x_k}\,M$. Ya que $\{\lambda_k\}_{k\in \mathrm{K}} \to 0$ y la sucesión $\{d_k\}$ es gradiente-relacionada, por tanto acotada, resulta que $\{\widehat{\lambda}_k\}_{k\in \mathrm{K}} \to 0$. Más aún, ya que $\widehat{d}_k$ es de norma uno, éste pertenece a un conjunto compacto, con lo que existe un conjunto de índice $\widehat{\mathrm{K}} \subseteq \mathrm{K}$ para el cual $\{\widehat{d}_k\}_{k\in \widehat{\mathrm{K}}} \to \widehat{d}_*$ para algún $\widehat{d}_*$ con $\|\widehat{d}_*\| = 1$. Tomando limite en (4.30) sobre $\widehat{\mathrm{K}}$, y ya que la métrica Riemanniana es continua, $f \in \mathcal{C}^1$, con $Df_{R_{x_k}}(0_{x_k})\,\widehat{d}_k = \big\langle grad\, f(x_k),\, \widehat{d}_k\,\big\rangle_{x_k}$, tenemos que

$$\big\langle grad\, f(x_*),\, \widehat{d}_*\,\big\rangle_{x_*} \geqslant \sigma\,\big\langle grad\, f(x_*),\, \widehat{d}_*\,\big\rangle_{x_*}.$$

De donde se tiene que

$$\left\langle grad\, f(x_*),\, \widehat{d_*}\, \right\rangle_{x_*}(1-\sigma) \geqslant 0.$$

Pero $\sigma < 1$, así que

$$\left\langle grad\, f(x_*),\, \widehat{d_*}\, \right\rangle_{x_*} \geqslant 0.$$

Por otra parte, la sucesión $\{d_k\}$ es gradiente-relacionada, con lo que

$$\left\langle grad\, f(x_*),\, \widehat{d_*}\, \right\rangle_{x_*} < 0,$$

obteniéndose una contradicción.$\diamond$

4.4. Aplicación

Ejemplo1.[Con Descenso (ver [7]). *Buscando el mínimo sobre la esfera unitaria* S^2] Sea $S^2 : x^2 + y^2 + z^2 = 1$ la esfera en $\mathbb{R}^3$. Ya que S^2 es un conjunto compacto, cualquier función que sea por lo menos continua sobre S^2 tiene un mínimo y un máximo global.

Sea $f : \mathbb{R}^3 \to \mathbb{R}, \quad f(x, y, z) = e^z$. Si queremos determinar los extremos de f sobre la esfera $S^2 : x^2 + y^2 + z^2 = 1$, podemos utilizar los multiplicadores de Lagrange y encontrar que, $p = (0, 0, 1)$ y $q = (0, 0, -1)$ serían los puntos críticos. En el punto p tendríamos el máximo de f, $max\, f = f(p) = e$, en el punto q el mínimo, $min\, f = f(q) = 1/e$.

Ahora de acuerdo a lo estudiado anteriormente es este capítulo, el problema se puede resolver como sigue.Sea $f : S^2 \to \mathbb{R}, \quad f(x, y, z) = e^z$. Encontramos, aplicando la teoría de

subvariedades que,(ver [3.20])

$$gra\, f = (0,0,e^z),\ \xi = (x,y,z)\ a = g(grad_{\mathbb{R}^3} f, \xi) = df(\xi) = (0,0,e^z)\cdot(x,y,z) = z\, e^z.$$

Luego

$$grad_{S^2} f = grad_{\mathbb{R}^3} f - a\,\xi = (0,0,e^z) - z\, e^z\,(x,y,z),$$

de donde tenemos que

$$grad_{S^2} f = e^z\,(-xz, -yz, 1-z^2).$$

De aquí que los puntos críticos son $p = (0,0,1)$ y $q = (0,0,-1)$ como se dijo antes.

Con respecto a la geodésica en la esfera S^2, tenemos las siguientes ecuaciones paramétricas

$$\begin{cases} x = a\,cos\,t + d\,sen\,t \\ y = b\,cos\,t + e\,sen\,t \\ z = c\,cos\,t + f\,sen\,t, \quad t \in \mathbb{R}, \end{cases}$$

donde (a,b,b) y (a,b,b) son versores ortogonales ,es decir, de módulo uno. Colocando el punto inicial como

$$p_1 = \Big(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0\Big).$$

Ya que $grad_{S^2} f(p) = \vec{k} = (0,0,1)$, la dirección y el sentido de descenso de f por p_1 son indicado por $-\vec{k}$. La geodésica que inicia en p_1 en dirección de $-\vec{k}$ es

$$x(0) = \frac{1}{\sqrt{2}},\ y(0) = \frac{1}{\sqrt{2}},\ z(0) = 0,\ x'(0) = 0,\ y'(0) = 0,\ z'(0) = -1$$

o equivalente a

$$a = \frac{1}{\sqrt{2}},\ b = \frac{1}{\sqrt{2}},\ c = 0,\ d = 0,\ e = 0,\ f = -1;$$

con lo que

$$x = \frac{1}{\sqrt{2}}\, cost,\ y = \frac{1}{\sqrt{2}}\, cost,\ z = -sent,\ t \in [0, \infty).$$

Sea

$$\omega(t) = f(x(t), y(t), z(t)) = e^{-sent},\ t \in [0, \infty).$$

Así que, en $t = \pi/2$,$\omega(\pi/2) = f(0, 0, -1) = f(q) = e^{-1}$, obtiene su mínimo por q. Ahora, como $grad_{S^2} f(q) = 0$, se tiene que q es un punto crítico de f; más, es un mínimo global, pues el Hessiano en este punto es definido positivo. Este extremo se obtuvo después de realizar un único paso. Bien para chequear este resultado, fijamos un punto de inicio $p_2 = \left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right)$. Encontramos que

$$grad_{S^2} f(p_2) = e^{1/\sqrt{2}}\Big(-\frac{1}{2}\,\vec{i} + \frac{1}{2}\,\vec{k}\Big).$$

De esta forma el versor $\vec{v} = \frac{\vec{i}}{\sqrt{2}} - \frac{\vec{k}}{\sqrt{2}}$, indica la dirección y el sentido de f por p_2. Aquí, la geodésica que comienza por p_2 en dirección de $\vec{v}$ es

$$x = \frac{1}{\sqrt{2}}\,(cost + sent),\ y = 0,\ z = \frac{1}{\sqrt{2}}\,(cost - sent)\ t \in [0, \infty).$$

Ya que la función

$$\phi(t) = f(x(t), y(t), z(t)) = e^{sen(\pi/4 - t)}$$

obtiene su mínimo e^{-1} por $t = 3\pi/4$, es decir, por $x = 0$, $y = 0$, $z = -1$; de nuevo llegamos al punto q (ver figura 4.3).

Ejemplo 2.[Con Armijo (ver [16]). *Minimizando el Cociente de Rayliegh sobre* $\mathcal{S}^{n-1}$]

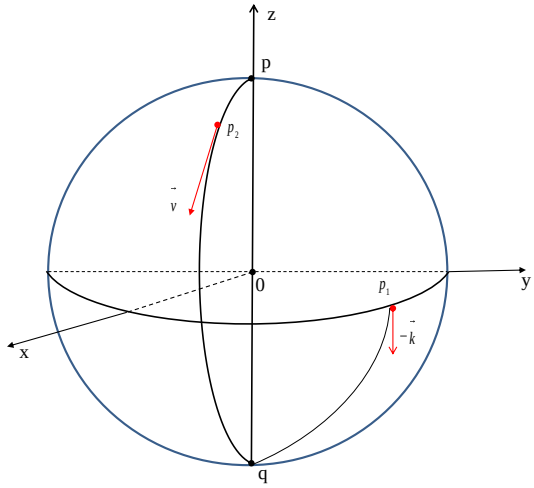

Figura 4.3: Método de descenso en S^2.

Definición 4.4.1 *(Cociente de Rayleigh) Si x es autovector de la matriz A, entonces su correspondiente autovalor es dado por*

$$\mu = \frac{x^T A\, x}{x^T \ \ x}.$$

Este cociente se llama el ***Cociente de Rayleigh****.*

Se sabe que este cociente de Rayleigh puede acelerar convergencia en los métodos iterativos. Con el objetivos de presentar una aplicación del **Algoritmo 1**, vamos a utilizarlo aquí.

Se aplicara el **Algoritmo 1** al problema de encontrar un minimizador de

$$f : \mathcal{S}^{n-1} \to \mathbb{R} \quad donde \quad f(x) = x^T A\, x, \tag{4.31}$$

que denota el cociente de Rayleigh sobre la esfera unitaria. La matriz A es simétrica ($A = A^T$) y no necesariamente definida positiva (ver 1.2.2). Sea μ_1 el autovalor más

pequeño de A cuyo autovector asociado de norma unitaria es v_1.

Considere la función costo

$$\widehat{f} : \mathcal{S}^{n-1} \to \mathbb{R} \quad donde \quad \widehat{f}(x) = x^T\, A\, x, \tag{4.32}$$

cuya restricción a la esfera unitaria $\mathcal{S}^{n-1}$ admite (4.31). Aquí, estamos viendo a $\mathcal{S}^{n-1}$ como una subvariedad Riemanniana del espacio euclídeo $\mathbb{R}^n$, dotada con la métrica Riemanniana

$$\widehat{g}(u, v) = u^T\, v.$$

Dado $x \in \mathcal{S}^{n-1}$, tendremos que

$$D\widehat{f}\, v = v^T\, A\, x + x^T\, A\, v = 2\, v^T\, A\, x,$$

para todo $v \in T_x\mathbb{R}^n \simeq \mathbb{R}^n$. De la definición de gradiente se tiene que

$$grad\, \widehat{f}(x) = 2\, A\, x.$$

El espacio tangente a $\mathcal{S}^{n-1}$, visto como subvariedad de $T_x\mathbb{R}^n \simeq \mathbb{R}^n$ es

$$T_x\mathcal{S}^{n-1} = \{v \in \mathbb{R}^n : x^T\, v = 0\}.$$

El espacio normal es

$$\left(T_x\mathcal{S}^{n-1}\right)^{\perp} = \{x\, \rho : \rho \in \mathbb{R}\}.$$

Las proyecciones ortogonales al espacio tangente y al normal son

$$\mathrm{P}_x\, v = v - x\, x^T\, v, \qquad \mathrm{P}_x^{\perp}\, v = x\, x^T\, v.$$

Recordemos que cualquier elemento $v \in T_x\mathbb{R}^n$ se puede escribir como la suma directa

$$v = \mathrm{P}_x\, v + \mathrm{P}_x^{\perp}\, v.$$

Ahora, ya que en general, si $\widehat{f}$ es la función costo sobre una variedad $\overline{M}$ y f su restricción a la subvariedad M, el gradiente de f es el gradiente de $\widehat{f}$ proyectado sobre T_xM, es decir,

$$grad\, f(x) = \mathrm{P}_x\, grad\, \widehat{f}(x).$$

Así que en nuestro caso

$$grad\, f(x) = 2\,\mathrm{P}_x(A\,x) = 2\left(A\,x - x\,x^T\,A\,x\right). \tag{4.33}$$

La retracción R_x sobre $T_x\mathbb{R}^n \simeq \mathbb{R}^n$ es definida como $R_x\, v = x + v$ así que en $T_x\mathcal{S}^{n-1}$, la podemos definir como

$$R_x\, v = \frac{x+v}{\|x+v\|}.$$

Pero ya que $\mathcal{S}^{n-1}$ pertenece al grupo de matrices ortogonales, una alternativa para la retracción también es

$$R_x\, v = q\, f(x+v),$$

donde $p\, f(A)$ denota el Q factor de la descomposición de $A \in \mathbb{R}_*^{n\times p}$ como $A = Q\,R$ donde Q pertenece a las $St(p,n) = \{X \in \mathbb{R}^{n\times p} : X^T\,X = I_p\}$ y R es una matriz triangular superior $n \times p$ con los elementos de la diagonal estrictamente positivos.

Todas las formulas dadas arriba están resumidas en el cuadro (4.1).

	Variedad ($\mathcal{S}^{n-1}$)	Espacio de embedding ($\mathbb{R}^n$)
Función Costo	$f(x) = x^T A\, x,\ x \in \mathcal{S}^{n-1}$	$\widehat{f}(x) = x^T A\, x,\ x \in \mathbb{R}^n$
Métrica	Métrica inducida	$\widehat{g}(u, v) = u^T\, v$
Espacio Tangente	$v \in \mathbb{R}^n : x^T\, v = 0$	$\mathbb{R}^n$
Espacio Normal	$v \in \mathbb{R}^n : v = \alpha\, x$	$\emptyset$
Proyección sobre el Espacio Tangente	$\mathrm{P}_x\, v = (I - x\, x^T)\, v$	Identidad
Gradiente	$grad\, f(x) = \mathrm{P}_x\, grad\, \widehat{f}(x)$	$grad\, \widehat{f}(x) = 2\, A\, x$
Retracción	$R_x\, v = q\, f(x + v)$	$R_x\, v = x + v$

Cuadro 4.1: Cociente de Rayleigh sobre la esfera unitaria

Las siguientes proposiciones nos hablan sobre los puntos críticos, mínimos, máximos y puntos sillas del cociente de Rayleigh. Las pruebas respectivas se pueden ver en [5].

Proposición 4.4.1 *Sea $A = A^T$ una matriz simétrica de orden $n \times n$. Un vector de norma unitaria $x \in \mathbb{R}^n$ es un autovector de A si y sólo si este es un punto crítico del cociente de Rayleigh.*

Proposición 4.4.2 *Sea $A = A^T$ una matriz simétrica de orden $n \times n$ con autovalores $\mu_1 \leqslant ... \leqslant \mu_n$ y autovectores ortonormales asociados $v_1, ..., v_n$. Entonces*

- *$\pm\, v_1$ son minimizadores locales y globales del Cociente de Rayleigh (4.31) si el autovalor μ_1 es simple (multiplicidad 1), con lo que serían los únicos minimizadores.*

- $\pm\, v_n$ *son máximizadores locales y globales del Cociente de Rayleigh (4.31) si el autovalor* μ_n *es simple, con lo que serían los únicos minimizadores.*

- $\pm\, v_t$ *son los autovectores correspondientes a los autovalores interiores(que están estrictamente entre* μ_1 *y* μ_n *), son los puntos sillas del Cociente de Rayleigh (4.31) .*

Se desprende de la Proposición (4.4.1) y el análisis de convergencia global del métodos de búsqueda lineal (Proposición 4.3.2), que todos los métodos dentro de la clase del **Algoritmo 1**, producen iteraciones que convergen en el conjunto de vectores propios de A. Y como estamos considerando el método de descenso, si el autovalor μ_1 es simple, se tendrá convergencia estable para $\pm\, v_1$ e inestable para todos los demás autovectores.

Luego, en el caso que nos interesa para el **Algoritmo 1**, vamos a colocar la dirección (gradiente-relacionada)como

$$d_k = -grad\, f(x_k) = -2\left(A\, x_k - x_k\, x_k^T\, A\, x_k\right).$$

En lo siguiente, para no complicar el asunto referente a la retracción ya comentada, tomemos

$$R_x\, v = \frac{x+v}{\|x+v\|}, \tag{4.34}$$

donde $\|\,,\,\|$ representa la norma euclidiana de $\mathbb{R}^n$ dada por,$\|\, z\,\| = \sqrt{z^T\, z}$. Otra posibilidad es

$$R_x\, v = x\, \cos\, \|\, v\,\| + \frac{v}{\|v\|}\, \sin\, \|\, v\,\|,$$

para la cual la curva $t \to R_x(t\,v)$ en un circulo grande de la esfera.

Bien, ya con todos los ingredientes necesario planteamos el **Algoritmo 1** en el siguiente cuadro: Este algoritmo presenta resultados numéricos ya probados (ver [5]), por ejemplo

---

Algoritmo 1* Búsqueda lineal de Armijo para el Cociente de Rayleigh sobre $\mathcal{S}^{n-1}$

Requiere: La matriz simétrica A, escalares $\widehat{\lambda} > 0$, β, $\sigma \in (0,1)$.

Entrada: Iterado inicial $x_0, \|x_0\|$.

Salida: Sucesión de iterados $\{x_k\}$.

1: **Para** $k = 0,\ 1,\ 2, ...$**hacer**

2: Calcular $d_k = -2\left(A\,x_k - x_k\,x_k^T\,A\,x_k\right)$.

3: Encontrar el entero más pequeño $m \geqslant 0$ tal que

$$f\big(R_{x_k}(\widehat{\lambda}\,\beta^m\,d_k)\big) \leqslant f\big(x_k\big) - \sigma\,\widehat{\lambda}\,\beta^m\,d_k^T\,d_k,$$

con f definida en (4.31) y R en (4.34).

4: Colocar

$$x_{k+1} = R_{x_k}(\widehat{\lambda}\,\beta^m\,d_k).$$

5: **Finalizar**

---

se puede tomar la matriz A como, $A = diag(1, 2, ..., 100), \sigma = 0{,}5, \lambda = 1$ y $\beta = 0{,}5$; el punto de inicio x_0 se escoge normalizando un vector cuya entradas son seleccionadas de una distribución normal. Pero el asunto aquí, es que éste mismo algoritmo, por la teoría expuesta, también presenta convergencia al agregar la segunda condición de Wolfe (4.13b);

inclusive la condición fuerte. El mismo quedaría de la forma siguiente:

Algoritmo 1* Búsqueda lineal de Armijo para el Cociente de Rayleigh sobre $\mathcal{S}^{n-1}$

Requiere: La matriz simétrica A, escalares $\widehat{\lambda} > 0$, β, $\sigma \in (0, 1)$.

Entrada: Iterado inicial $x_0, \|x_0\|$.

Salida: Sucesión de iterados $\{x_k\}$.

1: **Para** $k = 0, 1, 2, ...$**hacer**

2: Calcular $d_k = -2\left(A\,x_k - x_k\,x_k^T\,A\,x_k\right)$.

3: Encontrar el entero más pequeño $m \geqslant 0$ tal que $0 < c_1 = \sigma\,\beta^m < \dfrac{c_2}{m} < 1$

$$f\big(R_{x_k}(\widehat{\lambda}\,\beta^m\,d_k)\big) \leqslant f\big(x_k\big) - \sigma\,\widehat{\lambda}\,\beta^m\,d_k^T\,d_k,$$

y

$$D\,f\big(R_{x_k}(\widehat{\lambda}\,\beta^m\,d_k)\big)\,T^{R_{x_k}}_{x_k,\,R_{x_k}(\widehat{\lambda}\,\beta^m\,d_k)}\,d_k \leqslant c_2\,d_k^T\,d_k,$$

con f definida en (4.31), R en (4.34) y $T^{R_{x_k}}$ es el transporte vectorial escalar que representa la derivada $D\,R_{x_k}$ de la retracción R_{x_k} sobre M.

4: Colocar

$$x_{k+1} = R_{x_k}(\widehat{\lambda}\,\beta^m\,d_k).$$

5: **Finalizar**

[Se hace la siguiente observación, el valor de m dado en el punto de Armijo (4.3.2), no es fácil de conseguir. En general se toma de tal forma que $0 < c_1 = \sigma\,\beta^m < \dfrac{c_2}{m} < 1$.]

Conclusión

En cuanto al objetivo planteado en este trabajo, hemos presentado la búsqueda lineal con condición de Wolfe dada en $\mathbb{R}^n$, sobre variedades Riemannianas. Valiéndonos para ello, de instrumentos de la geometría diferencial como son las retracciones junto con el transporte vectorial paralelo, que nos permitieron caminar sobre la estructura no euclidiana e implementar las estrategias necesarias para conseguir convergencia en estos tipos de métodos. Por su puesto, se buscó destacar los detalles de las demostraciones que, en general, cuando se trata de variedades se opta por decir, que si se cumple en el $\mathbb{R}^n$ correspondiente, de igual forma debe ocurrir en la misma.

Bibliografía

[1] W. Ring and B. wirth, *Optimization Methods Riemannian Manifold and their application to shape space,* SIAMJ. Optim, 22(2):596-627, 2012.

[2] J. Nocedal and S. Wright, *Numerical Optimization* , Springer Series in Operations Research and Financial Engineering Springer,New York,Second Edition,2006.

[3] D.Gabay, *Minimizing a Diferentiable Function over a Diferential Manifold,* J. Optim. Theory Appl.,37(2):177-219, 1982.

[4] R.E. Mahony, *The Constrained Newton Method on a Lie Grupo and the Symmetric Eigenvalue Problem,* Linear Algebra Appl. 23(3):309-327, 2002.

[5] R. Mahony and J. Manton, *The Geometry of the Newton onnon-compact Lie Grupos,* J Global Optim. 248:367-89, 1996.

[6] In: A. Bloch(ed), *Halmition and Gradient Flows* , Algorithms and Control, Fields Institute Communication, Vol 3,Amer. Math. Soc,pp:113-136, 1994.

[7] C. Udriste, *Convex Funtions and Optimización Methods on Riemannian Manifold,* Kluwer Acad, Publ.,Pordre 1994) pp.1-8 Rio de Janeiro: Sociedad Brasilera de Matemítica, tercera edición (2005).

[8] C. Baker, P. Absil and K. Gallivan *An implicit trust-región method on Riemannian Manifold,* IMA-J. Number Anal, 28(4):665-207 (1988).

[9] Y. Yang, *Globally Convergent Opimization Algorithms on Riemannian Manifolds: Uniform framework for unconstrained and constrained optimization,* J. Optim. Theory Appl.,132(2): 245-265, 2007.

[10] M. Epelman, *Continuos Optimization Methods,* Section 1.IOE 511 . Math 652. Fall,2000.

[11] L. Gripo,F. Lampariello and S. Lucidi, *A Nonmonotone Line Search Technique For Newton's Method,* Istituto di analisi dei Sistemi ed Informatica del CNR (1985).

[12] F. Camargo, *Estudo Comparativo de Passos Espectrais e Buscas Lineares não monótonas* , Instituto De matemítica E Estatística Da Universidade De São Paulo (2008).

[13] E. Quiroz, E. Quispe and P. oliveira, *steepest Descent Method for Quasiconvex Minimization on Riemannian Manifold,* Federal University of Rio de Janeiro, PESC-COPPE (2006).

[14] S. do Carmo, Manfredo Perdigao, *Geometria Riemanniana,* Rio de Janeiro: IMPA-Projeto Euclides, segunda edición, (1988).

[15] Erik Quiroz, *Un método no Euclidiano para Problemas de Optimización en Espacios Euclidianos,* Universidad Nacional Del Callao (2007).

[16] A. Absil, R. Mahony and R. Sepulchre, *Riemanniana Geometry of Grassmann Manifold With a View on Algorithmic Computation,* Florida State University Tallahassee,FL 32306-4120,Usa (2000).

[17] BOOTHBY WILLIAM, *An Introduction to Differentiable Manifold and Riemannian Geometry,* Florida USA: Edit. Adacemic Press, First Edition, 1986.

[18] Lawrence Conlon, *Differentiable Manifold a First Course,* Birkhäuser Advanced Texs.Boston.Basel.Berlin: Edited by Herbert Amann,Zürich (1993).

[19] B. González,D. Benavides, *Nociones de Análisis Funcional,* UNIVERSIDAD DE SEVILLA.Departamento de Análisis Matemático.Sevilla,España (2010).