Primavera-2021

Apuntes de la Escuela de Ingeniería

MMD3702 - Optimización No-lineal

INSTITUTO DE CIENCIAS DE LA INGENIERÍA

David Salas V.

Universidad de O'Higgins

Aclaraciones sobre el apunte y agradecimientos

Se concede permiso para imprimir o almacenar copias de este documento a cualquier integrante de la Universidad de O'Higgins. Salvo por las excepciones más abajo señaladas, este permiso no autoriza fotocopiar o reproducir copias para otro uso que no sea el personal, o distribuir o dar acceso a copias electrónicas de este documento sin permiso previo por escrito del Director de la Escuela de Ingeniería de la Universidad de O'Higgins.

Las excepciones al permiso por escrito del párrafo anterior son:

1. Las copias electrónicas disponibles en ucampus.uoh.cl,
2. Las copias distribuidas por el cuerpo docente de Universidad de O'Higgins en el ejercicio de las funciones que le son propias.

Cualquier reproducción parcial de este documento debe hacer referencia a su fuente de origen.

Este documento fue confeccionado como material de estudio para el curso de Optimización No-Lineal MMD3702 de la Universidad de O'Higgins, sin fines de lucro. Está basado sobre los apuntes del curso dictado por el profesor David Salas V. en los semestres de Primavera 2020 y Primavera 2021.

Índice general

Universidad de O'Higgins

Universidad de O'Higgins

Capítulo 1

Introducción a la Optimización no-lineal

1.1 Problemas de Optimización y noción de solución

En este curso estamos interesados en resolver problemas de optimización, que tienen la forma general

$$(\mathcal{P}) := \begin{cases} \min\limits_{x\in\mathbb{R}^n} & f(x) \\ s.a. & x \in X. \end{cases} \tag{1.1}$$

El problema $(\mathcal{P})$ se conoce como *problema de minimización* y está compuesto por tres elementos:

1. La función objetivo $f : \mathbb{R}^n \to \mathbb{R}$. Es la función cuyo valor queremos minimizar.
2. La variable de decisión $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Es lo que controlamos. Podemos decidir el valor de x y por lo tanto no está predefinido.
3. El conjunto factible $X \subset \mathbb{R}^n$. La variable que controlamos está restringida a este conjunto. Solo podemos elegir valores para x que estén en el conjunto factible X.

En general, el conjunto factible X está descrito por una serie de *restricciones*, es decir,

$$X = \left\{ x \in \mathbb{R}^n \ : \ \begin{array}{l} h_i(x) = 0, \forall i \in I \\ g_j(x) \leq 0, \forall j \in J \end{array} \right\} \tag{1.2}$$

donde I y J son conjuntos de índices finitos, y $h_i : \mathbb{R}^n \to \mathbb{R}$ y $g_j : \mathbb{R}^n \to \mathbb{R}$ son funciones continuas para todo $i \in I$ y todo $j \in J$. Cuando tenemos esta estructura, el problema (1.1) se reescribe como

$$(\mathcal{P}) := \begin{cases} \min\limits_{x\in\mathbb{R}^n} & f(x) \\ s.a. & \begin{cases} h_i(x) = 0, & \forall i \in I, \\ g_j(x) \leq 0, & \forall j \in J. \end{cases} \end{cases} \tag{1.3}$$

En algunos problemas, aparecen restricciones de desigualdad de la forma $g(x) \geq \alpha$. Sin embargo, reemplazando la función g por $x \mapsto \alpha - g(x)$, siempre podemos expresar el conjunto factible X en su forma estándar dada por (1.2).

Universidad de O'Higgins

Volviendo a la forma general (1.1), la siguiente definición nos da el concepto de solución del problema $(\mathcal{P})$.

Definición 1.1 (Solución)**.** *Un punto $\bar{x} \in \mathbb{R}^n$ se dice* solución *del problema de optimización (1.1) si cumple que*

- *$\bar{x}$ es un punto factible del problema, es decir, $\bar{x} \in X$.*
- *$f(\bar{x})$ es el valor mínimo que puede alcanzar f en el conjunto factible X, es decir,*

$$\forall x \in X, \quad f(\bar{x}) \leq f(x). \tag{1.4}$$

El conjunto de todas las soluciones del problema (1.1) se denota como $\operatorname{argmin}\{f(x) \ : \ x \in X\}$*, o simplemente* $\operatorname{argmin}_X f$*. Los puntos $\bar{x} \in \operatorname{argmin}_X f$ también se les denomina* mínimos globales *de f en X.*

El concepto de mínimo global está dado por la ecuación (1.5): Un punto $\bar{x} \in X$ es mínimo global de f en X si $f(\bar{x})$ toma el valor más pequeño en todo X. Como veremos más adelante, la gran dificultad de la optimización no-lineal versus su contraparte lineal, es que encontrar mínimos globales en nuestro contexto no siempre es posible computacionalmente. Debido a esta obstrucción, muchos métodos de optimización buscan lo que se conoce como mínimos locales.

Definición 1.2 (Mínimo local)**.** *Un punto $\bar{x} \in \mathbb{R}^n$ se dice* mínimo local *de f en X cumple que*

- *$\bar{x}$ es un punto factible del problema, es decir, $\bar{x} \in X$.*
- *Existe $\delta > 0$ suficientemente pequeño tal que*

$$\forall x \in X \cap B(\bar{x}, \delta), \quad f(\bar{x}) \leq f(x). \tag{1.5}$$

Es claro que todo mínimo global es a su vez mínimo local, pero que la conversa no se tiene. La noción de mínimo local se puede interpretar como una "solución débil" del problema (1.1). La Figura 1.1 ilustra esta diferencia para la función $f : \mathbb{R}^2 \to \mathbb{R}$ dada por

$$f(x, y) = \ln(x^4 + 2(x-2)^3 + 100\exp(-x^2) + 300) + \ln(1 + y^2),$$

en el conjunto factible $X = \{(x, y) \ : \ -10 \leq x, y \leq 10\}$.

En este apunte, nos enfocaremos en cómo encontrar soluciones globales, o al menos locales, para el problema (1.1), cuando los datos del problema son no-lineales.

1.2 Problemas de maximización

En muchas aplicaciones, nos encontramos con otra familia de problemas, llamados *problemas de maximización*, que tienen la forma general

$$(\mathcal{P}) := \begin{cases} \max\limits_{x \in \mathbb{R}^n} & f(x) \\ s.a. & x \in X. \end{cases} \tag{1.6}$$

En este tipo de problemas tenemos exactamente los mismos elementos que en el (1.1), es decir, la función objetivo f, la variable de decisión $x \in \mathbb{R}^n$ y el conjunto factible X. La diferencia está en el concepto de solución, pues aquí queremos encontrar un punto $\bar{x} \in \mathbb{R}^n$ que sea factible pero tal que $f(\bar{x})$ sea el valor *máximo* que puede alcanzar f en X. Es decir, buscamos $\bar{x} \in X$ tal que

$$\forall x \in X, \quad f(\bar{x}) \geq f(x). \tag{1.7}$$

Universidad de O'Higgins

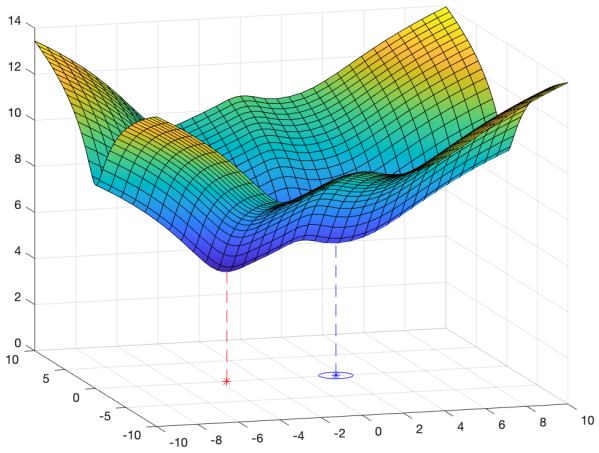


Figura 1.1: Ejemplo de Mínimo local versus Mínimo global. En rojo, el mínimo global de la función, alcanzado aproximadamente en $(-3.6173, 0)$. En azul, un óptimo local, alcanzado aproximadamente en $(1.6847, 0)$. El círculo azul es una bola donde se verifica (1.5).

El conjunto de todas las soluciones del problema (1.6) se denota como $\operatorname{argmax}\{f(x) \ : \ x \in X\}$ o simplemente $\operatorname{argmax}_X f$. Un elemento $\bar{x} \in \operatorname{argmax}_X f$ se llama también *máximo global* de f en X.

Análogamente, también podemos definir máximos locales: Un punto $\bar{x} \in X$ es un *máximo local* de f en X si existe $\delta > 0$ lo suficientemente pequeño tal que

$$\forall x \in X \cap B(\bar{x}, \delta),\ f(\bar{x}) \geq f(x). \tag{1.8}$$

Sin embargo, los problemas de maximización y minimización están relacionados, en el sentido que uno siempre puede transformarse en el otro manteniendo las mismas soluciones.

Teorema 1.3. *El problema de maximización (1.3) es equivalente al problema de minimización*

$$\begin{cases} \min\limits_{x \in \mathbb{R}^n} & -f(x) \\ s.a. & x \in X, \end{cases}$$

en el sentido que ambos problemas tienen las mismas soluciones fuertes y débiles, es decir,

- $\operatorname{argmin}_X(-f) = \operatorname{argmax}_X f$*; y*
- $\bar{x}$ *es máximo local de* f *en* X *si y sólo si* $\bar{x}$ *es mínimo local de* $-f$ *en* X*.*

Demostración. Mostraremos solamente que $\operatorname{argmin}_X(-f) = \operatorname{argmax}_X f$. Para esto, escribimos:

$$\begin{aligned} x^* \in \operatorname{argmin}_X(-f) &\iff \forall x \in X,\ -f(x^*) \leq -f(x) \\ &\iff \forall x \in X,\ f(x^*) \geq f(x) \\ &\iff x^* \in \operatorname{argmax}_X f. \end{aligned}$$

En el desarrollo anterior, la primera y la tercera equivalencia son las definiciones de solución para el problema de minimización de $-f$ y de maximización de f, respectivamente. La segunda equivalencia es directa del hecho que $-f(x^*) \leq -f(x)$ si y sólo si $f(x^*) \geq f(x)$.

La demostración de la segunda afirmación sobre óptimos locales es análoga, reemplazando X por $X \cap B(\bar{x}, \delta)$ y razonando por doble implicancia. □

Universidad de O'Higgins

El Teorema 1.3 nos dice que basta desarrollar la teoría que necesitamos para problemas de minimización solamente, pues siempre podemos transformar los problemas de maximización reemplazando la función objetivo f por $-f$. Recordemos que estamos interesados en encontrar soluciones en el sentido de la Definición 1.1, y por lo tanto podemos manipular los datos de los problemas, siempre y cuando preservemos el conjunto de soluciones. Así, en este apunte sólo trabajaremos con problemas de minimización.

1.3 Algunos ejemplos de Optimización No-Lineal

1.3.1. Modelo Uninodal de mercado eléctrico

En Chile, así como en muchas partes del mundo, el sistema de energía eléctrica es un mercado privado, donde varios productores participan en lo que se conoce como Mercado SPOT. En este, los productores declaran su capacidad y costos de producción a un Operador central (Conocido como Operador Independiente de Sistema), que debe asignar las producciones de cada productor satisfaciendo la demanda energética D a menor costo posible. Supongamos que en el mercado SPOT tenemos N productores $P_1, \dots, P_n$, y que cada productor P_i tiene una función de costos dada por

$$c_i(q_i) = a_i q_i + b_i q_i^2,$$

donde a_1, b_i son constantes, y q_i representa la potencia eléctrica (medida en MWh) que debe producir el productor P_i. Claramente, q_i debe ser no-negativa, pero además, el productor P_i tiene una capacidad máxima de producción, que denotaremos por Q_i. Con esta información, el Operador Central debe resolver el problema de optimización

$$\begin{cases} \min\limits_{q \in \mathbb{R}^N} & \sum_{i=1}^N a_i q_i + b_i q_i^2 \\ \\ s.a. & \sum_{i=1}^N q_i = D, \\ & 0 \leq q_i \leq Q_i, \quad \forall i \in \{1, \dots, N\}. \end{cases} \tag{1.9}$$

Este problema se conoce como *modelo uninodal*, pues no considera la transmisión de corriente eléctrica a través de las líneas de transmisión, y es el modelo más simple que existe para el problema del operador central. Como vemos, la función objetivo de este problema es no-lineal.

1.3.2. Ajuste no-lineal de parámetros (Data-Fitting)

En muchos problemas de estadística (incluido el Aprendizaje de Máquinas), queremos ajustar un modelo predictivo a partir de datos medidos. Este modelo predictivo muchas veces está dado por una función parametrizada

$$\begin{aligned} F : \mathbb{R}^p \times \mathbb{R}^n &\to \mathbb{R}^m \\ (\theta, x) &\mapsto F(\theta, x) = y, \end{aligned}$$

donde $x \in \mathbb{R}^n$ es la entrada, $y \in \mathbb{R}^m$ es la salida, y $\theta \in \mathbb{R}^p$ es el vector de parámetros que queremos ajustar.

Los datos están dados por un conjunto de pares ordenados $\{(x_k, y_k) \ : \ k = 1, \dots, N\} \subset \mathbb{R}^n \times \mathbb{R}^m$, dadas por mediciones empíricas: Para cada muestra $x_k \in \mathbb{R}^n$, se midió la salida $y_k \in \mathbb{R}^m$, pero sin conocer la función $F(\theta, \cdot)$.

Universidad de O'Higgins

El modelo más sencillo para ajuste de parámetros se conoce como método de *mínimos cuadrados*. Aquí, lo que queremos es encontrar el vector de parámetros θ^* que minimiza el error cuadrático medio del modelo predictivo con respecto a los datos empíricos. Es decir, queremos resolver

$$\min_{\theta \in \mathbb{R}^p} \sum_{k=1}^{N} \|F(\theta, x_k) - y_k\|^2. \tag{1.10}$$

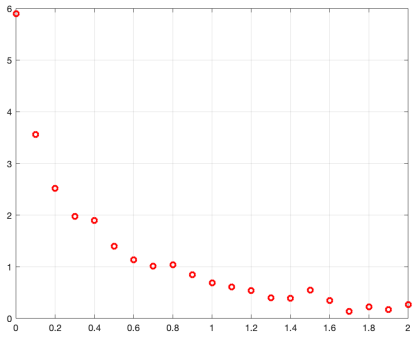
La Figura 1.2 muestra el ajuste de parámetros para la función $F : \mathbb{R}^4 \times \mathbb{R} \to \mathbb{R}$ dada por

$$F(\theta, x) = \theta_1 \exp(-\theta_2 x) + \theta_3 \exp(-\theta_4 x),$$

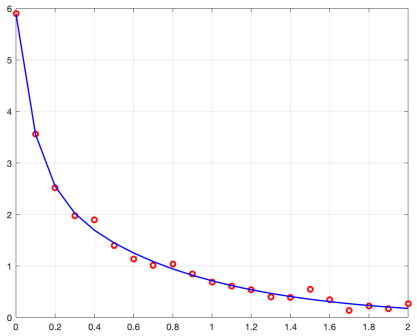
con los datos empíricos de la Tabla 1.1.

x	y
0.0000	5.8955
0.1000	3.5639
0.2000	2.5173
0.3000	1.9790
0.4000	1.8990
0.5000	1.3938
0.6000	1.1359
0.7000	1.0096
0.8000	1.0343
0.9000	0.8435
1.0000	0.6856
1.1000	0.6100
1.2000	0.5392
1.3000	0.3946
1.4000	0.3903
1.5000	0.5474
1.6000	0.3459
1.7000	0.1370
1.8000	0.2211
1.9000	0.1704
2.0000	0.2636

Tabla 1.1: Datos para ajuste por mínimos cuadrados



(a) Datos Emíricos



(b) Curva Ajustada

Figura 1.2: Ajuste de Datos por mínimos cuadrados

1.4 Preliminares de topología general

Esta sección está dedicada a repasar algunos conceptos fundamentales que ocuparemos en el desarrollo del curso. La mayoría de ellos forman parte del programa del curso *Cálculo Avanzado* (ING2001), o bien del programa del curso *Análisis para Ciencia de Datos* (MMD2002).

Antes de empezar, es bueno recordar algunas notaciones:

Universidad de O'Higgins

1. Para un punto $x \in \mathbb{R}^n$, escribimos $\|x\|$ para denotar su norma euclidiana, es decir,

$$\|x\| = \sqrt{\sum_{i=1}^{n} |x_i|^2}.$$

2. La bola abierta y la bola cerrada centrada en $x \in \mathbb{R}^n$ y de radio $\rho > 0$ están dadas, respectivamente, por

$$B(x,\rho) = \{y \in \mathbb{R}^n \ : \ \|y - x\| < \rho\},$$
$$\overline{B}(x,\rho) = \{y \in \mathbb{R}^n \ : \ \|y - x\| \leq \rho\}.$$

3. Denotamos por $\mathbb{B}$ la bola unitaria cerrada, es decir, $\mathbb{B} = \overline{B}(0,1)$. Similarmente, denotamos por $\mathbb{S}$ a la esfera unitaria, que es el conjunto dado por

$$\mathbb{S} = \{x \in \mathbb{R}^n \ : \ \|x\| = 1\} = \overline{B}(0,1) \setminus B(0,1).$$

4. Para un conjunto $A \subset \mathbb{R}^n$, denotamos su interior, adherencia y frontera por $\text{int}(A)$, $\overline{A}$ y $\text{Fr}(A)$, respectivamente.

Definición 1.4 (Sucesiones)**.** *Una sucesión en $\mathbb{R}^n$ es una función $x : \mathbb{N} \to \mathbb{R}^n$, donde a cada natural $k \in \mathbb{N}$ se asigna un punto $x_k = x(k)$ en el espacio. En lo que sigue, denotaremos la sucesión $x : \mathbb{N} \to \mathbb{R}^n$ como $(x_k)_{k\in\mathbb{N}}$, $(x_k)_k$ o simplemente (x_k).*

Dado que el conjunto $\mathbb{N}$ de los números naturales está naturalmente ordenado, la notación $(x_k)_{k\in\mathbb{N}}$ nos dice que podemos interpretar la sucesión como una secuencia de puntos, dada por $x_1, x_2, x_3, \ldots$.

Cabe destacar que la notación x_k puede ser ambigua: No es claro si nos referimos al k-ésimo elemento de una sucesión, o bien a la k-ésima coordenada de un vector en $\mathbb{R}^n$. Usualmente es claro por contexto a lo que nos estamos refiriendo cuando escribimos x_k, pero vale la pena estar atento. En caso que x_k sea un elemento de una sucesión, escribiremos $x_{k,i}$ para denotar la i-ésima coordenada de x_k, es decir,

$$x_k = \begin{pmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,n} \end{pmatrix} \in \mathbb{R}^n.$$

Las sucesiones están intrínsecamente relacionadas con la topología de $\mathbb{R}^n$ y serán la base sobre la cual construiremos algoritmos para aproximar soluciones de problemas de optimización.

Definición 1.5 (Convergencia)**.** *Decimos que una sucesión $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ converge a un punto $\bar{x} \in \mathbb{N}$ si se cumple que*

$$\forall \varepsilon > 0, \exists k_0 \in \mathbb{N}, \forall k \geq k_0, \|x_k - \bar{x}\| \leq \varepsilon. \tag{1.11}$$

*En tal caso, decimos que $\bar{x}$ es el **límite** de $(x_k)_{k\in\mathbb{N}}$, y escribimos $x_k \xrightarrow{k\to\infty} \bar{x}$ o bien $\bar{x} = \lim_{k\to\infty} x_k$.*

Cuando $x_k \to \bar{x}$, significa que para un error $\varepsilon > 0$ arbitrario, siempre existe un elemento $k_0 \in \mathbb{N}$ tal que la cola de la sucesión $(x_k)_{k\geq k_0}$ está a distancia a lo más ε del límite. La figura 1.3 ilustra esta idea para una sucesión en $\mathbb{R}^2$: Para el error ε, la sucesión en algún momento (es decir, a partir de k_0) queda contenida en la bola de centro $\bar{x}$ y radio ε.

Proposición 1.6. *Sea $(x_k)_{k\in\mathbb{N}}$ una sucesión en $\mathbb{R}^n$ y sea $\bar{x} \in \mathbb{R}^n$. Las siguientes afirmaciones son equivalentes:*

(i) $x_k \to \bar{x}$.

Universidad de O'Higgins

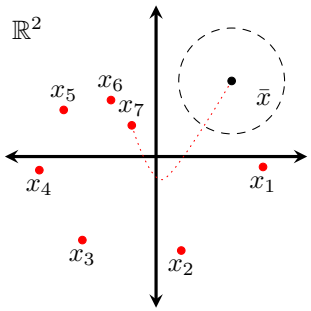


Figura 1.3: Sucesión convergente en $\mathbb{R}^2$

(ii) *Para todo* $\varepsilon > 0$, *existe* $k_0 \in \mathbb{N}$ *tal que* $(x_k)_{k \geq k_0} \subset B(\bar{x}, \varepsilon)$.

(iii) $\|x_k - \bar{x}\| \to 0$.

(iv) *Para todo índice* $i \in \{1, \ldots, n\}$, *la sucesión* $(x_{k,i})_{k \in \mathbb{N}} \subset \mathbb{R}$ *converge a* $\bar{x}_i$.

Demostración. La equivalencia $(i) \Leftrightarrow (ii) \Leftrightarrow (iii)$ es directa. Mostraremos que estas afirmaciones son equivalentes a (iv) por doble implicancia:

- $(iii) \Rightarrow (iv)$: Fijemos $i \in \{1, \ldots, n\}$ y sea $\varepsilon > 0$. Se tiene que

$$0 \leq |x_{k,i} - \bar{x}_i| \leq \|x_k - \bar{x}\| \xrightarrow{k \to \infty} 0.$$

  En vista del teorema del Sandwich, se concluye que $|x_{k,i} - \bar{x}_i| \xrightarrow{k \to \infty} 0$, lo que equivale a que $x_{k,i} \xrightarrow{k \to \infty} \bar{x}_i$. Como esto se tiene para cualquier coordenada $i \in \{1, \ldots, n\}$, se concluye (iv).

- $(iv) \Rightarrow (i)$: Sea $\varepsilon > 0$. Para cada $i \in \{1, \ldots, n\}$, tenemos que $x_{i,k} \to \bar{x}_i$. Por lo tanto, en vista de (1.11), existe $k_i \in \mathbb{N}$ tal que

$$\forall k \geq k_i,\ |x_{k,i} - \bar{x}_i| \leq \frac{\varepsilon}{n}.$$

  Sea $k_0 = \text{máx}\{k_i \ : \ i = 1, \ldots, n\}$. Tenemos entonces que

$$\forall k \geq k_0,\ r_k = \text{máx}\{|x_{k,i} - \bar{x}| \ : \ i = 1, \ldots, n\} \leq \frac{\varepsilon}{\sqrt{n}}.$$

  Luego, para todo $k \geq k_0$, podemos escribir

$$\|x_k - \bar{x}\| = \sqrt{\sum_{i=1}^{n} |x_{k,i} - \bar{x}_i|^2} \leq \sqrt{\sum_{i=1}^{n} r_k^2} \leq \sqrt{n \left(\frac{\varepsilon}{\sqrt{n}}\right)^2} = \varepsilon.$$

  Como ε es arbitrario, (1.11) se verifica, lo que demuestra que $x_k \xrightarrow{k \to \infty} \bar{x}$.

Hemos demostrado que $(i) \Leftrightarrow (ii) \Leftrightarrow (iii) \Rightarrow (iv) \Rightarrow (i)$, lo que concluye la demostración. □

Muchas veces, las sucesiones no poseen límite, es decir, divergen. Sin embargo, esta divergencia puede ocurrir pues las sucesiones están oscilando entre varios puntos candidatos a límites. Un ejemplo típico de esta situación es la sucesión (x_k) en $\mathbb{R}$ dada por $x_k = (-1)^k$. Para capturar este caso, se introduce la noción de punto de acumulación.

Universidad de O'Higgins

Definición 1.7 (Punto de acumulación)**.** *Sea $(x_k)_{k\in\mathbb{N}}$ una sucesión en $\mathbb{R}^n$ y $\bar{x}\in\mathbb{R}^n$. Decimos que $\bar{x}$ es punto de acumulación de $(x_k)_{k\in\mathbb{N}}$ si*

$$\forall \varepsilon > 0, \forall k_0 \in \mathbb{N}, \exists k \geq k_0, \|x - x_k\| \leq \varepsilon. \tag{1.12}$$

Denotamos el conjunto de todos los puntos de acumulación de $(x_k)_{k\in\mathbb{N}}$ como $\operatorname{acc}(x_k \ : \ k\in\mathbb{N})$, o simplemente como $\operatorname{acc}(x_k)$.

La idea de un punto de acumulación es que para un error $\varepsilon > 0$ arbitrario, y para cualquier índice $k_0 \in \mathbb{N}$, hay un punto de la sucesión x_k más adelante (es decir, $k \geq k_0$) que está a distancia a lo más ε del punto $\bar{x}$.

Sea $\bar{x} \in \operatorname{acc}(x_k)$. Si nos quedamos solo con los elementos que se van acercando a $\bar{x}$, tendríamos una sucesión que de hecho converge a $\bar{x}$. Esta idea de seleccionar algunos elementos de una sucesión para construir otra se formaliza con el concepto de subsucesiones.

Definición 1.8 (Subsucesiones)**.** *Sea $(x_k)_{k\in\mathbb{N}}$ una sucesión en $\mathbb{R}^n$. Una sucesión $(y_j)_{j\in\mathbb{N}}$ se dice subsucesión de $(x_k)_{k\in\mathbb{N}}$ si existe una función estrictamente creciente $\varphi : N \to \mathbb{N}$ tal que*

$$y_j = x_{\varphi(j)}, \quad \forall j \in \mathbb{N}.$$

Dicho de otro modo, la sucesión $y : \mathbb{N} \to \mathbb{R}^n$ está dada por $y = x \circ \varphi$.

Muchas veces, en vez de escribir $(y_j)_{j\in\mathbb{N}}$, escribimos directamente $(x_{\varphi(j)})_{j\in\mathbb{N}}$ para denotar la subsucesión de (x_k) dada por la función φ. También escribimos $(x_{k_j})_{j\in\mathbb{N}}$ en vez de $(x_{\varphi(j)})_{j\in\mathbb{N}}$, entendiendo que $k_j = \varphi(j)$. Esta última notación es útil para explicitar que la subsucesión está dada por una subsecuencia creciente de índices $\{k_j \ : \ j \in \mathbb{N}\}$. Cabe destacar que si $x_k \xrightarrow{k} \bar{x}$, entonces toda subscuesión (x_{k_j}) de (x_k) es convergente con $x_{k_j} \xrightarrow{j} \bar{x}$.

Proposición 1.9. *Sea $(x_k)_{k\in\mathbb{N}}$ una sucesión en $\mathbb{R}^n$ y $\bar{x}\in\mathbb{R}^n$. Se tiene que*

$$\bar{x} \in \operatorname{acc}(x_k) \iff \textit{Existe una subsucesión } (x_{k_j})_j \textit{ de } (x_k)_k \textit{ tal que } x_{k_j} \xrightarrow{j\to\infty} \bar{x}.$$

Demostración. Mostraremos el resultado por doble implicancia:

- $\Rightarrow$: Construiremos inductivamente una subsucesión $(x_{k_j})_j$ converegente a $\bar{x}$. Para el caso base, tomemos $j = 1$. Como $\bar{x} \in \operatorname{acc}(x_k)$, sabemos que

  $$\exists k_1 \geq 1 \text{ tal que } \|x_{k_1} - \bar{x}\| \leq 1.$$

  Para el paso inductivo, asumamos que hemos construido la colección $\{x_{k_1}, \ldots, x_{k_j}\}$ tal que $k_1 < k_2 < \ldots < k_j$ y que verifica que

  $$\|x_{k_t} - \bar{x}\| \leq \frac{1}{t}, \quad \forall t \in \{1, \ldots, j\}.$$

  Tomando $k_0 = 1 + k_j$ y $\varepsilon = \frac{1}{j+1}$, podemos aplicar (1.12) y concluir que

  $$\exists k_{j+1} > k_j, \ \|x_{k_{j+1}} - \bar{x}\| \leq \frac{1}{j+1}.$$

  Hemos construido inductivamente una sucesión $(x_{k_j})_j$ tal que $\|x_{k_j} - \bar{x}\| \leq \frac{1}{j}$, para todo $j \in \mathbb{N}$. Como además el mapeo $\varphi : j \to k_j$ es estrictamente creciente, concluimos que de hecho $(x_{k_j})_j$ es subsucesión de $(x_k)_k$. Finalmente, tenemos que

  $$0 \leq \|x_{k_j} - \bar{x}\| \leq \frac{1}{j} \xrightarrow{j\to\infty} 0,$$

  que, por Teorema del Sándwich, implica que $\|x_{k_j} - \bar{x}\| \to 0$. Aplicando la Proposición 1.6, concluimos que $x_{k_j} \to \bar{x}$.

Universidad de O'Higgins

- $\Leftarrow$: Supongamos que existe una subsucesión $(x_{k_j})_j$ convergiendo a $\bar{x}$. Sean $\varepsilon > 0$ y $k_0 \in \mathbb{N}$ fijos.

  Por un lado, como $x_{k_j} \to \bar{x}$, existe $j_1 \in \mathbb{N}$ tal que $\|x_{k_j} - \bar{x}\| \leq \varepsilon$, para todo $j \geq j_1$. Por el otro, como el mapeo $j \mapsto k_j$ es estrictamente creciente, existe $j_2 \in \mathbb{N}$ tal que $k_{j_2} \geq k_0$. Tomando $j_0 = \text{máx}\{j_1, j_2\}$, tenemos que

  $$k_{j_0} \geq k_0 \quad \text{y} \quad \|x_{k_{j_0}} - \bar{x}\| \leq \varepsilon.$$

  Como ε y k_0 son arbitrarios, concluimos que (1.12) se verifica, demostrando que $\bar{x} \in \text{acc}(x_k)$ y finalizando la demostración.

□

Para cerrar este repaso, recordaremos lo que son los conjuntos compactos, algunas de sus caracterizaciones, y enunciaremos el teorema fundamental de la optimización: Los problemas de optimización con función objetivo continua y conjunto factible compacto siempre tienen solución.

Recordemos que un conjunto $K \subset \mathbb{R}^n$ se dice *acotado* si existe $M > 0$ tal que $K \subset \overline{B}(0, M)$ o, equivalentemente, si

$$\forall x \in K, \|x\| \leq M.$$

Definición 1.10. *Un conjunto $K \subset \mathbb{R}^n$ se dice compacto si es no-vacío, cerrado y acotado.*

Vamos a ver que, así como los conjuntos cerrados se pueden caracterizar por sucesiones, los conjuntos compactos están relacionados con subsucesiones. Esta caracterización se conoce como el Teorema de Bolzano-Weierstrass. Para probarlo, necesitaremos el siguiente lema, que es la versión de una variable del teorema que queremos probar. Este lema fue demostrado en primera instancia por Bolzano en 1817 como parte de la demostración del teorema de valor intermedio.

Lema 1.11 (Bolzano)**.** *Sea $K \subset \mathbb{R}$ un conjunto acotado. Entonces toda sucesión en K admite una subsucesión convergente.*

Demostración. Como K es acotado, existe un intervalo $[a_1, b_1]$ tal que $K \subset [a_1, b_1]$. Ahora, construiremos la subsucesión $(x_{k_j})_{j \in \mathbb{N}}$ convergente.

- Para $j = 1$, tomamos $I_1 = [a_1, b_1]$ y elegimos $k_1 = 1$, sabiendo que $x_{k_1} \in I_1$.
- Para $j = 2$, dividimos el intervalo $I_1 = [a_1, b_1]$ en dos intervalos, dados por $\left[a_1, \frac{a_1+b_1}{2}\right]$ y $\left[\frac{a_1+b_1}{2}, b_1\right]$. Definimos I_2 como uno de los intervalos que contenga infinitos puntos de la sucesión (x_k) (al menos uno de los dos hace esto). Tomamos $k_2 > k_1$ tal que $x_{k_2} \in I_2$.
- Para el caso general $j \in \mathbb{N}$, asumamos que ya hemos construido $\{x_{k_1}, \ldots, x_{k_{j-1}}\}$ y los subintervalos $I_1, \ldots, I_{j-1}$. Nuevamente dividimos el intervalo I_{j-1} en dos intervalos a través de su punto medio. Definimos I_j como uno de los intervalos que contenga infinitos puntos de la sucesión (x_k), lo cual se puede hacer pues I_{j-1} contenía infinitos puntos.

Hemos construido así una secuencia de intervalos $(I_j : j \in \mathbb{N})$ y una subsucesión $(x_{k_j})_{j \in \mathbb{N}}$ con las siguientes propiedades:

1. Para todo $j \in \mathbb{N}$, $I_{j+1} \subset I_j$ y $\text{diam}(I_{j+1}) = \frac{1}{2}\text{diam}(I_j)$.
2. Para todo $j \in \mathbb{N}$, $x_{k_j} \in I_j$.

Consideremos ahora las sucesiones $(a_j)_{j \in \mathbb{N}}, (b_j)_{j \in \mathbb{N}} \subset \mathbb{R}$ dadas por $a_j = \text{mín}\{t : t \in I_j\}$ y $b_j = \text{máx}\{t : t \in I_j\}$. La inclusión $I_{j+1} \subset I_j$ para todo $j \in \mathbb{N}$ nos dice que (a_j) es una sucesión creciente y que (b_j) es una sucesión decreciente. Como ambas sucesiones son acotadas, tenemos que

Universidad de O'Higgins

son convergentes, con

$$a = \lim_j a_j = \sup_j a_j \qquad \text{y} \qquad b_j = \lim_j b_j = \inf_j b_j.$$

Más aún, se tiene que

$$|a-b| = \lim_j |a_j - b_j| = \lim_j \operatorname{diam}(I_j) = \lim_j \frac{\operatorname{diam}(I_1)}{2^{j-1}} = 0.$$

Por lo tanto, $a = b$. Finalmente, como $x_{k_j} \in I_j$, tenemos que $a_j \leq x_{k_j} \leq b_j$ para todo $j \in \mathbb{N}$. Por Teorema del Sándwich, concluimos que $x_{k_j} \to a$, lo que concluye la demostración. □

Teorema 1.12 (Bolzano-Weierstrass)**.** *Sea $K \subset \mathbb{R}^n$ un conjunto no-vacío. Las siguientes afirmaciones son equivalentes:*

(i) K es compacto.

(ii) Toda sucesión $(x_k)_{k\in\mathbb{N}}$ contenida en K admite una subsucesión $(x_{k_j})_{j\in\mathbb{N}}$ convergente con límite en K.

Demostración. Demostraremos el teorema por doble implicancia.

- $(i) \Rightarrow (ii)$: Demostraremos el resultado por inducción en la dimensión del espacio $\mathbb{R}^n$. El caso base está dado por el Lema 1.11. Ahora, para el caso general, tomemos $K \subset \mathbb{R}^n$ compacto y una sucesión $(x_k)_{k\in\mathbb{N}} \subset K$. Para cada x_k, podemos escribir $x_k = (y_k, z_k) \in \mathbb{R}^{n-1} \times \mathbb{R}$, donde $y_k = (x_{k,1}, \ldots, x_{k,n-1})$, y $z_k = x_{k,n}$. Por inducción, podemos asumir que $(y_k)_{k\in\mathbb{N}}$ admite una subsucesión convergente, pues es fácil ver que

  $$(y_k)_{k\in\mathbb{N}} \subset \overline{\{y \in \mathbb{R}^{n-1} \ : \ \exists z \in \mathbb{R}, (y,z) \in K\}},$$

  y este conjunto es compacto pues K lo es. Sea $\varphi_1 : \mathbb{N} \to \mathbb{N}$ la función estrictamente creciente tal que $(y_{\varphi_1(j)})_{j\in\mathbb{N}}$ es convergente. Consideremos la sucesión $(a_j)_{j\in\mathbb{N}} \subset \mathbb{R}$ dada por $a_j = z_{\varphi_1(j)}$. Aplicando el Lema 1.11 nuevamente, sabemos que (a_j) admite una subsucesión convergente. Sea $\varphi_2 : \mathbb{N} \to \mathbb{N}$ la función estrictamente creciente tal que $(a_{\varphi_2(j)})$ es convergente. Definamos $\varphi : \mathbb{N} \to \mathbb{N}$ como $\varphi_1 \circ \varphi_2$. Tenemos que φ también es estrictamente creciente y por lo tanto

  - $(y_{\varphi(j)}) = (y_{\varphi_1 \circ \varphi_2(j)})$ es una subsucesión de $(y_{\varphi_1(j)})$, y por lo tanto es convergente con $\lim_j y_{\varphi(j)} = \lim_j y_{\varphi_1(j)}$.
  - $(z_{\varphi(j)}) = (a_{\varphi_2(j)})$, que por construcción es convergente.

  Luego, aplicando la Proposición 1.6, tenemos que $\lim_j x_{\varphi(j)}$ existe y está dado por

  $$\bar{x} = \lim_j x_{\varphi(j)} = (\lim_j y_{\varphi(j)}, \lim_j z_{\varphi(j)}) \in K,$$

  donde la inclusión $\bar{x} \in K$ sigue del hecho que K es cerrado. Esto concluye la demostración de (ii).

- $(ii) \Rightarrow (i)$: Sea K un conjunto no-vacío que verifica (ii). Veamos primero que K es cerrado. Para esto, sea $(x_k)_k \subset K$ una sucesión convergente con límite $\lim_k x_k = \bar{x}$. Por hipótesis, sabemos que (x_k) admite una subsucesión $(x_{k_j})_j$ convergente con $\lim_j x_{k_j} \in K$. Sin embargo, como (x_{k_j}) es subsucesión de una sucesión convergente, necesariamente tenemos que

  $$\bar{x} = \lim_j x_{k_j} \in K.$$

  Como $(x_k)_k$ es arbitraria, hemos demostrado que toda sucesión convergente en K tiene su límite en K, lo que demuestra que K es cerrado.

Universidad de O'Higgins

Ahora, nos resta ver que K es acotado. Razonando por contradicción, asumamos que K no es acotado. Esto implica que para cada $k \in \mathbb{N}$, existe un punto $x_k \in K$ con $\|x_k\| \geq k$. Consideremos la sucesión $(x_k)_{k\in\mathbb{N}}$ dada por la construcción anterior. Por hipótesis, tenemos que $(x_k)_k$ admite una subsucesión $(x_{k_j})_j$ convergente. Sea $\bar{x} = \lim_j x_{k_j}$. Entonces, como la función $\|\cdot\|$ es continua, tenemos que

$$\|\bar{x}\| = \lim_j \|x_{k_j}\| \geq \lim_j k_j = +\infty,$$

lo cual es una contradicción. Concluimos que K tiene que ser acotado y, como ya vimos que también es cerrado, entonces K tiene que ser compacto. Esto concluye la demostración.

□

Terminaremos esta sección con lo que podríamos denominar el teorema fundamental de la optimización, que nos da condiciones suficientes para asegurar que un problema de optimización tiene solución. Este teorema también se conoce como Teorema de Weierstrass, o Teorema de valores extremos.

Teorema 1.13 (Teorema fundamental de la optimización)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función continua y $K \subset \mathbb{R}^n$ un conjunto compacto. Se tiene que f alcanza su máximo y su mínimo en K, es decir,*

$$\exists \bar{x} \in K,\ f(\bar{x}) = \min_K f \qquad y \qquad \exists \bar{z} \in K,\ f(\bar{z}) = \max_K f.$$

Demostración. Vamos a demostrar solamente que f alcanza su mínimo en K, pues el otro caso se obtiene directamente reemplazando f por $-f$. Sea $\alpha = \inf\{f(x) \ : \ x \in K\}$. Notemos que, como K es no-vacío, $\alpha < +\infty$. Por definición de ínfimo, para todo $k \in \mathbb{N}$ podemos encontrar $x_k \in K$ tal que

$$\begin{cases} f(x_k) \leq \alpha + \frac{1}{k} & \text{si } \alpha > -\infty \\ f(x_k) \leq -k & \text{si } \alpha = -\infty. \end{cases}$$

Consideremos la sucesión $(x_k)_{k\in\mathbb{N}}$ dada por la construcción anterior. Como K es compacto, (x_k) admite una subsucesión $(x_{k_j})_{j\in\mathbb{N}}$ convergente. Sea $\bar{x} = \lim_j x_{k_j}$, que pertenece a K pues K es cerrado.

Como f es continua, tenemos que $f(x_{k_j}) \xrightarrow{j\to\infty} f(\bar{x})$. Por lo tanto, si $\alpha = -\infty$ podemos escribir

$$-\infty \leq f(\bar{x}) = \lim_j f(x_{k_j}) \leq \lim_j -k_j = -\infty.$$

Esto nos diría que $f(\bar{x}) = -\infty$, lo cual es una contradicción pues f toma valores en $\mathbb{R}$. Concluimos entonces que $\alpha > -\infty$. Por lo tanto, podemos escribir,

$$\alpha \leq f(\bar{x}) = \lim_j f(x_{k_j}) \leq \lim_j (\alpha + \frac{1}{k_j}) = \alpha.$$

Se tiene entonces que $f(\bar{x}) = \alpha$, lo que concluye la demostración. □

1.5 Algoritmos de optimización

El Teorema 1.13 nos entrega un marco teórico para trabajar el problema de optimización genérico (1.1): Necesitamos que la función objetivo $f : \mathbb{R}^n \to \mathbb{R}$ sea al menos continua y que el conjunto factible X sea al menos compacto. Con estos dos ingredientes podemos asegurar existen soluciones,

Universidad de O'Higgins

en el sentido de la Definición 1.1. Sin embargo, el Teorema 1.13 no es constructivo, y por lo tanto no nos dice cómo encontrar dicha solución.

Para buscar soluciones de un problema de optimización, necesitamos desarrollar algoritmos. En optimización no-lineal, consideraremos lo que se conoce como *algoritmos iterativos*.

Definición 1.14 (Algoritmo Iterativo)**.** *Un algoritmo iterativo es una serie de instrucciones que, a partir de un punto inicial $x_0 \in \mathbb{R}^n$, nos permite generar una sucesión de puntos $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ con las siguientes propiedades:*

- *Para todo $k \in \mathbb{N}$, x_k se construye a partir de los puntos $\{x_0, x_1, \ldots, x_{k-1}\}$.*
- *La sucesión $(x_k)_{k\in\mathbb{N}}$ es convergente.*

En algoritmos iterativos, llamaremos al punto x_k el k-ésimo iterando*, pues se construye en la k-ésima iteración.*

Idealmente, cuando un algoritmo iterativo está bien adaptado a un problema de optimización, el límite $\bar{x} = \lim_k x_k$ es una solución (global, o al menos local). Esto depende fuertemente de los datos del problema (la función objetivo f y el conjunto factible X), por lo que existen diversos algoritmos para diferentes problemas.

Dado que la sucesión de iterandos (x_k) producida por un algoritmo iterativo es, en general, infinita, estos son incapaces de entregar el límite $\bar{x}$ final. Para hacerlo, un computador tendría que realizar infinitas iteraciones, lo cual no es posible. Por lo tanto, los algoritmos iterativos están acoplados con un *criterio de parada*, que es un conjunto de condiciones lógicas (verdadero-falso) sobre el iterando x_k. Cuando las condiciones del criterio de parada se cumplen en un iterando x_k, el algoritmo asume que este punto aproxima "lo suficientemente bien" al límite $\bar{x}$, y se detiene entregando x_k como solución aproximada. Normalmente un criterio de parada incluye las siguientes condiciones:

- El punto x_k se parece al punto x_{k-1} (la k-ésima iteración produjo poca variación del iterando).
- El valor $f(x_k)$ se parece al valor $f(x_{k-1})$ (la k-ésima iteración produjo poca mejora de la función objetivo).
- El punto x_k está lo suficientemente cerca del conjunto factible X.

El Algoritmo 1.1 muestra la estructura general de un algoritmo iterativo con criterio de parada.

Algoritmo 1.1: Esquema General de Algoritmos Iterativos

Entrada: Función objetivo f; Conjunto factible X; punto inicial $x_0 \in \mathbb{R}^n$.
Fijar $k = 0$;
while *CriterioParada(x_k) == FALSE* **do**
    - Construir x_{k+1} a partir de $\{x_0, x_1, \ldots, x_k\}$.
    - Actualizar $k = k + 1$.
end
Salida: $\bar{x} = x_k$.

Qué tan bueno es un algoritmo de optimización se mide utilizando los siguientes criterios:

1. **Robustez:** Un buen algoritmo debería funcionar para varios problemas de optimización. La robustez se evalúa identificando las hipótesis sobre los datos del problema (f, X y x_0) para los cuales el algoritmo produce una buena aproximación de una solución (fuerte o débil).
2. **Precisión:** Un buen algoritmo debería ser estable, en el sentido que entrega buenas aproximaciones de solución, sin ser demasiado sensible a variaciones en los datos del problema (f, X y x_0) o errores computacionales asociados a su ejecución computacional.
3. **Eficiencia:** Un buen algoritmo debería ser eficiente, en el sentido que el tiempo de cómputo

Universidad de O'Higgins

y el espacio de memoria computacional no debieran ser muy grandes.

Para estudiar la eficiencia de un algoritmo, nos concentraremos en el "tiempo de cómputo" que necesita, dependiendo de los datos del problema f, X y x_0. Como el tiempo depende del computador en donde el algoritmo se ejecuta, consideraremos una medición "normalizada": contaremos el número de iteraciones que necesita un algoritmo para satisfacer los criterios de parada.

Definición 1.15. *(Radio de convergencia) Sea $(x_k)_{k\in\mathbb{N}}$ una sucesión en $\mathbb{R}^n$ convergente a un punto $\bar{x} \in \mathbb{R}^n$. Definimos el* radio de convergencia de orden p *de $(x_k)_{k\in\mathbb{N}}$ como la sucesión*

$$r_{k+1}^p = \frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|^p}.$$

Diremos que la sucesión $(x_k)_{k\in\mathbb{N}}$ converge con

1. ***radio de convergencia lineal**, si existe una constante $\beta \in (0,1)$ tal que*

$$r_k^1 \leq \beta, \qquad \text{para } k \text{ lo suficientemente grande.}$$

2. ***radio de convergencia superlineal**, si*

$$\lim_k r_k^1 = 0.$$

3. ***radio de convergencia de orden p**, con $p > 1$, si existe $M > 0$ tal que*

$$r_k^p \leq M, \qquad \text{para } k \text{ lo suficientemente grande.}$$

La complejidad (tiempo de cómputo) de un algoritmo se mide calculando los siguientes indicadores:

- El número de iteraciones necesarias de tal manera que se cumpla el criterio de parada. Este número se estima usando los radios de convergencia de la sucesión de iterandos (x_k) y de la sucesión de valores (v_k), donde $v_k = f(x_k)$.
- El tiempo necesario para computar el iterando x_k en la iteración k. Esto depende normalmente del tamaño del problema (número de variables y número de restricciones).

En este curso, nos concentraremos sólo en el primer criterio, es decir, en el número de iteraciones que necesitamos para cumplir el criterio de parada.

1.6 Ejercicios Capítulo 1

P1. Para una función $f : \mathbb{R}^n \to \mathbb{R}$ y $\alpha \in \mathbb{R}$ definimos el subnivel de f de valor α como el conjunto

$$[f \leq \alpha] := \{x \in \mathbb{R}^n \ : \ f(x) \leq \alpha\}.$$

Muestre que si f es continua, entonces para todo $\alpha > \text{ínf}_{\mathbb{R}^n} f$ se tiene que $[f \leq \alpha]$ es cerrado.

P2. Una función $f : \mathbb{R}^n \to \mathbb{R}$ se dice *inf-compacta* si existe $\alpha > \text{ínf}_{\mathbb{R}^n} f$ tal que el subnivel $[f \leq \alpha]$ es un conjunto compacto. Muestre que si una función $f : \mathbb{R}^n \to \mathbb{R}$ es continua e inf-compacta, entonces alcanza su mínimo en $\mathbb{R}^n$.

P3. Una función $f : \mathbb{R}^n \to \mathbb{R}$ se dice *coercitiva* si

$$\lim_{\|x\|\to+\infty} f(x) = +\infty.$$

Universidad de O'Higgins

Muestre que si una función f es coercitiva y continua, entonces alcanza su mínimo en $\mathbb{R}^n$.
Hint: Recuerde que

$$\lim_{\|x\|\to+\infty} f(x) = +\infty \iff \begin{array}{c}\forall (x_k) \text{ sucesión en } \mathbb{R}^n \text{ tal que } \|x_k\| \to +\infty, \\ \text{se tiene que } f(x_k) \to +\infty.\end{array}$$

Utilice la caracterización anterior para mostrar que cualquier subnivel $[f \leq \alpha]$ debe ser un conjunto acotado.

P4. Sea $K \subset \mathbb{R}^n$ un conjunto no-vacío, $f : K \to \mathbb{R}$ y $x \in K$. Muestre que si toda sucesión $(x_k) \subset K$ convergente a x admite una subsucesión (x_{k_j}) tal que $f(x_{k_j}) \xrightarrow{j} f(x)$, entonces f es continua en x.

Universidad de O'Higgins

Capítulo 2

Elementos de Análisis Convexo

2.1 Conjuntos convexos y funciones convexas

Dentro del análisis no-lineal, la convexidad es una de las propiedades geométricas más deseables al momento de resolver un problema de optimización. A lo largo del curso, haremos siempre la diferencia entre problemas convexos y problemas no-convexos.

Definición 2.1 (Convexidad)**.** *Un conjunto $C \subset \mathbb{R}^n$ se dice* convexo *si para todo $x, y \in C$ y para todo $t \in [0,1]$, se tiene que*

$$tx + (1-t)y \in C. \tag{2.1}$$

Si C es un conjunto convexo, una función $f : C \to \mathbb{R}$ se dice convexa *si para todo $x, y \in C$ y para todo $t \in [0,1]$,*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y). \tag{2.2}$$

La definición anterior nos dice que un conjunto C es convexo si, para cualquier par de puntos $x, y \in C$, el segmento

$$[x, y] = \{tx + (1-t)y \ : \ t \in [0,1]\}$$

está contenido en C. La Figura 2.1 ilustra la idea anterior.

(a) Conjunto convexo (b) Conjunto no-convexo

Figura 2.1: (a) Conjunto convexo: El segmento $[x, y]$ queda contenido en el conjunto. (b) Conjunto no-convexo: El segmento $[x, y]$ se sale del conjunto.

Ejemplo 2.2 Los siguientes conjuntos son convexos:

1. Toda bola abierta $B(x, \rho)$ o cerrada $\overline{B}(x, \rho)$.

Universidad de O'Higgins

2. El conjunto vacío $\emptyset \subset \mathbb{R}^n$ y el espacio completo $\mathbb{R}^n$ son convexos.
3. Todo conjunto dado por restricciones lineales, es decir, de la forma $P = \{x \in \mathbb{R}^n \ : \ Ax \leq b\}$.
4. Todo subespacio vectorial H de $\mathbb{R}^n$.

Las siguientes funciones son convexas:

1. Toda función lineal $\ell : \mathbb{R}^n \to \mathbb{R}$ es convexa.
2. La norma euclidiana $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ es convexa.
3. La 1-norma y la ∞-norma dadas por

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$
$$\|x\|_\infty = \text{máx}\{|x_i| \ : \ i = 1, \ldots, n\}$$

   son funciones convexas.
4. La función $f : (0, +\infty) \to \mathbb{R}$ dada por $f(x) = \frac{1}{x}$ es convexa. La función exponencial $\exp : \mathbb{R} \to \mathbb{R}$ es convexa.

La convexidad de una función también tiene una interpretación geométrica: El conjunto dado por "lo que está sobre el grafo de la función" debe ser convexo. La Proposición 2.3 formaliza esta idea usando el concepto del *epígrafo* de f. la Figura 2.2 ilustra la convexidad de funciones.

Proposición 2.3. *Sea $C \subset \mathbb{R}^n$ un conjunto convexo. Una función $f : C \to \mathbb{R}$ es convexa si y solamente si su epígrafo, dado por el conjunto*

$$\text{epi}f := \{(x, r) \in \mathbb{R}^n \times \mathbb{R} \ : \ x \in C, \ f(x) \leq r\}, \tag{2.3}$$

es un conjunto convexo en $\mathbb{R}^{n+1}$.

Demostración. Haremos la demostración por doble implicancia:

- $\Rightarrow$: Sean $(x_1, r_1), (x_2, r_2) \in \text{epi}f$ y sea $t \in [0, 1]$. Queremos demostrar que

$$(x_t, r_t) := t(x_1, r_1) + (1 - t)(x_2, r_2)$$

  también está en epi$f$. Usando la definición de epígrafo, tenemos que $r_1 \geq f(x_1)$ y que $r_2 \geq f(x_2)$. Por lo tanto, notando que $tx + (1 - t)y \in C$, podemos escribir

$$r_t = tr_1 + (1 - t)r_2 \geq tf(x_1) + (1 - t)f(x_2) \geq f(tx_1 + (1 - t)x_2) = f(x_t),$$

  donde la segunda desigualdad se obtiene del hecho que f es convexa. Concluimos entonces que $(x_t, r_t) \in \text{epi}f$. Como (x_1, r_1), (x_2, r_2) y $t \in [0, 1]$ son arbitrarios, concluimos que epi$f$ es un conjunto convexo.
- $\Leftarrow$: Sean $x, y \in C$ y $t \in [0, 1]$. Queremos demostrar que $f(tx + (1 - t)y)$ es menor o igual a $tf(x) + (1 - t)f(y)$. Para esto, notemos que $(x, f(x))$ e $(y, f(y))$ ambos pertenecen al epígrafo de f. Como epi$f$ es convexo, tenemos que

$$(x_t, r_t) = t(x, f(x)) + (1 - t)(y, f(y)) \in \text{epi}f.$$

  Por lo tanto, podemos escribir

$$f(tx + (1 - t)y) = f(x_t) \leq r_t = tf(x) + (1 - t)f(y).$$

  Como x, y y t son arbitrarios, concluimos que f es convexa.

Universidad de O'Higgins

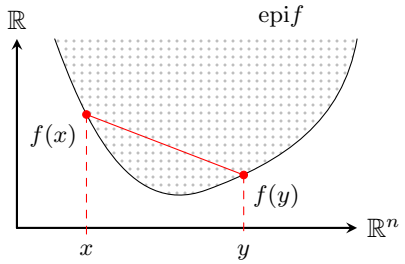


Figura 2.2: Función convexa. El segmento rojo corresponde a $[(x, f(x)), (y, f(y))]$ en $\mathbb{R}^{n+1}$, o equivalentemente, al grafo de la función $t \in [0,1] \mapsto tf(x) + (1-t)f(y)$. El segmento está por sobre el grafo de la función f. La zona punteada corresponde al epígrafo de f.

□

Para construir conjuntos convexos (y funciones convexas) necesitamos entender primero qué tipo de operaciones preservan la convexidad. La siguiente proposición resume algunas de estas operaciones.

Proposición 2.4. *Las siguientes operaciones de conjuntos preservan convexidad:*

1. *Si $\{C_i \ : \ i \in I\}$ es una familia cualquiera de conjuntos convexos en $\mathbb{R}^n$, entonces*

$$C = \bigcap_{i \in I} C_i$$

   también es convexo en $\mathbb{R}^n$.

2. *Si $C \subset \mathbb{R}^n$ es un conjunto convexo y $T : \mathbb{R}^n \to \mathbb{R}^m$ es una transformación lineal, entonces $T(C)$ es un conjunto convexo en $\mathbb{R}^m$.*

3. *Si A, B son dos conjuntos convexos en $\mathbb{R}^n$, entonces la suma de Minkowski*

$$A + B = \{a + b \ : \ a \in A, \, b \in B\}$$

   es un conjunto convexo en $\mathbb{R}^n$.

4. *Si $f : \mathbb{R}^n \to \mathbb{R}$ es una función convexa y $C \subset \mathbb{R}^n$ es un conjunto convexo, entonces los conjuntos*

$$C \cap [f \leq \alpha] = \{x \in C \ : \ f(x) \leq \alpha\}$$
$$C \cap [f < \alpha] = \{x \in C \ : \ f(x) < \alpha\}$$

   son conjuntos convexos en $\mathbb{R}^n$, para todo $\alpha \in \mathbb{R}$.

5. *Si $\ell : \mathbb{R}^n \to \mathbb{R}$ es una función lineal y $C \subset \mathbb{R}^n$ es un conjunto convexo, entonces*

$$C \cap [\ell = \alpha] = \{x \in C \ : \ \ell(x) = \alpha\}$$

   es un conjunto convexo en $\mathbb{R}^n$, para todo $\alpha \in \mathbb{R}$.

Demostración. Demostraremos cada afirmación por separado.

1. Sean $x, y \in C = \bigcap_{i \in I} C_i$ y sea $t \in [0,1]$. Como $x, y \in C$, entonces tenemos que

$$x, y \in C_i, \qquad \text{para todo } i \in I.$$

   Sea $i \in I$. Como C_i es convexo, tenemos que $tx + (1-t)y \in C_i$. Como esto es cierto para todo $i \in I$, concluimos que $tx + (1-t)y \in \bigcap_{i \in I} C_i = C$, lo que demuestra que C es convexo.

Universidad de O'Higgins

2. Sean $x, y \in T(C)$ y $t \in [0,1]$. Como $T(C)$ es la imagen de C a través de T, sabemos que existen $a, b \in C$ tal que $T(a) = x$ y $T(b) = y$. Luego, como T es lineal, podemos escribir

$$tx + (1-t)y = tT(a) + (1-t)T(b) = T(ta + (1-t)b).$$

Finalmente, como C es convexo, tenemos que $ta+(1-t)b \in C$ y por lo tanto $tx+(1-t)y \in T(C)$. Esto demuestra que $T(C)$ es convexo.

3. Sean $x, y \in A + B$ y $t \in [0,1]$. Por la definición de $A + B$, sabemos que existen $a_1, a_2 \in A$ y $b_1, b_2 \in B$ tal que $x = a_1 + b_1$ y $y = a_2 + b_2$. Luego, podemos escribir

$$\begin{aligned} tx + (1-t)y &= t(a_1 + b_1) + (1-t)(a_2 + b_2) \\ &= \underbrace{ta_1 + (1-t)a_2}_{=a} + \underbrace{tb_1 + (1-t)b_2}_{=b}. \end{aligned}$$

Como A y B son conjuntos convexos, tenemos que $a \in A$ y $b \in B$, por lo tanto

$$tx + (1-t)y = a + b \in A + B,$$

lo que demuestra que $A + B$ es convexo.

4. Sea $\alpha \in \mathbb{R}$. Si $C \cap [f \leq \alpha]$ es vacío, entonces es un conjunto convexo trivialmente. Por lo tanto, sin perder generalidad, asumamos que $C \cap [f \leq \alpha]$ es no-vacío. Sean entonces $x, y \in C \cap [f \leq \alpha]$ y $t \in [0,1]$. Como C es convexo, tenemos que $tx + (1-t)y \in C$. Por otro lado, como f es convexa, tenemos que

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \leq t\alpha + (1-t)\alpha = \alpha,$$

donde la segunda desigualdad sigue del hecho que $x, y \in [f \leq \alpha]$. Concluimos entonces que $f(tx+(1-t)y) \leq \alpha$ y por lo tanto $tx+(1-t)y \in [f \leq \alpha]$. Concluimos que $tx+(1-t)y \in C \cap [f \leq \alpha]$, lo que demuestra la convexidad de este conjunto. La demostración de que $C \cap [f < \alpha]$ es convexo es análoga.

5. Sean $x, y \in C \cap [\ell = \alpha]$ y $t \in [0,1]$. Como C es convexo, tenemos que $tx + (1-t)y \in C$. Por lo tanto, para demostrar que $C \cap [\ell = \alpha]$ es convexo, basta ver que $tx + (1-t)y \in [\ell = \alpha]$. Sin embargo, esto es directo pues, usando la linealidad de ℓ, podemos escribir

$$\ell(tx + (1-t)y) = t\ell(x) + (1-t)\ell(y) = t\alpha + (1-t)\alpha = \alpha.$$

Esto concluye la demostración.

□

Vale la pena destacar que, mezclando los incisos 1., 4. y 5. de la Proposición 2.4, podemos decir que los conjuntos de la forma

$$X = \left\{ x \in \mathbb{R}^n \ : \ \begin{array}{l} h_i(x) = 0, \ \forall i \in I \\ g_j(x) \leq 0, \ , \forall j \in J \end{array} \right\}$$

son convexos cuando las funciones $\{g_j \ : \ j \in J\}$ son convexas y las funciones $\{h_i \ : \ i \in I\}$ son lineales.

Así como hay operaciones que preservan convexidad de conjuntos, también existen operaciones que preservan la convexidad de funciones. La siguiente proposición nos muestra algunas de esas operaciones.

Proposición 2.5. *Las siguientes funciones son convexas:*

Universidad de O'Higgins

1. *Si* $T : \mathbb{R}^n \to \mathbb{R}^m$ *es una transformación lineal y* $g : \mathbb{R}^m \to \mathbb{R}$ *es una función convexa, entonces* $f : \mathbb{R}^n \to \mathbb{R}$ *dada por* $f = g \circ T$ *es una función convexa.*

2. *Sea* $C \subset \mathbb{R}^n$ *convexo. Si* $\{f_i : C \to \mathbb{R} \mid i = 1, \dots, k\}$ *es una familia de funciones convexas, entonces el* máximo puntual *dado por*

$$\begin{aligned} \max f_i : C &\to \mathbb{R} \\ x &\mapsto \max\{f_i(x) \ : \ i = 1, \dots, k\} \end{aligned}$$

es una función convexa.

3. *Sea* $C \subset \mathbb{R}^n$ *convexo. Si* $\{f_i : C \to \mathbb{R} \mid i = 1, \dots, k\}$ *es una familia finita de funciones convexas y* $\{\lambda_1, \dots, \lambda_k\} \subset \mathbb{R}_+$ *es una colección de escalares no-negativos, entonces*

$$f = \sum_{i=1}^{k} \lambda_i f_i$$

es una función convexa.

4. *Sean* $C \subset \mathbb{R}^n$ *y* $Z \subset \mathbb{R}^m$ *dos conjuntos convexos, y sea* $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *es una función convexa tal que para todo* $x \in C$ *se tiene que* $\inf_Z F(x, z) > -\infty$. *Entonces la* función marginal

$$\begin{aligned} f : \mathbb{R}^n &\to \mathbb{R} \\ x &\mapsto \inf_{z \in Z} F(x, z) \end{aligned}$$

es convexa.

Demostración. Vamos a demostrar cada caso por separado.

1. Sean $x, y \in \mathbb{R}^n$ y $t \in [0, 1]$. Ocupando la linealidad de T y la convexidad de g, podemos escribir

$$\begin{aligned} f(tx + (1-t)y) &= g\big(T(tx + (1-t)y\big) \\ &= g\big(tT(x) + (1-t)T(y)\big) \\ &\leq tg(T(x)) + (1-t)g(T(y)) = tf(x) + (1-t)f(y). \end{aligned}$$

   Como $x, y \in \mathbb{R}^n$ y $t \in [0, 1]$ son aribitrarios, se concluye que $f = g \circ T$ es convexa.

2. Denotemos $f = \max f_i$. Mostraremos, en vista de la Proposición 2.3, que $\mathrm{epi} f$ es convexo. Para esto, podemos ver que

$$\begin{aligned} (x, r) \in \mathrm{epi} f &\Leftrightarrow x \in C, \text{ y } r \geq f(x) \\ &\Leftrightarrow x \in C, \text{ y } \forall i \in \{1, \dots, k\},\ r \geq f_i(x) \\ &\Leftrightarrow \forall i \in \{1, \dots, k\},\ (x, r) \in \mathrm{epi} f_i \\ &\Leftrightarrow (x, r) \in \bigcap_{i=1}^{k} \mathrm{epi} f_i. \end{aligned}$$

   La serie de equivalencias anterior nos permite concluir que $\mathrm{epi} f = \bigcap_{i=1}^{k} \mathrm{epi} f_i$. Como las funciones $\{f_i \mid i = 1, \dots, k\}$ son convexas, entonces sus epígrafos son conjuntos convexos. Así, aplicando la Proposición 2.4, tenemos que $\mathrm{epi} f$ es convexo al ser intersección de conjuntos convexos. Concluimos entonces que f es convexa.

Universidad de O'Higgins

3. Sean $x, y \in C$ y $t \in [0,1]$. Usando que las funciones $\{f_i \mid i = 1, \ldots, k\}$ son convexas y que $\{\lambda_i \ : \ i = 1, \ldots, k\}$ son no-negativos, tenemos que

$$\begin{aligned} f(tx + (1-t)y) &= \sum_{i=1}^{k} \lambda_i f_i(tx + (1-t)y) \\ &\leq \sum_{i=1}^{k} \lambda_i \big(t f_i(x) + (1-t) f_i(y)\big) \\ &= t \sum_{i=1}^{k} \lambda_i f_i(x) + (1-t) \sum_{i=1}^{k} \lambda_i f_i(y) \\ &= t f(x) + (1-t) f(y). \end{aligned}$$

Como $x, y \in C$ y $t \in [0,1]$ son arbitrarios, concluimos que f es convexa.

4. Sean $x, y \in C$ y $t \in [0,1]$. Queremos demostrar que $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$. Fijemos $\varepsilon > 0$, y sean $z_1, z_2 \in Z$ tal que

$$f(x) = \inf_{Z} F(x,z) \geq F(x, z_1) - \varepsilon \qquad \text{y} \qquad f(y) = \inf_{Z} F(y,z) \geq F(y, z_2) - \varepsilon.$$

Como Z es convexo, tenemos que $tz_1 + (1-t)z_2 \in Z$. Luego, como F es convexa, podemos escribir

$$\begin{aligned} f(tx + (1-t)y) &= \inf_{Z} F(tx + (1-t)y, z) \\ &\leq F(tx + (1-t)y, tz_1 + (1-t)z_2) \\ &= F(t(x, z_1) + (1-t)(y, z_2)) \\ &\leq tF(x, z_1) + (1-t)F(y, z_2) \\ &\leq t(f(x) + \varepsilon) + (1-t)(f(y) + \varepsilon) = tf(x) + (1-t)f(y) + \varepsilon. \end{aligned}$$

Como ε es arbitrario, concluimos que

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

Como $x, y \in \mathbb{R}^n$ y $t \in [0,1]$ son arbitrarios, se concluye que la función marginal f es convexa.

□

2.2 Propiedades topológicas heredadas de la convexidad

En este curso, trabajaremos muchas veces con conjuntos que tienen interior vacío. Sin embargo, en el caso de los conjuntos convexos, estos siempre tienen interior, pero relativo a un espacio afín de dimensión más pequeña. Este espacio se conoce como la envoltura afín de un conjunto, y se construye intersectando todos los espacios afines que contienen al conjunto.

Para formalizar esta idea, recordemos primero lo que es un espacio afín.

Definición 2.6 (Espacio afín)**.** *Un conjunto H en $\mathbb{R}^n$ se dice espacio afín si existe un subespacio vectorial V y un punto $x_0 \in \mathbb{R}^n$ tal que*

$$H = x_0 + V = \{x_0 + v \ : \ v \in V\}.$$

El subespacio vectorial V se denomina subespacio paralelo a H y es único. Definimos entonces la dimensión afín de H como la dimensión de V, es decir, $\operatorname{adim}(H) = \dim(V)$.

Universidad de O'Higgins

De hecho, si H es un espacio afín con subespacio paralelo V, entonces se tiene que

$$\forall h_0 \in H, \quad H = h_0 + V. \tag{2.4}$$

De hecho, para cualquier punto $h_0 \in H$, se tiene que $V = H - h_0 = \{h - h_0 \ : \ h \in H\}$. La Figura 2.3 muestra un espacio afín en $\mathbb{R}^2$. Note que un subespacio vectorial es un espacio afín que contiene al 0.

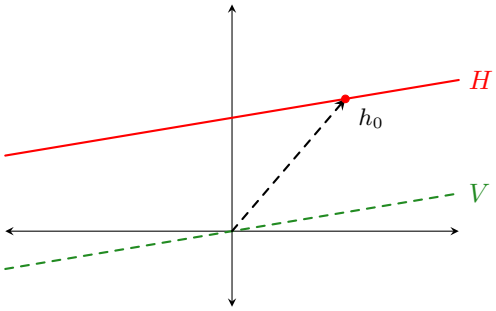


Figura 2.3: Espacio Afín en $\mathbb{R}^2$. H se obtiene desplazando V a través del vector h_0.

Proposición 2.7. *Sea $\{H_i \ : \ i \in I\}$ una familia cualquiera de espacios afines en $\mathbb{R}^n$. Si $H = \cap_{i\in I} H_i$ es no-vacío, entonces H es un espacio afín.*

Demostración. Sea $h_0 \in H$. Para cada $i \in I$, tenemos que $V_i = H_i - h_0$ es el subespacio paralelo a H_i. Sea $V = \bigcap_{i\in I} V_i$. Como V es una intersección de espacios vectoriales, entonces V es también un subespacio vectorial. Luego, podemos escribir

$$\begin{aligned} h \in H &\Leftrightarrow h \in H_i, \quad \forall i \in I \\ &\Leftrightarrow h - h_0 \in V_i, \quad \forall i \in I \\ &\Leftrightarrow h - h_0 \in V \\ &\Leftrightarrow h \in h_0 + V. \end{aligned}$$

Concluimos que $H = h_0 + V$, lo que demuestra que H es un espacio afín con espacio paralelo V. □

Observación 2.8. *Es importante destacar que para un conjunto cualquiera $A \subset \mathbb{R}^n$ y para un punto $x_0 \in \mathbb{R}^n$, siempre se tiene que*

$$(A + x_0) - x_0 = A.$$

Esto se tiene pues al sumar x_0 estamos desplazando el conjunto A. Esta operación, que la podemos pensar como la transformación lineal afín $T : \mathbb{R}^n \to \mathbb{R}^n$ dada por $T(x) = x + x_0$, es invertible, precisamente tomando $T^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ como $T^{-1}(x) = x - x_0$. Sin embargo, en general, para dos conjuntos $A, B \subset \mathbb{R}^n$ no podemos cancelar la suma de Minkowski, es decir,

$$(A + B) - B = \{a + b_1 - b_2 \ : \ a \in A, b_1, b_2 \in B\} \neq A.$$

La Proposición 2.7 nos permite introducir la definición de envoltura afín como el espacio afín "más pequeño"que contiene al conjunto.

Definición 2.9 (Envoltura Afín)**.** *Sea K un conjunto no-vacío de $\mathbb{R}^n$. Definimos la envoltura afín de K como la intersección de todos los espacios afines que contienen a K, es decir,*

$$\text{aff}(K) = \bigcap \{H \ : \ H \text{ espacio afín, } K \subset H\}. \tag{2.5}$$

Universidad de O'Higgins

Como la intersección del lado derecho es una intersección no-vacía de espacios afines, entonces $\text{aff}(K)$ *es también un espacio afín y es el más pequeño que contiene a* K*, es decir,*

$$H \text{ espacio afín con } K \subset H \implies \text{aff}(K) \subset H.$$

Con la definición de envoltura afín, podemos introducir ahora la noción de interior que necesitamos.

Definición 2.10 (Interior relativo)**.** *Sea* H *un espacio afín de* $\mathbb{R}^n$ *y* $A \subset H$*. Decimos que* A *es abierto relativo a* H *si*

$$\forall x \in A, \exists \rho > 0, B(x, \rho) \cap H \subset A. \tag{2.6}$$

Sea K *un conjunto no-vacío de* $\mathbb{R}^n$*. Definimos la el interior relativo de* K*, denotado por* $\text{rint}(K)$*, como el interior de* K *con respecto a* $\text{aff}(K)$*, es decir,*

$$x \in \text{rint}(K) \Leftrightarrow \exists \rho > 0, B(x, \rho) \cap \text{aff}(K) \subset K. \tag{2.7}$$

La idea de abierto relativo e interior relativo se ilustra en la Figura 2.4. Para un espacio afín, podemos mirar este conjunto como el "nuevo espacio completo", y considerar conjuntos abiertos e interiores con respecto a él. El interior relativo de un conjunto K es simplemente el interior de K cuando consideramos $\text{aff}(K)$ como el "nuevo espacio completo". Cabe destacar entonces que $\text{rint}(K)$ se puede definir alternativamente como "el abierto relativo a $\text{aff}(K)$ más grande contenido en K", es decir,

$$\text{rint}(K) = \bigcup \{A \ : \ A \text{ abierto relativo a } \text{aff}(K),\ A \subset K\}. \tag{2.8}$$

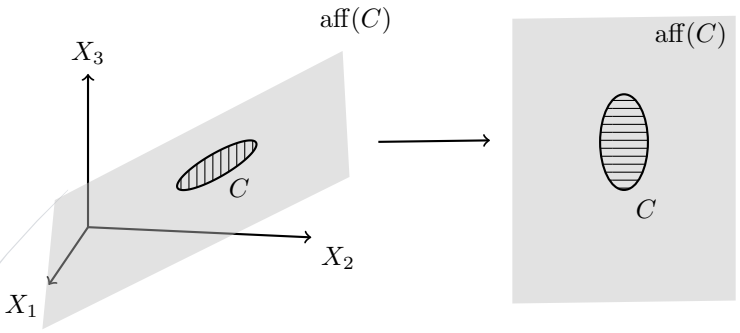


Figura 2.4: Conjunto convexo y envoltura afín en $\mathbb{R}^3$. La dimensión de $\text{aff}(C)$ es una menos que el espacio completo. En $\text{aff}(C)$, el conjunto C tiene interior no-vacío.

Proposición 2.11. *Si* C *es un conjunto convexo no-vacío de* $\mathbb{R}^n$*, entonces,* $\text{rint}(C)$ *es no-vacío.*

Demostración. Sea $x_0 \in C$ y sea V el subespacio paralelo a $\text{aff}(C)$. Definamos $K = C - x_0$. Es fácil ver que $V = \text{aff}(K)$. Si $\dim(V) = 0$, eso quiere decir que $C = \text{aff}(C) = \{x_0\}$, y por lo tanto tenemos que $\text{rint}(C) = \{x_0\}$. Si $\dim(V) = m > 0$, entonces existen $z_1, \ldots, z_m \in K$ linealmente independientes tal que $\text{span}\{z_1, \ldots, z_m\} = V$. En efecto, esta base se puede construir inductivamente:

- Como $\dim(V) > 0$, existe $z_1 \in K$ distinto de 0. Si $\dim(V) = 1$, entonces la construcción termina.
- Para el paso inductivo, supongamos que hemos construido $z_1, \ldots, z_{k-1} \in K$ linealmente independientes y que $\dim(V) > k - 1$. Entonces, $V_{k-1} = \text{span}\{z_1, \ldots, z_{k-1}\}$ no contiene a a K, pues de lo contrario $V = \text{aff}(K) \subset V_{k-1}$ y $\dim(V) \leq k - 1$. Por lo tanto existe $z_k \in K \setminus V_{k-1}$. Se tiene entonces que $\{z_1, \ldots, z_k\}$ es linealmente independiente y $V_k = \text{span}\{z_1, \ldots, z_k\} \subset V$. Si $\dim(V) = k$, entonces la construcción termina, pues necesariamente $V = V_k$.

Universidad de O'Higgins

- Repitiendo esta construcción hasta $k = m$, terminamos con el conjunto $\{z_1, \ldots, z_m\}$ deseado.

Ahora, sea

$$X = \left\{ x = \sum_{i=1}^{m} \lambda_i z_i \ : \ \lambda_1, \ldots, \lambda_m > 0 \text{ y } \sum_{i=1}^{m} \lambda_i < 1 \right\}.$$

Claramente $X \subset K$, pues K es convexo, $0 \in K$ y los elementos de X son combinaciones convexas de la forma

$$\sum_{i=1}^{m} \lambda_i z_i + \left(1 - \sum_{i=1}^{m} \lambda_i\right) 0.$$

Demostraremos que X es un conjunto abierto de V. Tomemos $Z = \begin{bmatrix} z_1 & \cdots & z_m \end{bmatrix}$. Tenemos que $T_Z : \mathbb{R}^m \to V$ dada por $T_Z(\lambda) = Z\lambda = \sum_{i=1}^{m} \lambda_i z_i$ es una transformación lineal biyectiva y por lo tanto, existe una constante $R > 0$ tal que

$$\|T_Z^{-1}(x)\| \leq R\|x\|, \quad \forall x \in V \tag{2.9}$$

(Basta notar que $(V, \|\cdot\|)$ es una espacio normado de dimensión finita, lo que implica que la función lineal $T_Z^{-1} : V \to \mathbb{R}^m$ es continua, y por lo tanto acotada en el sentido que verifica (2.9)). Sea ahora $\bar{x} \in X$. Entonces, $\bar{x} = Z\bar{\lambda}$ para algún $\bar{\lambda}$ en

$$A = \left\{ \lambda \in \mathbb{R}^m \ : \ \lambda_1, \ldots, \lambda_m > 0 \text{ y } \sum_{i=1}^{m} \lambda_i < 1 \right\}.$$

Como A es un conjunto abierto en $\mathbb{R}^m$, existe $\rho > 0$ tal que $B(\bar{\lambda}, \rho) \subset A$. Luego, para todo $x \in B(\bar{x}, \rho/R) \cap V$, tenemos que

$$\|T_Z^{-1}(x) - \bar{\lambda}\| = \|T_Z^{-1}(x) - T^{-1}(\bar{x})\| \leq R\|x - \bar{x}\| < \rho.$$

Por lo tanto, $T_Z^{-1}(x) \in A$. Finalmente, notando que $T_Z^{-1}(X) = A$, concluimos que

$$B(\bar{x}, \rho/R) \cap V = T_Z\Big(T_Z^{-1}\big(B(\bar{x}, \rho/R) \cap V\big)\Big) \subset T_Z(A) = X.$$

Concluimos así que X es abierto en V y por lo tanto, $X \subset \operatorname{rint}(K)$. La demostración termina notando que $\operatorname{rint}(C) = \operatorname{rint}(K) + x_0$. □

Teorema 2.12. *Sea C un conjunto convexo no-vacío de $\mathbb{R}^n$. Se tiene que*

1. *Los conjuntos $\operatorname{rint}(C)$ y $\overline{C}$ son convexos.*
2. *Para todo $x_0 \in \operatorname{rint}(C)$ y todo $x \in \overline{C}$, el segmento semiabierto*

$$[x_0, x) = \{tx + (1-t)x_0 \ : \ t \in [0, 1)\}$$

   está contenido en $\operatorname{rint}(C)$.
3. *Se tiene que $\overline{C} = \overline{\operatorname{rint}(C)}$ y que $\operatorname{rint}(\overline{C}) = \operatorname{rint}(C)$.*

Demostración. Demostraremos cada afirmación por separado.

1. Veamos que $\operatorname{rint}(C)$ es convexo. Sean $x_1, x_2 \in \operatorname{rint}(C)$ y $t \in [0, 1]$. Como $x_1, x_2 \in \operatorname{rint}(C)$, existen $\rho_1, \rho_2 > 0$ tal que

$$B(x_i, \rho_i) \cap \operatorname{aff}(C) \subset C, \quad \text{para } i = 1, 2.$$

   Tomemos $\rho = \min\{\rho_1, \rho_2\}$. Sea V el subespacio paralelo a $\operatorname{aff}(C)$. Se tiene que

$$B(x_i, \rho) \cap \operatorname{aff}(C) = B(0, \rho) \cap V + x_i, \quad \text{para } i = 1, 2.$$

Universidad de O'Higgins

Denotemos $x_t = tx_1 + (1-t)x_2$ y veamos que $B(x_t, \rho) \cap \text{aff}(C) \subset C$. En efecto, si $y \in B(x_t, \rho) \cap \text{aff}(C)$, entonces existe $v \in B(0, \rho) \cap V$ tal que $y = x_t + v$. Luego, notando que $x_i + v \in B(x_i, \rho) \cap \text{aff}(C)$ para $i = 1, 2$, podemos escribir

$$y = x_t + v = tx_1 + (1-t)x_2 + v = t(x_1 + v) + (1-t)(x_2 + v) \in C.$$

Concluimos entonces que $B(x_t, \rho) \cap \text{aff}(C) \subset C$, y por lo tanto $x_t \in \text{rint}(C)$. Esto muestra que $\text{rint}(C)$ es convexo.

Ahora veamos que $\overline{C}$ es convexo. Sean $x_1, x_2 \in \overline{C}$ y sea $t \in [0, 1]$. Usando la caracterización de la adherencia por sucesiones, sabemos que existen dos sucesiones $(x_{1,k})_k, (x_{2,k})_k \subset C$ tal que $x_{1,k} \to x_1$ y $x_{2,k} \to x_2$. Luego, podemos escribir

$$tx_1 + (1-t)x_2 = \lim_k \underbrace{(tx_{1,k} + (1-t)x_{2,k})}_{\in C} \in \overline{C}.$$

Concluimos entonces que $\overline{C}$ es convexo.

2. Sea $x_0 \in \text{rint}(C)$ y $x \in \overline{C}$. Sea $t \in [0, 1)$ y sea $x_t = tx + (1-t)x_0$. Queremos demostrar que $x_t \in \text{rint}(C)$.

   Supongamos primero que $x \in C$. Sea V el subespacio paralelo a $\text{aff}(C)$. Sabemos que existe $\rho > 0$ tal que

   $$B(x_0, \rho) \cap \text{aff}(C) = x_0 + B(0, \rho) \cap V \subset C.$$

   Vamos a demostrar que $B(x_t, (1-t)\rho) \cap \text{aff}(C) \subset C$. Sea $y \in B(x_t, (1-t)\rho) \cap \text{aff}(C)$. Se tiene que existe $v \in B(0, \rho)$ tal que $y = x_t + (1-t)v$. Luego, notando que $x_0 + v \in B(x_0, \rho) \cap \text{aff}(C) \subset C$, podemos escribir

   $$y = x_t + (1-t)v = tx + (1-t)x_0 + (1-t)v = tx + (1-t)(x_0 + v) \in C.$$

   Concluimos entonces que $y \in C$, lo que demuestra que $B(x_t, (1-t)\rho) \cap \text{aff}(C) \subset C$. Así, $x_t \in \text{rint}(C)$, como queríamos demostrar.

   Para el caso general $x \in \overline{C}$, consideremos una sucesión $(x_k) \subset C$ tal que $x_k \to x$. Por la parte anterior, sabemos que

   $$\forall k \in \mathbb{N}, B(tx_k + (1-t)x_0, (1-t)\rho) \cap \text{aff}(C) \subset C.$$

   Ahora, sea $k \in \mathbb{N}$ lo suficientemente grande tal que $\|x_t - (tx_k + (1-t)x_0)\| < (1-t)\rho/3$. Tenemos entonces que

   $$B(x_t, (1-t)\rho/3) \cap \text{aff}(C) \subset B(tx_k + (1-t)x_0, (1-t)\rho) \cap \text{aff}(C) \subset C,$$

   y por lo tanto, $x_t \in \text{rint}(C)$. Esto concluye la demostración.

3. Claramente $\overline{\text{rint}(C)} \subset \overline{C}$. Vamos a demostrar la otra inclusión. Sea $x \in \overline{C}$. Para ver que $x \in \overline{\text{rint}(C)}$, nos basta encontrar una sucesión $(x_k) \subset \text{rint}(C)$ tal que $x_k \to x$. Sea $(y_k) \subset C$ tal que $y_k \to x$ y sea $x_0 \in \text{rint}(C)$. Por la parte anterior, sabemos que

   $$x_k = \left(1 - \tfrac{1}{k}\right) y_k + \tfrac{1}{k} x_0 \in \text{rint}(C).$$

   Luego, como la sucesión (y_k) es convergente, tenemos que $\lim_k \frac{1}{k} y_k = 0$ y por lo tanto, podemos escribir

   $$\lim_k x_k = \lim_k y_k + \underbrace{\lim_k \tfrac{1}{k}(y_k - x_0)}_{=0} = x.$$

Universidad de O'Higgins

Concluimos entonces que $x \in \overline{\text{rint}(C)}$, que era lo que queríamos demostrar.

Ahora, veamos que $\text{rint}(\overline{C}) = \text{rint}(C)$. Claramente, se tiene que $\text{rint}(C) \subset \text{rint}(\overline{C})$, por lo que solo necesitamos demostrar la inclusión reversa. Sea $x \in \text{rint}(\overline{C})$. Queremos demostrar que $x \in \text{rint}(C)$. Por definición, existe $\rho > 0$ tal que $B(x, \rho) \cap \text{aff}(\overline{C}) \subset \overline{C}$. Como $\text{aff}(C)$ es cerrado y contiene a C, sabemos que $\overline{C} \subset \text{aff}(C)$ y por lo tanto, tenemos que

$$\text{aff}(C) \subset \text{aff}(\overline{C}) \subset \text{aff}(C),$$

es decir, $\text{aff}(C) = \text{aff}(\overline{C})$. Sea ahora $x_0 \in \text{rint}(C)$ y $d = x - x_0$. Si $d = 0$, entonces $x_0 = x$ y por lo tanto $x \in \text{rint}(C)$, lo que concluiría la demostración. Si $d \neq 0$, entonces $d \in V$, donde V es el subespacio paralelo a $\text{aff}(C)$. Luego, existe $\alpha > 0$ lo suficientemente pequeño tal que $\alpha d \in B(0, \rho)$ y por lo tanto

$$x_1 = x + \alpha d \in B(x, \rho) \cap \text{aff}(C) \subset \overline{C}.$$

Ahora, podemos escribir

$$\frac{\alpha}{1+\alpha} x_0 + \frac{1}{1+\alpha} x_1 = \frac{1}{1+\alpha}(\alpha x_0 + x + \alpha d) = \frac{1}{1+\alpha}((1+\alpha)x) = x.$$

Por lo tanto, x se puede escribir como combinación convexa de $x_1 \in \overline{C}$ y $x_0 \in \text{rint}(C)$. Usando la parte 2., concluimos que $x \in \text{rint}(C)$, lo que demuestra que $\text{rint}(\overline{C}) \subset \text{rint}(C)$.

□

El interior relativo de un conjunto convexo nos permite concluir una de las propiedades topológicas más importantes de la convexidad: Las funciones convexas son continuas en el interior relativo de sus dominios.

Teorema 2.13. *Sea $C \subset \mathbb{R}^n$ un conjunto convexo y $f : C \to \mathbb{R}$ una función convexa. Entonces, f es localmente acotada en $\text{rint}(C)$, es decir, para todo $x \in \text{rint}(C)$, existen $M, \rho > 0$ tales que*

$$|f(y)| \leq M, \quad \forall y \in B(x, \rho) \cap \text{rint}(C).$$

Más aún, f es localmente Lipschitz en $\text{rint}(C)$, es decir, para todo $x \in \text{rint}(C)$ existen constantes $L, r > 0$ tal que

$$|f(y_1) - f(y_2)| \leq L\|y_1 - y_2\|, \quad \forall y_1, y_2 \in B(x, r) \cap \text{rint}(C).$$

En particular, f es continua en $\text{rint}(C)$.

Demostración. Veamos primero que f es localmente acotada en $\text{rint}(C)$. Sea $\bar{x} \in \text{rint}(C)$. Reemplazando C por $C - \bar{x}$ y f por $g(x) = f(x + \bar{x})$, podemos suponer, sin perder generalidad, que $\bar{x} = 0$ y que $V = \text{aff}(C)$ es subespacio vectorial de $\mathbb{R}^n$. Luego, como $0 \in \text{rint}(C)$, existe $\rho_0 > 0$ tal que $B(0, \rho_0) \cap V \subset C$. Ahora, tomando $\rho = \frac{\rho_0}{4\sqrt{n}}$ podemos verificar que

$$B(0, 2\rho) \subset A = \{x \in \mathbb{R}^n \ : \ \|x\|_\infty \leq 2\rho\} \subset B(0, \rho_0).$$

y por lo tanto, $P = A \cap V \subset B(x, \rho) \cap V$. Como P es un polítopo convexo (está dado por un número finito de restricciones lineales), tenemos que P tiene un conjunto finito vértices $\{x_1, \ldots, x_k\}$ y además

$$P = \left\{ \sum_{i=1}^{k} \lambda_i x_i \ : \ \lambda_1, \ldots, \lambda_k \geq 0, \sum_{i=1}^{k} \lambda_i = 1 \right\}.$$

Universidad de O'Higgins

Luego, ocupando la desigualdad de Jensen (ver Problema **P2** de la Sección 2.6) tenemos que

$$\forall x \in P,\ f(x) = f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leq \sum_{i=1}^{k} \lambda_i f(x_i) \leq \underbrace{\text{máx}\{f(x_i)\ :\ i = 1, \ldots, k\}}_{=M_1}.$$

Concluimos que f es acotada superiormente en P, es decir, $f(x) \leq M_1$ para todo $x \in P$. Ahora, veamos que f es acotada inferiormente en $B(0,\rho) \cap V$, es decir, que $M_2 = \text{ínf}_{B(0,\rho)\cap V} f > -\infty$. Si no lo fuera, entonces existiría una sucesión $(x_k) \subset B(0,\rho) \cap V$ tal que $f(x_k) \leq -k^2$. Como A es compacto y $B(0,\rho) \subset A$, reemplazando (x_k) por una subsucesión si es necesario, podemos asumir sin perder generalidad que (x_k) es convergente, digamos con $\lim_k x_k = z \in A$ Más aún, $\|z\| \leq \rho$. Luego, tomando $y_k = z + (1/k)(z - x_k)$ es fácil verificar que para $k \in \mathbb{N}$ lo suficientemente grande, $y_k \in B(0, 2\rho) \cap V \subset P$ y además,

$$z = \frac{1}{1+1/k} y_k + \frac{1/k}{1+1/k} x_k.$$

Luego, podemos escribir,

$$f(z) \leq \frac{1}{1+1/k} f(y_k) + \frac{1/k}{1+1/k} f(x_k) \leq M_1 + \frac{1/k}{1+1/k}(-k^2) \leq M_1 - \frac{k}{2} \to -\infty.$$

Concluimos que $f(z) = -\infty$, lo cual es una contradicción. Por lo tanto, $M_2 > -\infty$. Tomando $M = \text{máx}\{M_1, -M_2\}$, concluimos que

$$\forall y \in B(0,\rho) \cap V, \quad |f(y)| \leq M.$$

Para la segunda parte, vamos a demostrar que f es Lipschitz en $B(0,r) \cap V$ con $r = \rho/2$. Sean $x, y \in B(0,r) \cap V$ con $x \neq y$, y sea $\alpha = \|x - y\|$. Como $y - x \in V$, podemos definir

$$z = y + \frac{r}{\alpha}(y - x) \in B(0, 2r) \cap V.$$

Ahora, podemos escribir y como combinación convexa de x y z. Concretamente,

$$\frac{\alpha}{r+\alpha} z + \frac{r}{r+\alpha} x = \frac{\alpha}{r+\alpha}\left(y + \frac{r}{\alpha}(y - x)\right) + \frac{r}{r+\alpha} x = y.$$

Por lo tanto, como f es convexa, concluimos que

$$f(y) \leq \frac{\alpha}{r+\alpha} f(z) + \frac{r}{r+\alpha} f(x).$$

Luego, restando $f(x)$ a ambos lados de la desigualdad, y recordando que $f(x), f(z) \leq M$ concluimos que

$$\begin{aligned} f(y) - f(x) &\leq \frac{\alpha}{\alpha + r}(f(z) - f(x)) \\ &\leq \alpha\left(\frac{|f(x)| + |f(z)|}{r}\right) \\ &\leq \frac{2M}{r}\alpha = \underbrace{\frac{2M}{r}}_{=L} \|x - y\|. \end{aligned}$$

Intercambiando los roles de x e y en el desarrollo anterior, concluimos que $f(x) - f(y) \leq L\|x - y\|$ y por lo tanto,

$$|f(x) - f(y)| \leq L\|x - y\|.$$

Como $x, y \in B(0,r) \cap V$ eran arbitrarios, concluimos el resultado. □

Universidad de O'Higgins

2.3 Distancia a conjuntos convexos y proyección métrica

En análisis convexo y en optimización, un problema recurrente es tratar de encontrar el punto "más cercano" en un conjunto K a un punto de referencia $x \in \mathbb{R}^n$. Cuando interpretamos K como el conjunto factible, lo que estamos buscando es la mejor aproximación factible de x. Recordando que la distancia entre dos puntos se mide como la norma de su diferencia, este problema se traduce en resolver

$$\begin{cases} \text{mín} & \|z - x\| \\ s.a. & z \in K. \end{cases} \tag{2.10}$$

El problema anterior se conoce como problema de mejor aproximación, y guarda profunda relación con la geometría del conjunto K. Cuando hacemos variar el punto de referencia x, el valor de la solución del problema de mejor aproximación se conoce como función distancia a K.

Definición 2.14 (Distancia a un conjunto)**.** *Sea $K \subset \mathbb{R}^n$ un conjunto no-vacío. Se define la función* ***distancia a*** *K como*

$$\begin{aligned} d_K : \mathbb{R}^n &\to \mathbb{R} \\ x &\mapsto d_K(x) = \inf_{z \in K} \|z - x\|. \end{aligned}$$

Cuando sea conveniente, ocuparemos la notación alternativa $d(\cdot; K) = d_K$.

Proposición 2.15. *Sea $K \subset \mathbb{R}^n$ un conjunto no-vacío. Se tiene que*

1. *$d_K = d_{\overline{K}}$.*
2. *d_K es 1-Lipschitz continua, es decir, $|d_K(x) - d_K(y)| \leq \|x - y\|$, para todo $x, y \in \mathbb{R}^n$.*
3. *Si K es convexo, entonces $d(\cdot; \operatorname{rint}(K)) = d(\cdot; K)$.*
4. *Si K es convexo, entonces d_K es convexa en $\mathbb{R}^n$.*

Demostración. Demostraremos cada afirmación por separado.

1. Sea $x \in \mathbb{R}^n$. Como $K \subset \overline{K}$, tenemos que

   $$d_K(x) = \inf_{z \in K} \|z - x\| \geq \inf_{z \in \overline{K}} \|z - x\| = d_{\overline{K}}(x).$$

   Por lo tanto, basta demostrar que $d_K(x) \leq d_{\overline{K}}(x)$. En efecto, sea $\varepsilon > 0$ fijo. Por definición, existe $z \in \overline{K}$ tal que $\|z - x\| \leq d_{\overline{K}}(x) + \varepsilon$. Luego, podemos considerar una sucesión $(z_k) \subset K$ tal que $z_k \to z$. Así, podemos escribir

   $$d_K(x) \leq \|z_k - x\| \xrightarrow{k \to \infty} \|z - x\| \leq d_{\overline{K}}(x) + \varepsilon.$$

   Como ε es arbitrario, concluimos que $d_K(x) \leq d_{\overline{K}}(x)$, concluyendo que $d_K(x) = d_{\overline{K}}(x)$.

2. Sean $x, y \in \mathbb{R}^n$. Ocupando desigualdad triangular, sabemos que

   $$d_K(x) \leq \|x - z\| \leq \|x - y\| + \|y - z\|, \quad \forall z \in K.$$

   Luego, podemos escribir,

   $$\begin{aligned} d_K(x) - d_K(y) &= d_K(x) - \inf_{z \in K} \|y - z\| \\ &= \sup_{z \in K} (d_K(x) - \|y - z\|) \leq \|x - y\|. \end{aligned}$$

   Intercambiando los roles de x e y en el desarrollo anterior, tenemos también que $d_K(y) - d_K(x) \leq \|x - y\|$. Luego, mezclando ambas desigualdades, concluimos que

   $$|d_K(x) - d_K(y)| \leq \|x - y\|,$$

   concluyendo así que d_K es 1-Lipschitz.

Universidad de O'Higgins

3. Directo de la parte 1, recordando que por el Teorema 2.12 tenemos que $\overline{\text{rint}(K)} = \overline{K}$.

4. Sean $x_1, x_2 \in \mathbb{R}^n$ y $t \in [0,1]$. Como K es convexo, tenemos que

$$K = \{tz_1 + (1-t)z_2 \ : \ z_1, z_2 \in K\}.$$

Luego, podemos escribir

$$\begin{aligned} d_K(tx_1 + (1-t)x_2) &= \inf_{z\in K} \|tx_1 + (1-t)x_2 - z\| \\ &= \inf_{z_1,z_2\in K} \|tx_1 + (1-t)x_2 - (tz_1 + (1-t)z_2)\| \\ &\leq \inf_{z_1,z_2\in K} (t\|x_1 - z_1\| + (1-t)\|x_2 - z_2\|) \\ &= t \inf_{z_1\in K} \|x_1 - z_1\| + (1-t) \inf_{z_2\in K} \|x_2 - z_2\| = td_K(x_1) + (1-t)d_K(x_2). \end{aligned}$$

Concluimos entonces que d_K es convexa, lo que concluye la demostración.

□

Si bien la función distancia siempre está bien definida, el problema de mejor aproximación puede no tener solución. Sin embargo, cuando el conjunto K es convexo cerrado, la solución siempre existe y además es única.

Teorema 2.16. *Sea $K \subset \mathbb{R}^n$ un conjunto convexo cerrado no-vacío. Entonces, para todo $x \in \mathbb{R}^n$, existe un único punto $\bar{x} \in K$ tal que*

$$\|\bar{z} - x\| = d_K(x).$$

Este punto se denomina ***la proyección de*** *x* ***en*** *K, y se denota por $\bar{z} = P_K(x)$ o bien por $\bar{z} = P(x; K)$.*

Demostración. Primero veamos que existe $\bar{z} \in K$ tal que $d_K(x) = \|z - x\|$. Para esto, notemos que como K es no-vacío, podemos tomar $y \in K$ cualquiera y definir $\rho = \|y - x\|$. Luego, tomando $C = K \cap \overline{B}(x, \rho)$, tenemos que C es convexo compacto no-vacío y que

$$d_K(x) = \inf_{z\in K} \|z - x\| = \inf_{z\in K\cap C} \|z - x\| = d_C(x).$$

Como C es compacto y $\|x - \cdot\|$ es continua, concluimos que existe $\bar{z} \in C \subset K$ tal que

$$\|\bar{z} - x\| = \min_{z\in C} \|z - x\| = d_C(x) = d_K(x).$$

Ahora, veamos que $\bar{z}$ es único. Supongamos que existen dos puntos $\bar{z}_1, \bar{z}_2 \in K$ con $\bar{z}_1 \neq \bar{z}_2$ tal que $\|\bar{z}_1 - x\| = \|\bar{z}_2 - x\| = d_K(x)$. Como K es convexo, tenemos que $z^* = \frac{1}{2}\bar{z}_1 + \frac{1}{2}\bar{z}_2$ está en K. Notando que los puntos $\{x, \bar{z}_1, \bar{z}_2\}$ forman un triángulo isósceles, tenemos que $\|x - z^*\|$ corresponde a la altura de este triángulo, y por lo tanto,

$$\|z^* - x\| < \|\bar{z}_1 - x\| = d_K(x),$$

lo cual es una contradicción con la definición de distancia. Concluimos entonces que existe un único punto $\bar{z} \in K$ tal que $\|\bar{z} - x\| = d_K(x)$. □

La existencia y unicidad de la proyección métrica es una de las grandes virtudes de los conjuntos convexos. Más aún, para estos conjuntos, podemos caracterizar la proyección mediante los ángulos que se forman entre el vector $x - P_K(x)$ y los vectores $z - P_K(x)$ con $z \in K$: El ángulo siempre debe ser obtuso, como ilustra la Figura 2.5.

Universidad de O'Higgins

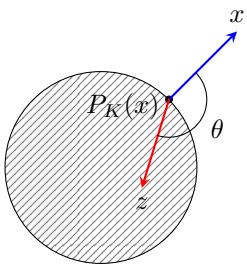


Figura 2.5: Ángulo θ entre $x - P_K(x)$ (en azul) y $z - P_K(x)$ (en rojo) es obtuso, es decir, $\pi/2\theta \leq \pi$.

El ángulo θ entre dos vectores $v_1, v_2 \in \mathbb{R}^n$ y el producto entre ellos se relaciona de la siguiente forma:

$$\langle v_1, v_2 \rangle = \|v_1\| \|v_2\| \cos(\theta).$$

El ángulo θ siempre se considera entre 0 y π, es decir, es el ángulo interno en 0 del triángulo formado por $\{v_1, v_2, 0\}$. Notando que un ángulo en $[0, \pi]$ es obtuso si y sólo si $\cos(\theta) \leq 0$, la relación de ángulos entre $x - P_K(x)$ y $z - P_K(x)$ se puede traducir en términos del producto entre ellos, lo que se conoce como caracterización variacional de la proyección.

Teorema 2.17 (Caracterización variacional de la proyección)**.** *Sea $K \subset \mathbb{R}^n$ un conjunto convexo cerrado no-vacío. Para todo $x \in \mathbb{R}^n$ y todo $\bar{x} \in K$ se tiene que*

$$\bar{x} = P_K(x) \iff \forall z \in K,\ \langle x - \bar{x}, z - \bar{x} \rangle \leq 0. \tag{2.11}$$

Demostración. En el caso que $x \in K$, tenemos que $P_K(x) = x$ y la equivalencia se cumple trivialmente. Por lo tanto, supondremos que $x \in \mathbb{R}^n \setminus K$. Razonemos por doble implicancia.

- $\Leftarrow$: Tomemos $z \in K$ con $z \neq \bar{x}$. Podemos escribir

$$\begin{aligned} \|z - x\|^2 &= \|(z - \bar{x}) - (x - \bar{x})\|^2 \\ &= \|z - \bar{x}\|^2 - 2\langle z - \bar{x}, x - \bar{x} \rangle + \|x - \bar{x}\|^2 \\ &\geq \|x - \bar{x}\|^2. \end{aligned}$$

  Concluimos entonces que $\|\bar{x} - x\| \leq \|z - x\|$ para todo $z \in K$, y por lo tanto $\bar{x} = P_K(x)$.

- $\Rightarrow$: Para esta dirección, consideremos $f(z) = \frac{1}{2}\|z - x\|^2$. Sabemos que f es diferenciable y que $\nabla f(z) = z - x$. Más aún, para todo $z \in K$, se tiene que $f(\bar{x}) \leq f(z)$. Luego, para $z \in K$ fijo podemos escribir

$$\begin{aligned} \langle x - \bar{x}, z - \bar{x} \rangle &= -\langle \nabla f(\bar{x}), z - \bar{x} \rangle \\ &= -f'(\bar{x}; z - \bar{x}) \\ &= -\lim_{t \to 0} \frac{f(\bar{x} + t(z - \bar{x})) - f(\bar{x})}{t} \leq 0 \end{aligned}$$

  donde $f'(\bar{x}; z - \bar{x})$ es la derivada direccional de f en $\bar{x}$ en la dirección $z - \bar{x}$. La última desigualdad se obtiene notando que cuando $t \leq 1$, $\bar{x} + t(z - \bar{x}) = tz + (1 - t)\bar{x} \in K$, y por lo tanto $f(\bar{x} + t(z - \bar{x})) \geq f(\bar{x})$.

□

Universidad de O'Higgins

2.4 Teoremas de separación

Uno de los resultados más importantes del análisis convexo está dado por los teoremas de separación, también conocidos como teoremas de Hahn-Banach. Este resultado nos asegura que, para dos conjuntos convexos, podemos encontrar un hiperplano que los separa, como ilustra la Figura 2.6. Aquí, la convexidad juega un rol fundamental, puesto que sin esta hipótesis geométrica es fácil construir ejemplos donde no es posible separar conjuntos por hiperplanos.

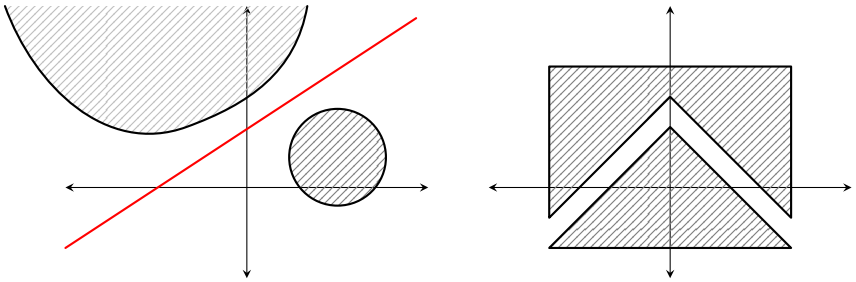


Figura 2.6: A la izquierda: Dos conjuntos convexos en $\mathbb{R}^2$ separados por una recta (en rojo). A la derecha: Dos conjuntos en $\mathbb{R}^2$ que no pueden ser separados por una recta, precisamente porque no son ambos convexos.

Para formalizar esta idea, recordemos la definición de hiperplano.

Definición 2.18 (Hiperplanos y semiespacios)**.** *Un conjunto H de $\mathbb{R}^n$ se dice* hiperplano *si existen un vector $x^* \in \mathbb{R}^n \setminus \{0\}$ y un valor $\alpha \in \mathbb{R}$ tal que*

$$H = [\langle x^*, \cdot \rangle = \alpha] = \{z \in \mathbb{R}^n \ : \ \langle x^*, z \rangle = \alpha\}.$$

Un conjunto E de $\mathbb{R}^n$ se dice semiespacio *si existen un vector $x^* \in \mathbb{R}^n \setminus \{0\}$ y un valor $\alpha \in \mathbb{R}$ tal que*

$$E = [\langle x^*, \cdot \rangle \leq \alpha] = \{z \in \mathbb{R}^n \ : \ \langle x^*, z \rangle \leq \alpha\}.$$

Todo hiperplano H en $\mathbb{R}^n$ es un subespacio afín de dimensión $\operatorname{adim}(H) = n - 1$, y viceversa. El hierplano $H = [\langle x^*, \cdot \rangle = \alpha]$ induce dos semiespacios: $E^- = [\langle x^*, \cdot \rangle \leq \alpha]$ y $E^+ = [\langle x^*, \cdot \rangle \geq \alpha]$. El segundo semiespacio se puede expresar en la forma de la Definición 2.18, reemplazando x^* y α por $-x^*$ y $-\alpha$.

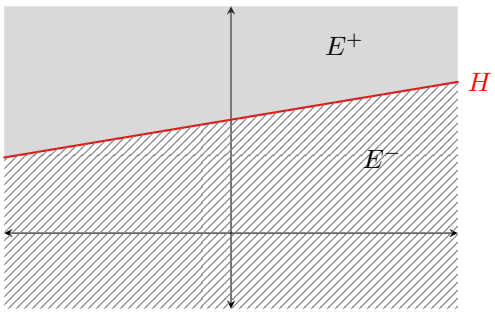


Figura 2.7: Hiperplano H en $\mathbb{R}^2$, y los semiespacios inducidos E^+ (en gris) y E^- (rayado).

Podemos ahora definir lo que significa separar conjuntos por hiperplanos.

Definición 2.19 (Hiperplano separador)**.** *Sean $A, B \subset \mathbb{R}^n$ dos conjuntos no-vacíos y sea $x^* \in \mathbb{R}^n \setminus \{0\}$. Decimos que*

Universidad de O'Higgins

1. x^* ***separa*** A *y* B *si existe* $\alpha \in \mathbb{R}$ *tal que*

$$\sup_{a\in A}\langle x^*, a\rangle \leq \alpha \leq \inf_{b\in B}\langle x^*, b\rangle \quad \text{o bien} \quad \sup_{b\in B}\langle x^*, b\rangle \leq \alpha \leq \inf_{a\in A}\langle x^*, a\rangle.$$

En tal caso, decimos que el hiperplano $H = [\langle x^*, \cdot\rangle = \alpha]$ *es un* ***hiperplano separador*** *de* A *y* B*.*

2. x^* ***separa estrictamente*** *si existe* $\alpha \in \mathbb{R}$ *tal que*

$$\sup_{a\in A}\langle x^*, a\rangle < \alpha < \inf_{b\in B}\langle x^*, b\rangle \quad \text{o bien} \quad \sup_{b\in B}\langle x^*, b\rangle < \alpha < \inf_{a\in A}\langle x^*, a\rangle.$$

En tal caso, decimos que el hiperplano $H = [\langle x^*, \cdot\rangle] = \alpha$ *es un* ***hiperplano separador estricto*** *de* A *y* B*.*

Cuando la separación no es estricta, los conjuntos A y B no necesariamente son disjuntos. Sin embargo, los puntos de intersección pertenecen al hiperplano separador $H = [\langle x^*, \cdot\rangle = \alpha]$.

También vale la pena notar que si en la definición de separación, tenemos que si x^* separa A y B satisfaciendo la primera desigualdad para algún $\alpha \in \mathbb{R}$, entonces $-x^*$ separa A y B satisfaciendo la segunda desigualdad para $-\alpha$. Lo mismo se tiene para separación estricta.

Por supuesto, no todo par de conjuntos admite separadores, como vimos en la Figura 2.6. La siguiente proposición nos dice que cuando el primer conjunto es convexo y el segundo es un singleton que no está en el interior, entonces podemos separarlos.

Proposición 2.20. *Sea* $C \subset \mathbb{R}^n$ *y sea* $\bar{x} \in \mathbb{R}^n \setminus \operatorname{int}(C)$*. Entonces, existe* $x^* \in \mathbb{R}^n \setminus \{0\}$ *tal que*

$$\sup_{y\in C}\langle x^*, y\rangle \leq \langle x^*, \bar{x}\rangle.$$

Es decir, existe $x^* \in \mathbb{R}^n \setminus \{0\}$ *que separa* C *y* $\{\bar{x}\}$*.*

Demostración. Sea $\bar{x} \in C \setminus \operatorname{int}(C)$. Como $\bar{x}$ está en la frontera de C, entonces $\bar{x} \in \overline{\mathbb{R}^n \setminus \overline{C}}$. Por lo tanto, existe una sucesión $(x_k)_k \subset \mathbb{R}^n \setminus \overline{C}$ tal que $x_k \to \bar{x}$. Como $\overline{C}$ es convexo cerrado (ver Teorema 2.12), podemos definir para cada $k \in \mathbb{N}$, el vector

$$x_k^* = \frac{x_k - P_{\overline{C}}(x_k)}{d(x_k; \overline{C})} = \frac{x_k - P_{\overline{C}}(x_k)}{d_C(x_k)} \in \mathbb{S}.$$

Como $\mathbb{S}$ es un conjunto compacto, podemos asumir (reemplazando (x_k^*) por una subsucesión si es necesario) que (x_k^*) es convergente, con límite en $\mathbb{S}$. Sea $x^* = \lim_k x_k^*$. Veamos que x^* soporta C en $\bar{x}$. Tomemos $y \in C$ cualquiera. Ocupando la caracterización variacional de la proyección para $\overline{C}$ (ver Teorema 2.17), sabemos que

$$\langle x_k^*, P_{\overline{C}}(x_k) - y\rangle = \frac{1}{d_C(x_k)}\langle x_k - P_{\overline{C}}(x_k) - x_k, P_{\overline{C}}(x_k) - y\rangle \geq 0.$$

Luego, podemos escribir

$$\begin{aligned}\langle x_k^*, y\rangle &\leq \langle x_k^*, P_{\overline{C}}(x_k)\rangle \\ &= \langle x_k^*, P_{\overline{C}}(x_k) - x_k\rangle + \langle x_k^*, x_k\rangle \\ &= -d_C(x_k) + \langle x_k^*, x_k\rangle \\ &\leq \langle x_k^*, x_k\rangle.\end{aligned}$$

Es decir, para todo $y \in C$, $\langle x_k^*, y\rangle \leq \langle x_k^*, x_k\rangle$. Notando que $x_k^* \to x^*$ y que $x_k \to \bar{x}$, tomando límite concluimos que

$$\langle x^*, y\rangle = \lim_k \langle x_k^*, y\rangle \leq \lim_k \langle x_k^*, x_k\rangle = \langle x^*, \bar{x}\rangle.$$

Concluimos que $\langle x^*, \bar{x}\rangle = \sup_{y\in C}\langle x^*, y\rangle$, concluyendo que x^* soporta a C en $\bar{x}$. □

Universidad de O'Higgins

Teorema 2.21 (de Separación). *Sean $C_1, C_2 \subset \mathbb{R}^n$ dos conjuntos convexos no-vacíos de $\mathbb{R}^n$ disjuntos. Entonces, existe un hiperplano $H = [\langle x^*, \cdot \rangle = \alpha]$ (con $x^* \neq 0$) que separa C_1 y C_2, es decir,*

$$\langle x^*, x_1 \rangle \leq \alpha \leq \langle x^*, x_2 \rangle, \quad \forall x_1 \in C_1, x_1 \in C_2. \tag{2.12}$$

Demostración. Tomemos el conjunto

$$C = C_1 - C_2 = \{x_1 - x_2 \ : \ x_1 \in C_1, x_2 \in C_2\}.$$

Es fácil ver que C es convexo. Más aún, dado que $C_1 \cap C_2 = \emptyset$, entonces $0 \notin C$. Por lo tanto, la Proposición 2.20 dice que existe $x^* \in \mathbb{R}^n \setminus \{0\}$ tal que

$$\sup_{x \in C} \langle x^*, x \rangle \leq 0.$$

Esto equivale a

$$\forall x_1 \in C_1, x_2 \in C_2, \quad \langle x^*, x_1 - x_2 \rangle \leq 0,$$

lo que a su vez equivale a (2.12) tomando $\alpha = \sup_{x_1 \in C_1} \langle x^*, x_1 \rangle$. □

Teorema 2.22 (de Separación Estricta). *Sean $C_1, C_2 \subset \mathbb{R}^n$ dos conjuntos convexos no-vacíos de $\mathbb{R}^n$ disjuntos, con C_1 cerrado y C_2 compacto. Entonces, existe un hiperplano $H = [\langle x^*, \cdot \rangle = \alpha]$ (con $x^* \neq 0$) que separa estrictamente C_1 y C_2, es decir,*

$$\sup_{x_1 \in C_1} \langle x^*, x_1 \rangle < \alpha < \inf_{x_2 \in C_2} \langle x^*, x_2 \rangle. \tag{2.13}$$

Demostración. Como la función distancia d_{C_1} es continua y el conjunto C_2 es compacto, por el teorema fundamental de la optimización (ver Teorema 1.13) sabemos que el problema

$$\begin{cases} \text{mín}_{x_2} & d_{C_1}(x_2) \\ s.a. & x_2 \in C_2 \end{cases} \tag{2.14}$$

tiene solución. Sea $\bar{x}_2$ una solución del problema y tomemos $\bar{x}_1 = P_{C_1}(\bar{x}_2)$. Veamos que de hecho $\bar{x}_2 = P_{C_2}(\bar{x}_1)$. En efecto, si así no lo fuera, entonces existiría $y \in C_2$ tal que

$$d_{C_1}(y) \leq \|y - \bar{x}_1\| < \|\bar{x}_2 - \bar{x}_1\| = d_{C_1}(\bar{x}_2).$$

Pero esto es una contradicción, pues $\bar{x}_2$ es solución del problema (2.14), y por lo tanto $d_{C_1}(\bar{x}_2) \leq d_{C_1}(y)$.

Definamos ahora

$$x^* = \frac{\bar{x}_2 - \bar{x}_1}{2} \qquad \text{y} \qquad \alpha = \frac{1}{2}\langle x^*, \bar{x}_2 + \bar{x}_1 \rangle.$$

Como C_1 y C_2 son disjuntos, tenemos que $x^* \neq 0$, o equivalentemente, que $\|x^*\| > 0$. Veamos que x^* y α satisfacen (2.13). Por un lado, ocupando la caracterización variacional de la proyección (ver Teorema 2.17), tenemos que para todo $y \in C_1$

$$\begin{aligned} \langle x^*, y \rangle &= \frac{1}{2}\langle \bar{x}_2 - \bar{x}_1, y \rangle \\ &\leq \frac{1}{2}\langle \bar{x}_2 - \bar{x}_1, \bar{x}_1 \rangle \\ &= \langle x^*, \bar{x}_1 \rangle \\ &= \left\langle x^*, \frac{\bar{x}_1 + \bar{x}_2}{2} \right\rangle + \left\langle x^*, \frac{\bar{x}_1 - \bar{x}_2}{2} \right\rangle \\ &= \alpha - \|x^*\|^2 < \alpha. \end{aligned}$$

Universidad de O'Higgins

Concluimos que $\sup_{y\in C_1}\langle x^*, y\rangle < \alpha$. Por otro lado, como $\bar{x}_2 = P_{C_2}(\bar{x}_1)$, podemos volver a ocupar la caracterización variacional de la proyección (ver Teorema 2.17), concluyendo que para todo $y \in C_2$

$$\begin{aligned}\langle x^*, y\rangle &= -\frac{1}{2}\langle \bar{x}_1 - \bar{x}_2, y\rangle \\ &\geq -\frac{1}{2}\langle \bar{x}_1 - \bar{x}_2, \bar{x}_2\rangle \\ &= \langle x^*, \bar{x}_2\rangle \\ &= \left\langle x^*, \frac{\bar{x}_1 + \bar{x}_2}{2}\right\rangle + \left\langle x^*, \frac{\bar{x}_2 - \bar{x}_1}{2}\right\rangle \\ &= \alpha + \|x^*\|^2 > \alpha.\end{aligned}$$

Concluimos entonces que $\inf_{y\in C_2}\langle x^*, y\rangle > \alpha$, lo que concluye la demostración. □

2.5 Diferenciabilidad y subdifferenciabilidad de funciones convexas

El gradiente de una función diferenciable $f : \mathbb{R}^n \to \mathbb{R}$ en un punto $x \in \mathbb{R}^n$ nos entrega información sobre la variación instantánea de la función f. Cuando tomamos un vector $v \in \mathbb{R}^n \setminus \{0\}$, el producto $\langle \nabla f(x), v\rangle$ nos dice cuanto "tiende a crecer o decrecer" la función f desde el punto x siguiendo la dirección v. Matemáticamente, esto se captura usando la expansión de Taylor de primer orden:

$$f(x + tv) = f(x) + t\langle \nabla f(x), v\rangle + o(t),$$

donde $o(t)$ es el error asociado a aproximar $t \mapsto f(x + tv)$ por la función lineal afín $t \mapsto f(x) + t\langle \nabla f(x), v\rangle$. Cerca de x, $o(t)$ es pequeño, y por lo tanto la variación de f cerca de x siguiendo la dirección v está dada por el producto $\langle \nabla f(x), v\rangle$.

En esta última sección estudiaremos como la convexidad influye la variación de una función. A priori, estos dos conceptos (diferenciabilidad y convexidad) no tendrían por qué estar relacionados. Sin embargo, la geometría de una función influye el comportamiento de primer orden y viceversa.

Proposición 2.23 (Derivada direccional de funciones convexas)**.** *Sea $C \subset \mathbb{R}^n$ un conjunto convexo no-vacío, sea $f : C \to \mathbb{R}$ una función convexa en C y sea $x \in C$. Entonces, para todo $v \in \mathbb{R}^n$ tal que tal que $x + tv \in C$ para algún $t > 0$, se tiene que la derivada direccional $f'(x; v)$ verifica que*

$$f'(x; v) = \inf_{t>0} \frac{f(x + tv) - f(x)}{t}. \tag{2.15}$$

Más aún, si $x \in \operatorname{int}(C)$, entonces $f'(x; v) > -\infty$ para todo $v \in \mathbb{R}^n$.

Demostración. Reemplazando C por $C - x$ y f por $g = f(\cdot - x) - f(x)$ si es necesario, podemos suponer que $x = 0 \in C$ y $f(x) = 0$, puesto que haciendo el reemplazo tenemos que $f'(x; \cdot) = g'(0; \cdot)$. Sea entonces $v \in \mathbb{R}^n \setminus \{0\}$ tal que $x + \delta v \in C$ para algún $\delta > 0$, y sean $s, t \in \mathbb{R}$ con $0 < t < s < \delta$. Entonces, podemos escribir

$$f(tv) = f\left(\tfrac{t}{s}(sv) + (1 - \tfrac{t}{s})0\right) \leq \frac{t}{s} f(sv).$$

Por lo tanto, tenemos que

$$\frac{f(tv) - f(0)}{t} = \tfrac{1}{t} f(tv) \leq \tfrac{1}{s} f(sv) = \frac{f(sv) - f(0)}{s}.$$

Universidad de O'Higgins

esto muestra que la función $t \in (0,\delta) \mapsto \frac{f(tv)-f(0)}{t}$ es monótona no-decreciente. Por lo tanto

$$\inf_{t>0} \frac{f(0+tv)-f(0)}{t} = \lim_{t\to 0} \frac{f(0+tv)-f(0)}{t} = f'(0;v) \in [-\infty, +\infty),$$

bajo la salvedad que el límite que define la derivada direccional puede valer $-\infty$.

Ahora, supongamos que $0 \in \operatorname{int}(C)$. Dado que (2.15) se verifica, para ver que $f'(0;v)$ existe, basta probar la función $t \mapsto \frac{f(tv)-f(0)}{t}$ es acotada inferiormente. Para esto, podemos aplicar el desarrollo anterior reemplazando v por $-v$, concluyendo que la función

$$t \mapsto -\frac{f(-tv)-f(0)}{t}$$

es no-creciente en $(0,\delta)$, para algún $\delta > 0$ lo suficientemente pequeño tal que $-\delta v$ y δv estén en C. Luego, ocupando convexidad, tenemos que para todo $t \in (0,\delta)$,

$$0 = f(0) = f\left(\tfrac{1}{2}(tv) + \tfrac{1}{2}(-tv)\right) \leq \tfrac{1}{2}f(tv) + \tfrac{1}{2}f(-tv).$$

Luego, concluimos que

$$-\frac{f(-tv)-f(0)}{t} \leq \frac{f(tv)-f(0)}{t}, \quad \forall t \in (0,\delta),$$

y tomando $s = \delta/2$, podemos escribir

$$-\infty < f(-sv) \leq \lim_{t\to 0} -\frac{f(-tv)-f(0)}{t} \leq \lim_{t\to 0} \frac{f(tv)-f(0)}{t} = f'(0;v).$$

Esto concluye la demostración. □

La proposición anterior nos dice que cuando hacemos aproximaciones de la forma $(f(x+tv)-f(x))/t$, estos valores corresponden a las pendientes de las rectas que pasan por $(x, f(x))$ y $(x+tv, f(x+tv))$. Estas pendientes van decreciendo cuando $t \to 0$, alcanzando el valor mínimo en el límite, que corresponde a la derivada direccional. La Figura 2.8 ilustra esta idea.

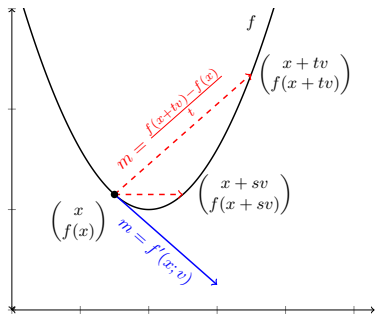


Figura 2.8: Función convexa y derivada direccional. En rojo, pendientes de las rectas dadas por diferencias parciales, con $0 < s < t$. En azul, la recta con pendiente igual a la derivada direccional.

Proposición 2.24 (Caracterización de primer orden)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función diferenciable y $C \subset \mathbb{R}^n$ un conjunto convexo. Se tiene que*

1. *f es convexa en C si y sólo si*

$$\forall x, z \in C, \quad f(z) \geq f(x) + \langle \nabla f(x), z - x \rangle. \tag{2.16}$$

Universidad de O'Higgins

2. *Si f es convexa en C y $x \in \operatorname{int}(C)$, entonces para $x^* \in \mathbb{R}^n$ se cumple que*

$$x^* = \nabla f(x) \iff \forall z \in C, \quad f(z) \geq f(x) + \langle x^*, z - x \rangle. \tag{2.17}$$

3. *Si f es convexa en C, entonces $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ es continua en $\operatorname{int}(C)$.*

Demostración. Estudiaremos cada afirmación por separado:

1. Supongamos primero que f es convexa, y fijemos $x \in C$. Entonces, para todo $z \in C$ con $z \neq x$, definiendo $v = \frac{z-x}{\|z-x\|}$ y $t = \|z - x\|$, la Proposición 2.23 nos asegura que

$$f(z) - f(x) = t\frac{f(x+tv) - f(x)}{t} \geq tf'(x; v) = t\langle \nabla f(x), v \rangle = \langle \nabla f(x), z - x \rangle.$$

   Esto muestra la implicancia directa ($\Rightarrow$) de (2.17), pues el caso $z = x$ es trivial.

   Para la implicancia reversa ($\Leftarrow$), supongamos que f verifica (2.16), y tomemos $x, z \in C$ con $z \neq x$ y $t \in (0, 1)$. Sea $x_t = tz + (1-t)x$. Por un lado, tenemos que

$$\begin{aligned}(1-t)f(x_t) &\leq (1-t)f(x) - (1-t)\langle \nabla f(x_t), x - tz - (1-t)x \rangle \\ &= (1-t)f(x) - (1-t)t\langle \nabla f(x_t), x - z \rangle.\end{aligned}$$

   Por otro lado, también tenemos que

$$\begin{aligned}tf(x_t) &\leq tf(z) - t\langle \nabla f(x_t), z - tz - (1-t)x \rangle \\ &= tf(z) + (1-t)t\langle \nabla f(x_t), x - z \rangle.\end{aligned}$$

   Sumando ambas desigualdades, concluimos que

$$f(tz + (1-t)x) \leq tf(z) + (1-t)f(x),$$

   lo que muestra que f es convexa en C.

2. La implicancia directa ($\Rightarrow$) es concecuencia de la parte 1. Para la implicancia reversa ($\Leftarrow$), supongamos que x^* satisface la desigualdad del lado derecho de (2.17), pero que $x^* \neq \nabla f(x)$. Como $x \in \operatorname{int}(C)$, existe $\delta > 0$ tal que para todo $v \in \mathbb{S}$ y todo $t \in (0, \delta)$, $x + tv \in C$. Entonces, como f es diferenciable, podemos escribir

$$\begin{aligned}0 &= \lim_{t \to 0^+} \frac{f(x+tv) - f(x) - t\langle \nabla f(x), v \rangle}{t} \\ &\geq \lim_{t \to 0^+} \frac{\langle x^*, tv \rangle - t\langle \nabla f(x), v \rangle}{t} \\ &= \langle x^* - \nabla f(x), v \rangle.\end{aligned}$$

   Como $x^* \neq \nabla f(x)$, podemos tomar $v = \frac{x^* - \nabla f(x)}{\|x^* - \nabla f(x)\|}$, lo que implicaría que

$$0 < \|x^* - \nabla f(x)\| = \langle x^* - \nabla f(x), v \rangle \leq 0,$$

   lo cual es una contradicción. Concluimos entonces que $x^* = \nabla f(x)$.

3. Sea $(x_k) \subset \operatorname{int}(C)$ una sucesión convergente con $x_k \to x \in \operatorname{int}(C)$. Mostraremos que $(x_k)_k$ tiene una subsucesión $(x_{k_j})_j$ tal que $\nabla f(x_{k_j}) \to \nabla f(x)$.

   Como f es localmente Lipschitz en $\operatorname{int}(C)$ gracias al Teorema 2.13, existen $L, \delta > 0$ tal que $\|\nabla f(z)\| \leq L$ para todo $z \in B(x, \delta) \subset \operatorname{int}(C)$. Sin perder generalidad, podemos asumir que

Universidad de O'Higgins

$(x_k) \subset B(x, \delta)$ y por lo tanto, la sucesión $(\nabla f(x_k))_k$ es acotada. Así, $(\nabla f(x_k))_k$ admite una subsucesión convergente $(\nabla f(x_{k_j}))_j$, con límite $\lim_j \nabla f(x_{k_j}) = x^*$.

Veamos que $x^* = \nabla f(x)$. Sea $z \in C$. Por continuidad de f y de la función $\langle \cdot, \cdot \rangle$, tenemos que

$$f(z) \geq f(x_{k_j}) + \langle \nabla f(x_{k_j}), z - x_{k_j} \rangle \xrightarrow{j \to \infty} f(x) + \langle x^*, z - x \rangle.$$

Luego, como $z \in C$ es arbitrario, la parte 2. anteriormente demostrada nos dice que necesariamente $x^* = \nabla f(x)$.

Como (x_k) era una sucesión arbitraria, se concluye que ∇f es continua en x (ver Ejercicio **P4** de la Sección 1.6).

□

Cuando una función f es de clase $\mathcal{C}^2$ (dos veces diferenciable con derivadas de primer y segundo orden continuas), entonces la convexidad también puede ser estudiada a partir de su matriz hessiana $\nabla^2 f$.

Proposición 2.25 (Caracterización de segundo orden)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^2$ y $C \subset \mathbb{R}^n$ un conjunto convexo. Se tiene que*

1. *$\nabla^2 f(x)$ es semidefinida positiva para todo $x \in C$, entonces f es convexa en C.*
2. *Si f es convexa en C y $\operatorname{int}(C) \neq \emptyset$, entonces $\nabla^2 f(x)$ es semidefinida positiva para todo $x \in C$.*

Demostración. Estudiemos cada afirmación por separado.

1. Sean $x, z \in C$. Ocupando el teorema de expansión de Taylor de segundo orden, sabemos que

$$f(z) = f(x) + \langle \nabla f(x), z - x \rangle + \frac{1}{2}(z - x)^T \nabla^2 f(x_t)(z - x),$$

con $x_t = tz + (1 - t)x$, para algún $t \in (0, 1)$. Luego, como $\nabla^2 f(x_t)(z - x)$ es semidefinida positiva, se tiene que

$$f(z) \geq f(x) + \langle \nabla f(x), z - x \rangle,$$

lo que muestra la convexidad de f aplicando la Proposición 2.24.

2. Razonemos por contradicción, y supongamos que existe $x \in C$ y $z > 0$ tal que $z^T \nabla^2 f(x) z < 0$. Como f es de clase $\mathcal{C}^2$, existe $\delta > 0$ tal que para todo $x' \in B(x, \delta)$, se tiene que $z^T \nabla^2 f(x') z < 0$. Gracias a esto, y el hecho que $\overline{C} = \overline{\operatorname{int}(C)}$, podemos suponer que $x \in \operatorname{int}(C)$.

   Ahora, reemplazando z por λz con $\lambda > 0$ lo suficientemente pequeño si es necesario, podemos asumir que $x + z \in \operatorname{int}(C)$. y que $z^T \nabla^2 f(x + tz) z < 0$ para todo $t \in [0, 1]$. Luego, ocupando la expansión de Taylor de orden 2 en $x + z$, tenemos que

$$f(x + z) = f(x) + \langle \nabla f(x), z \rangle + \frac{1}{2} z^T \nabla^2 f(x + tz) z < f(x) + \langle \nabla f(x), z \rangle,$$

lo cual es una contradicción, en vista de la Proposición 2.24.

□

La ecuación (2.17) se puede interpretar de la siguiente manera: Si f es convexa y diferenciable, entonces $(\nabla f(x), -1)$ es el único vector en $\mathbb{R}^{n+1}$ que separa $(x, f(x))$ del epígrafo de f. Esto quiere decir que la aproximación lineal $\ell(x) = f(x) + \langle \nabla f(x), z - x \rangle$ es la única función lineal "por debajo" de f que es tangente al epígrafo en $(x, f(x))$.

Universidad de O'Higgins

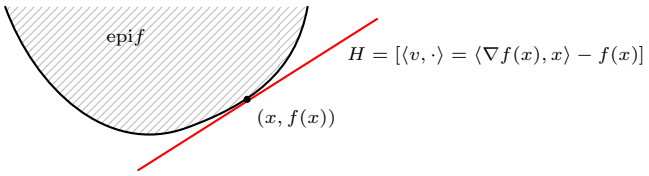


Figura 2.9: Hiperplano H dado por el vector $v = (\nabla f(x), -1) \in \mathbb{R}^n \times \mathbb{R}$, tangente a $(x, f(x))$.

Cuando una función $f : \mathbb{R}^n \to \mathbb{R}$ es convexa pero no es diferenciable en x, la propiedad anterior no se verifica. Sin embargo, los teoremas de separación del capítulo anterior (particularmente la Proposición 2.20) nos asegura que que siempre existe un vector $(a^*, s) \in \mathbb{R}^n \times \mathbb{R}$ que separa $(x, f(x))$ de epi$f$, es decir, tal que

$$\langle a^*, z \rangle + sr \leq \langle a^*, x \rangle + sf(x), \quad \forall (z, r) \in \text{epi} f. \tag{2.18}$$

Notando que el epígrafo es no-acotado en la dirección $(0, 1) \in \mathbb{R}^n \times \mathbb{R}$, este separador debe cumplir necesariamente que $s \leq 0$. Cuando $s \neq 0$, entonces podemos reemplazar (a^*, s) por $(x^*, -1)$ con $x^* = |s|^{-1} a$, y por lo tanto, tomando $r = f(z)$ en la ecuación (2.18), podemos escribir:

$$f(x) + \langle x^*, z - x \rangle \leq f(z), \quad \forall z \in \mathbb{R}^n.$$

Es decir, los vectores $x^* \in \mathbb{R}^n$ construidos de esta manera también inducen aproximaciones lineales por debajo de la función f y tangentes al epígrafo en $(x, f(x))$. El punto es que, cuando f no es diferenciable, pueden existir muchos de estos vectores y no solo uno.

Definición 2.26 (Subgradientes y subdiferencial)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función convexa y $x \in \mathbb{R}^n$. Un vector $x^* \in \mathbb{R}^n$ se dice **subgradiente** de f en x si verifica que*

$$f(x) + \langle x^*, z - x \rangle \leq f(z), \quad \forall z \in \mathbb{R}^n. \tag{2.19}$$

*Se define el **subdiferencial** de f en x, denotado por $\partial f(x)$, como el conjunto de todos los subgradientes de f en x, es decir,*

$$\partial f(x) := \left\{ x^* \in \mathbb{R}^n \ : \ f(x) + \langle x^*, z - x \rangle \leq f(z), \forall z \in \mathbb{R}^n \right\}. \tag{2.20}$$

Ejemplo 2.27 La función valor absoluto $|\cdot| : \mathbb{R} \to \mathbb{R}$ es convexa, y no es diferenciable en 0. Sin embargo, para $s \in [-1, 1]$, tenemos que

$$|z| \geq sz = s(z - 0) + |0|, \quad \forall z \in \mathbb{R}.$$

De hecho, lo anterior se cumple sólo si $s \in [-1, 1]$, y por lo tanto $\partial |\cdot|(0) = [-1, 1]$ (ver Figura 2.10). De forma más general, la norma euclidiana $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ es convexa, no es diferenciable en 0 y

$$\partial \|\cdot\|(0) = \mathbb{B}.$$

Así como el gradiente de una función convexa posee buenas propiedades, el subdiferencial también. La siguiente proposición enuncia alguna de estas propiedades donde quizás la más relevante para nosotros es la relación entre el subdiferencial y la derivada direccional, dada por la ecuación (2.21).

Proposición 2.28 (Propiedades del subdiferencial)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función convexa y sea $x \in \mathbb{R}^n$. Se tiene que*

1. *$\partial f(x)$ es un conjunto convexo, compacto y no-vacío.*
2. *Si f es diferenciable en x, entonces $\partial f(x) = \{\nabla f(x)\}$.*

Universidad de O'Higgins

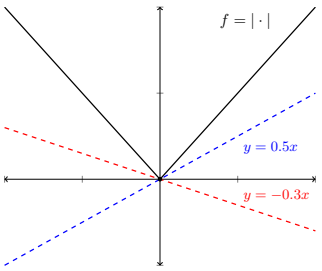


Figura 2.10: Función $f = |\cdot|$. En rojo, la función $y = -0.3x$, asociada al subgradiente $s = -0.3$. En azul, la función $y = 0.5x$, asociada al subgradiente $s = 0.5$.

3. *Para todo punto $x \in \mathbb{R}^n$ y toda dirección $v \in \mathbb{R}^n$, se tiene que*

$$f'(x;v) = \sup\{\langle x^*, v\rangle \ : \ x^* \in \partial f(x)\}. \tag{2.21}$$

Demostración. Estudiaremos cada afirmación por separado.

1. Fijemos $x \in \mathbb{R}^n$. Aplicando la Proposición 2.20 a $C = \text{epi}f$ y el punto $(x, f(x)) \in \mathbb{R}^n \times \mathbb{R}$, sabemos que existe $(a^*, s) \in \mathbb{R}^n \times \mathbb{R}$, con $s \leq 0$, tal que

$$\sup\{\langle a^*, z\rangle + sr \ : \ (z,r) \in \text{epi}f\} \leq \langle a^*, x\rangle - sf(x).$$

Si $s = 0$, entonces el supremo del lado izquierdo de la desigualdad sería $+\infty$, lo cual no puede ser. Por lo tanto $s < 0$ y así podemos reemplazar (a^*, s) por $(x^*, -1)$ como separador. Siguiendo la discusión previa a la Definición 2.26, concluimos que $x^* \in \partial f(x)$, lo que muestra que el subdiferencial es no-vacío. Ocupando la definición de subgradiente dada por la desigualdad (2.19), y el hecho que f es continua, es fácil verificar que $\partial f(x)$ es también convexo y cerrado. Solo nos queda demostrar que $\partial f(x)$ es acotado. Gracias al Teorema 2.13, sabemos que existen $\delta, L > 0$ tal que f es L-Lipschitz en $\overline{B}(x, \delta)$ (si bien el teorema establece una bola abierta, basta reemplazar δ por $\delta/2$ si es necesario). Más aún, f alcanza su máximo y su mínimo en $\overline{B}(x, \delta)$ (ver Teorema 1.13), por lo que podemos asegurar que existe $M > 0$ tal que

$$|f(z)| \leq M, \quad \forall z \in \overline{B}(x, \delta).$$

Tomemos $x^* \in \partial f(x)$. Si $x^* \neq 0$, podemos tomar $z = x + \frac{\delta}{\|x^*\|}x^* \in \overline{B}(x, \delta)$ y luego,

$$\delta\|x^*\| = \langle x^*, z - x\rangle \leq f(z) - f(x) \leq 2M.$$

Por lo tanto, $\|x^*\| \leq 2M/\delta$. Si $x^* = 0$, entonces también se tiene la desigualdad anterior, y por lo tanto $\partial f(x^*)$ es acotado. Esto concluye la demostración.

2. Directo de la Proposición 2.24, parte 2.

3. Sea $x \in \mathbb{R}^n$ y $v \in \mathbb{R}^n$ fijos. Denotemos $g(v) := \sup\{\langle x^*, v\rangle \ : \ x^* \in \partial f(x)\}$. Demostraremos la doble desigualdad, es decir que $f'(x;v) \geq g(v)$ y $f'(x,v) \leq g(v)$. Como $f'(x;0) = 0 = g(0)$, podemos suponer sin perder generalidad que $v \neq 0$.

Sea $x^* \in \partial f(x)$. Tenemos que para todo $t > 0$,

$$\langle x^* v\rangle = \tfrac{1}{t}\langle x^*, tv\rangle \leq \frac{f(x+tv) - f(x)}{t}.$$

Tomando ínfimo en $t > 0$, tenemos que $\langle x^* v\rangle \leq f'(x;v)$. Luego, tomando supremo en $x^* \in \partial f(x)$, concluimos que $g(v) \leq f'(x;v)$.

Universidad de O'Higgins

Para la otra desigualdad, notemos que la función $f'(x;\cdot):\mathbb{R}^n\to\mathbb{R}$ es convexa y por lo tanto continua (ver Proposición 2.13). Así, el conjunto $E=\text{epi}(f'(x;\cdot)$ es convexo cerrado y además $\partial(f'(x;\cdot))(v)\neq\emptyset$.Tomemos $x^*\in\partial(f'(x;\cdot))(v)$. Tenemos entonces dos observaciones:

- $(x^*,-1)$ separa E de $(v,f'(x;v))$, es decir,

$$\sup\{\langle x^*,u\rangle-r\ :\ (u,r)\in E\}\leq\langle x^*,v\rangle-f'(x;v).$$

- $\langle x^*,v\rangle-f'(x;v)=0$. En efecto, recordando que $f'(x;\cdot)$ es convexa y $x^*\in\partial(f'(x;\cdot))(v)$, tenemos que

$$f'(x;v)-\langle x^*,v\rangle=f'(x;v)+\langle x^*,0-v\rangle\leq f'(x;0)=0,$$

lo que muestra que $\langle x^*,v\rangle-f'(x;v)\geq 0$. Por otro lado, si $\langle x^*,v\rangle-f'(x;v)>0$ entonces tomando $\lambda>0$, tendríamos que $f'(x;\lambda v)=\lambda f'(x;v)$ y por lo tanto,

$$\langle x^*,\lambda v\rangle-f'(x;\lambda v)=\lambda(\langle x^*,v\rangle-f'(x;v))\xrightarrow{\lambda\to+\infty}+\infty.$$

Esto sería una contradicción, pues para todo $\lambda>0$, $(\lambda v,f'(x;\lambda v))\in E$.

Luego, notando que para todo $u\in\mathbb{R}$ el par $(u,f(x+u)-f(x))\in E$ tenemos que

$$\langle x^*,u\rangle-f(x)-f(x+u)=\langle x^*,u\rangle-(f(x+u)-f(x))\leq\langle x^*,v\rangle-f'(x;v)=0.$$

Esto demuestra que $x^*\in\partial f(x)$, y por lo tanto,

$$g(v)\geq\langle x^*,v\rangle=f'(x;v\rangle.$$

Concluimos entonces que $g(v)=f'(x;v)$, lo que termina la demostración.

□

Para cerrar esta sección, veremos que las funciones convexas son particularmente bien adaptadas para problemas de minimización, pues los mínimos globales pueden ser caracterizados mediante la ecuación $\nabla f(x)=0$. Los puntos que satisfacen la igualdad anterior (sea f convexa o no) se conocen como puntos críticos de f y ahondaremos más en ellos en los capítulos venideros. En el caso convexo no-diferenciable, decimos que x es punto crítico de f si $0\in\partial f(x)$, lo que generaliza la ecuación $\nabla f(x)=0$ para el caso diferenciable, gracias a la Proposición 2.28.

Para una función $f:\mathbb{R}^n\to\mathbb{R}$ diferenciable sabemos que, si un punto x es mínimo global de f, entonces es mínimo local también. A su vez, si x es mínimo local de f, entonces es punto crítico, es decir, verifica que $\nabla f(x)=0$. Sin embargo, en general, un punto crítico no tiene por qué ser mínimo local, y un mínimo local no tiene por que ser mínimo global. Lo impresionante es que para las funciones convexas estos tres conceptos son equivalentes.

Teorema 2.29 (Regla de Fermat para funciones convexas)**.** *Sea $f:\mathbb{R}^n\to\mathbb{R}$ una función convexa y $x\in\mathbb{R}^n$. Las siguientes afirmaciones son equivalentes:*

(i) *x es mínimo global de f en $\mathbb{R}^n$.*

(ii) *x es mínimo local de f en $\mathbb{R}^n$.*

(iii) $0\in\partial f(x)$.

Demostración. Claramente $(i)\Rightarrow(ii)$. Veamos ahora que $(ii)\Rightarrow(iii)$. Sea $x\in\mathbb{R}^n$ un mínimo local de f. Entonces, existe $\delta>0$ tal que para todo $z\in B(x,\delta)$, se tiene que

$$f(x)=\langle 0,z-x\rangle+f(x)\leq f(z).$$

Universidad de O'Higgins

Tomemos ahora $z \in \mathbb{R}^n$ y $t \in (0,1)$ lo suficientemente pequeño tal que $z_t = tz + (1-t)x \in B(x,\delta)$. Entonces, tenemos que

$$f(x) \leq f(z_t) \leq tf(z) + (1-t)f(x).$$

Luego, restando a ambos lados $(1-t)f(x)$ y dividiendo por t, podemos escribir

$$\langle 0, z-x\rangle + f(x) = f(x) \leq f(z),$$

por lo que concluimos que $0 \in \partial f(x)$. Para terminar, demostremos que $(iii) \Rightarrow (i)$. Supongamos que $x \in \mathbb{R}^n$ es un punto crítico de f, es decir, tal que $0 \in \partial f(x)$. Entonces, para todo $z \in \mathbb{R}^n$, tenemos que

$$f(x) = f(x) + \langle 0, z-x\rangle \leq f(z).$$

Esto concluye la demostración. □

2.6 Ejercicios Capítulo 2

P1. Muestre que $C \subset \mathbb{R}^n$ es un conjunto convexo si y solamente si para todo $k \in \mathbb{N}$, para todo $\{x_1, \ldots, x_k\} \subset C$ y para todo $\{\lambda_1, \ldots, \lambda_k\} \subset \mathbb{R}_+$ tal que $\sum_{i=1}^k \lambda_i = 1$, se tiene que $\sum_{i=1}^k \lambda_i x_i \in C$.

P2. *(Desigualdad de Jensen)* Sea $C \subset \mathbb{R}^n$ un conjunto convexo. Muestre que una función $f : C \to \mathbb{R}$ es convexa si y solo si, para todo $k \in \mathbb{N}$, para todo $\{x_1, \ldots, x_k\} \subset C$ y para todo $\{\lambda_1, \ldots, \lambda_k\} \subset \mathbb{R}_+$ tal que $\sum_{i=1}^k \lambda_i = 1$, se tiene que

$$f\left(\sum_{i=1}^k \lambda_i x_i\right) \leq \sum_{i=1}^n \lambda_i f(x_i).$$

P3. Una función $A : \mathbb{R}^n \to \mathbb{R}^m$ se dice *lineal afín* si existen $T : \mathbb{R}^n \to \mathbb{R}^m$ transformación lineal y $b \in \mathbb{R}^m$ tal que $A(x) = T(x) + b$ para todo $x \in \mathbb{R}^n$. Muestre que

- Si $C \subset \mathbb{R}^n$ es convexo y $A : \mathbb{R}^n \to \mathbb{R}^m$ es una función lineal afín, entonces $A(C)$ es convexo.
- Si $g : \mathbb{R}^m \to \mathbb{R}$ es una función convexa y $A : \mathbb{R}^n \to \mathbb{R}^m$ es una función lineal afín, entonces $f = g \circ A$ es convexa.

P4. *(Lema de prolongación)* Sea $C \subset \mathbb{R}^n$ y sea V el espacio paralelo a $\text{aff}(C)$. Muestre que si $x_0 \in \text{rint}(C)$ y d es un vector en V, entonces existe $\varepsilon > 0$ lo suficientemente pequeño tal que

$$x_0 + td \in \text{rint}(C), \qquad \forall t \in [0, \varepsilon].$$

P5. *(Cambio de Variables)* Sea C un conjunto convexo no-vacío en $\mathbb{R}^n$ con dimensión afín $\text{adim}(C) = m$ y $f : C \to \mathbb{R}$ una función convexa. El objetivo de este problema es mostrar que podemos hacer un cambio de variables $T : \mathbb{R}^m \to \text{aff}(C)$ y construir $D \subset \mathbb{R}^m$ y $g : D \to \mathbb{R}$ de tal manera que

- D es un conjunto convexo con $0 \in \text{int}D$.
- g es convexa.
- T es invertible, $T(D) = C$ y $f = g \circ T^{-1}$.

Universidad de O'Higgins

Para esto, siga los siguientes pasos:

a) Muestre que si $x_0 \in \text{rint}(C)$ y V es el subespacio paralelo de $\text{aff}(C)$, entonces $D_1 = C - x_0$ es convexo, $0 \in \text{rint}(D_1)$ y $\text{aff}(D_1) = V$.

b) Muestre que si $\{z_1, \ldots, z_m\}$ es una base del subespacio V, entonces el conjunto

$$D = \left\{ \lambda \in \mathbb{R}^m \ : \ \sum_{i=1}^{m} \lambda_i z_i \in D_1 \right\}$$

es convexo y que $0 \in \text{int}(D)$.

c) Muestre que $T : \mathbb{R}^m \to \text{aff}(C)$ dada por

$$T(\lambda) = x_0 + \sum_{i=1}^{m} \lambda_i z_i$$

es lineal afín e invertible.

d) Muestre que la función $g = f \circ T$ es convexa. Concluya el resultado.

P6. Sean C_1 y C_2 dos conjuntos convexos disjuntos de $\mathbb{R}^n$, tal que el conjunto

$$C = C_1 - C_2 = \{x_1 - x_2 \ : \ x_1 \in C_1, x_2 \in C_2\}$$

es cerrado. Muestre que en tal caso existe un hiperplano $H = [\langle x^*, \cdot \rangle = \alpha]$ (con $x^* \neq 0$) que separa estrictamente C_1 y C_2.

Hint: Considere $\bar{x} = P_C(0)$. Muestre que $\bar{x} = \bar{x}_1 - \bar{x}_2$ para algún par $\bar{x}_1 \in C_1$ y $\bar{x}_2 \in C_2$. Verifique que

$$\bar{x}_1 = P_{C_1}(\bar{x}_2) \qquad \text{y} \qquad \bar{x}_2 = P_{C_2}(\bar{x}_1),$$

y luego replique el argumento de la demostración del Teorema 2.22.

P7. Una función $\sigma : \mathbb{R}^n \to \mathbb{R}$ se dice *sublineal* si cumple que

- **(subaditividad)** Para todo $x, y \in \mathbb{R}^n$, $\sigma(x + y) \leq \sigma(x) + \sigma(y)$.
- **(homogeneidad positiva)** Para todo $x \in \mathbb{R}^n$ y todo $\lambda > 0$, $\sigma(\lambda x) = \lambda \sigma(x)$.

(a) Muestre que toda función sublineal es convexa.

(b) Sea $f : \mathbb{R}^n \to \mathbb{R}$ es una función convexa. Muestre que para todo $x \in \mathbb{R}^n$, la derivada direccional $f'(x; \cdot)$ es una función sublineal. **Hint:** Para demostrar la subaditividad, observe que para $u, v \in \mathbb{R}^n$ y $t > 0$ se tiene que

$$f(x + t(u + v)) = f\left(\tfrac{1}{2}(x + 2tu) + \tfrac{1}{2}(x + 2tv)\right)$$

Usando la identidad anterior, ocupe álgebra de límites y la convexidad de f para mostrar la desigualdad deseada.

P8. Muestre, ocupando las caracterizaciones de primer orden de convexidad que

a) La función $f : \mathbb{R} \to \mathbb{R}$ dada por $f(x) = e^x$ es convexa.

b) La función $f : (0, +\infty) \to \mathbb{R}$ dada por $f(x) = x^{-1}$ es convexa.

Universidad de O'Higgins

Capítulo 3

Optimización sin Restricciones

En este capítulo, nos enfocaremos en el problema de optimización sin restricciones, es decir, en el problema

$$\min_{x \in \mathbb{R}^n} f(x), \tag{3.1}$$

donde $f : \mathbb{R}^n \to \mathbb{R}$ es una función de clase $\mathcal{C}^1$. Recordemos que f es una función de clase $\mathcal{C}^1$ si es diferenciable en todo punto y su gradiente $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ es una función continua.

Salvo que tengamos una condición como la convexidad, que nos permita traducir comportamiento local a comportamiento global (como ocurre en el Teorema 2.29), lo mejor que podemos esperar en el Problema (3.1) es encontrar *soluciones locales*. La intuición detrás de esta limitación es que es imposible (computacionalmente hablando) explorar todo el espacio $\mathbb{R}^n$.

Resolver el problema (3.1) de manera local significa encontrar un punto $x \in \mathbb{R}^n$ (de manera exacta o aproximada) tal que

$$f(\bar{x}) \leq f(z), \quad \forall z \in B(\bar{x}, \delta). \tag{3.2}$$

Sin embargo, atacar este problema directamente puede ser difícil e incluso imposible. Básicamente, el problema nuevamente es que en general la condición de optimalidad local (3.2) no se puede verificar computacionalmente.

Por ende, en vez de buscar una solución local de manera directa, aprovecharemos la información de primer orden de la función f para buscar puntos "candidatos a solución": Es decir, a partir de un punto inicial $x_0 \in \mathbb{R}^n$, buscaremos un punto $\bar{x} \in \mathbb{R}^n$ tal que

- $f(\bar{x}) \leq f(x_0)$; y que
- cumpla alguna *condición de optimalidad verificable*.

Con algo de suerte, el punto $\bar{x}$ que encontremos será un óptimo local de f.

3.1 Condiciones de optimalidad

En esta sección, formalizaremos lo que significa *condición de optimalidad verificable*. Supongamos que tenemos una propiedad $P(f, x)$ que puede ser verdadera o falsa, dependiendo del punto $x \in \mathbb{R}^n$ y la función f. Diremos que la propiedad es *verificable* si es posible construir un algoritmo que, conociendo f y x, me permita evaluar $P(f, x)$.

- **Condiciones necesarias de optimalidad:** Son propiedades que verifican la siguiente impli-

Universidad de O'Higgins

cancia

$$\bar{x} \text{ es óptimo local de } f \implies P(f, \bar{x}) \text{ es verdadera.}$$

Una condición necesaria nos permite encontrar puntos candidatos a solución. Idealmente, las condiciones necesarias son útiles si buscar puntos que las satisfagan (o sea, tal que $P(f, x)$ es verdadera) es "más fácil" que buscar óptimos locales.

- **Condiciones suficientes de optimalidad:** Son propiedades que verifican la siguiente implicancia

$$P(f, \bar{x}) \text{ es verdadera} \implies \bar{x} \text{ es óptimo local de } f.$$

Una condición suficiente nos permite evaluar si puntos encontrados son soluciones locales. Idealmente, las condiciones suficientes son útiles si verificarlas es "más fácil" que verificar optimalidad local.

La estrategia general de un algoritmo de optimización es la siguiente: Buscar candidatos a solución utilizando una condición necesaria, y luego verificar si los candidatos son solución utilizando una condición suficiente. En el caso de funciones diferenciables, ocuparemos la condición necesaria de *puntos críticos*.

Definición 3.1 (Punto Crítico)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función diferenciable. Decimos que $x \in \mathbb{R}^n$ es un **punto crítico** de f si*

$$\nabla f(x) = 0.$$

*Si $x \in \mathbb{R}^n$ no es un punto crítico de f, entonces se dice **punto regular** de f.*

Observe que si f es una función diferenciable para la cual conocemos ∇f, la condición de punto crítico es verificable: Para un punto $x \in \mathbb{R}^n$ basta evaluar ∇f en x y ver si es igual o distinto de 0.

Proposición 3.2 (Condición Necesaria de Optimalidad de primer orden)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función diferenciable y $\bar{x} \in \mathbb{R}^n$. Se tiene que*

$$\bar{x} \text{ es mínimo local de } f \implies \bar{x} \text{ es punto crítico de } f.$$

Demostración. Supongamos que $\bar{x}$ es óptimo local de f. Eso significa que existe $\delta > 0$ lo suficientemente pequeño tal que

$$f(z) - f(\bar{x}) \geq 0, \quad \forall z \in B(x, \delta).$$

Sea ahora $v \in \mathbb{S}_n$. Como $\|v\| = 1$, tenemos que $x + tv \in B(x, \delta)$ para todo $t \in (0, \delta)$. Luego, tenemos que

$$\begin{aligned} \langle \nabla f(\bar{x}), v \rangle &= f'(\bar{x}; v) \\ &= \lim_{t \to 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t} \\ &= \lim_{(0,\delta) \ni t \to 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t} \geq 0. \end{aligned}$$

Por lo tanto, $\langle \nabla f(\bar{x}), v \rangle \geq 0$ para todo $v \in \mathbb{S}_n$, lo que implica que $\nabla f(\bar{x}) = 0$. □

Proposición 3.3 (Condición necesaria de segundo orden)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^2$ y $\bar{x} \in \mathbb{R}^n$. Se tiene que*

$$\bar{x} \text{ es mínimo local de } f \implies \nabla^2 f(\bar{x}) \text{ es semidefinida positiva.}$$

Demostración. Sea $\bar{x}$ mínimo local de f. Razonemos por contradicción, y supongamos que $\nabla^2 f(\bar{x})$ no es semidefinida positiva, es decir,

$$\exists d \in \mathbb{S}_n, \quad d^T \nabla^2 f(\bar{x}) d < 0.$$

Universidad de O'Higgins

Como f es de clase $\mathcal{C}^2$, entonces el hessiano $\nabla^2 f : \mathbb{R}^n \to \mathcal{M}_{n\times n}(\mathbb{R})$ es continua. Luego, existe $\delta > 0$ lo suficientemente pequeño tal que

$$d^T \nabla^2 f(z) d < 0, \quad \forall z \in B(x, \delta).$$

Achicando δ si es necesario, podemos asumir que $f(\bar{x}) \leq f(z)$ para todo $z \in B(\bar{x}, \delta)$. Luego, tomando $\varepsilon = \frac{\delta}{2}$ y ocupando el teorema de expansión de Taylor, sabemos que existe $t \in (0, 1)$ tal que

$$f(\bar{x}) \leq f(\bar{x} + \varepsilon d) = f(\bar{x}) + \underbrace{\langle \nabla f(\bar{x}), \varepsilon d \rangle}_{=0} + \underbrace{\frac{1}{2} d^T \nabla^2 f(\bar{x} + t\varepsilon d) d}_{<0} < f(\bar{x}),$$

lo cual es una contradicción. □

Ambas condiciones necesarias (Proposiciones 3.2 y 3.3) son verificables. En efecto, como ya mencionamos, para la condición necesaria de primer orden basta evaluar el gradiente. Por otro lado, para la condición necesaria de segundo orden, basta calcular los valores propios del hessiano, y recordar que para una matriz simétrica $A \in \mathcal{M}_{n\times n}(\mathbb{R})$ se tiene que

1. A es semidefinida positiva si y solo si sus valores propios son todos no-negativos.
2. A es definida positiva si y solo si sus valores propios son todos estrictamente positivos.

Finalmente, ocupando la segunda de las afirmaciones precedentes, podemos también enunciar una condición suficiente de segundo orden, también verificable.

Proposición 3.4 (Condición Suficiente de segundo orden)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^2$ y $x \in \mathbb{R}^n$. Se tiene que*

$$\begin{array}{c} x \text{ es punto crítico de } f \text{ y} \\ \nabla^2 f(x) \text{ es definida positiva} \end{array} \implies \exists \delta > 0 \text{ tal que } f(x) < f(z) \text{ para todo } z \in B(x, \delta) \setminus \{x\}.$$

En tal caso, decimos que x es un mínimo local estricto de f.

Demostración. Como $\nabla^2 f$ es continua y $\nabla^2 f(x)$ es definida positiva, entonces existe $\delta > 0$ tal que

$$\nabla^2 f(z) \text{ es definida positiva, para todo } z \in B(x, \delta).$$

En efecto, si así no lo fuera, existiría una sucesión $z_k \to x$ y una sucesión $(d_k) \subset \mathbb{S}_n$ tal que

$$\forall k \in \mathbb{N}, d_k^T \nabla^2 f(z_k) d_k < 0.$$

Como $\mathbb{S}_n$ es compacto, existe una subsucesión (d_{k_j}) convergente, digamos con $\lim_j d_{k_j} = d \in \mathbb{S}_n$. Luego,

$$d^T \nabla^2 f(x) d = \lim_j d_{k_j}^T \nabla^2 f(z_{k_j}) d_{k_j} \leq 0,$$

lo cual sería una contradicción con el hecho de que $\nabla^2 f(x)$ es definida positiva.

Luego, para todo $z \in B(x, \delta) \setminus \{x\}$, existe $\xi \in (x, z)$ tal que

$$f(z) = f(x) + \underbrace{\langle \nabla f(x), z - x \rangle}_{=0} + \underbrace{\tfrac{1}{2}(z - x)^T \nabla^2 f(\xi)(z - x)}_{>0} > f(x).$$

□

Universidad de O'Higgins

3.2 Algoritmos de búsqueda lineal

Supongamos que queremos resolver el problema (3.1) para una función $f : \mathbb{R}^n \to \mathbb{R}$ diferenciable, y que contamos con un punto inicial $x_0 \in \mathbb{R}^n$. Un algoritmo iterativo se dice *Método de búsqueda lineal* si en cada iteración $k \in \mathbb{N}$, construimos x_{k+1} como

$$x_{k+1} = x_k + \alpha_k \nu_k, \tag{3.3}$$

donde $\nu_k \in \mathbb{R}^n \setminus \{0\}$ y $\alpha_k > 0$. Si x_k no es un punto crítico, queremos que el siguiente iterando x_{k+1} sea un mejor punto que x_k, es decir, queremos que $f(x_{k+1}) < f(x_k)$. En cada iteración, el algoritmo debe seleccionar una dirección $\nu_k \in \mathbb{R}^n \setminus \{0\}$, y luego buscar el paso $\alpha_k > 0$ para "avanzar" en la dirección ν_k, produciendo el siguiente iterando x_{k+1}. El término búsqueda lineal se refiere a que x_{k+1} se busca a través de la línea $x_k + \mathbb{R}_+ \nu_k$. La Figura 3.1 muestra gráficamente las iteraciones de un método de búsqueda lineal para una función $f : \mathbb{R}^2 \to \mathbb{R}$.

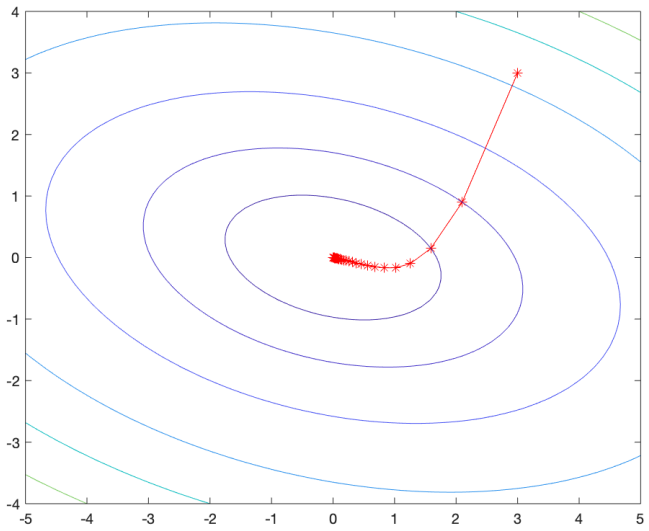


Figura 3.1: Puntos producidos por un método de búsqueda lineal (estrellas en rojo), para la función $f(x, y) = x^2 + xy + 3y^2$, con punto incial $(x_0, y_0) = (3, 3)$. Las elipses son curvas de nivel de la función f. El método converge al mínimo global $(x^*, y^*) = (0, 0)$.

Para que un algoritmo de búsqueda lineal tenga posibilidades de encontrar un mínimo local, necesitamos que la dirección ν_k de búsqueda nos entregue un mejor punto que el inicial. Para lograr esto, ν_k debe ser una *dirección de descenso* de f en x_k.

Definición 3.5 (Dirección de descenso)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función continua y $x \in \mathbb{R}^n$ tal que la derivada direccional $f'(x; \cdot)$ está bien definida. Decimos que $\nu \in \mathbb{R}^n \setminus \{0\}$ es una* dirección de descenso de f en x *si*

$$f'(x; \nu) = \lim_{t \to 0} \frac{f(x + t\nu) - f(x)}{t} < 0.$$

Cuando la función $f : \mathbb{R}^n \to \mathbb{R}$ es diferenciable, es posible caracterizar las direcciones de descenso a partir del gradiente de f.

Proposición 3.6. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función diferenciable y $x \in \mathbb{R}^n$ un punto regular de f. Entonces, se tiene que*

Universidad de O'Higgins

1. *La dirección* $\bar{\nu} := -\frac{\nabla f(x)}{\|\nabla f(x)\|}$ *es la* dirección de máximo descenso, *en el sentido que*

$$f'(x;\bar{\nu}) \leq f'(x;\nu), \quad \forall \nu \in \mathbb{S}_n.$$

2. *Para todo* $\nu \in \mathbb{R}^n \setminus \{0\}$, *se tiene que*

$$\theta = \arccos(\langle \bar{\nu}, \nu \rangle) < \frac{\pi}{2} \Leftrightarrow \nu \text{ es dirección de descenso de } f \text{ en } x.$$

Demostración. Tomemos $\nu \in \mathbb{S}_n$. Tenemos que

$$\begin{aligned} f'(x;\nu) &= \langle \nabla f(x), \nu \rangle \\ &\geq -\|\nabla f(x)\|\|\nu\| \\ &= -\|\nabla f(x)\| = \langle \nabla f(x), \bar{\nu} \rangle = f'(x;\nu). \end{aligned}$$

Concluimos entonces que $\bar{\nu}$ es la dirección de máximo descenso de f en x. Para la segunda parte, notemos que

$$f'(x;\nu) = \langle \nabla f(x), \nu \rangle = -\|\nabla f(x)\| \langle \bar{\nu}, \nu \rangle = -\|\nabla f(x)\|\|\nu\| \cos(\theta).$$

Por lo tanto, como $\theta \in [0, \pi]$ tenemos que

$$\nu \text{ es dirección de descenso} \iff \cos(\theta) > 0 \iff \theta < \pi/2.$$

Esto concluye la demostración. □

La proposición anterior se puede interpretar geométricamente ocupando los conjuntos de subnivel de la función f. En efecto, consideremos el conjunto $S = [f \leq f(x)]$. Cuando f es de clase $\mathcal{C}^1$, se puede demostrar que el hiperplano $H = [\langle \nabla f(x), \cdot \rangle = \langle \nabla f(x), x \rangle]$, es tangente a S en el punto x. Por lo tanto el vector $\nabla f(x)$ es el vector exterior a S y ortogonal a H. Similarmente, el vector $\bar{\nu} = -\nabla f(x)/\|\nabla f(x)\|$ es el vector ortogonal a H que apunta hacia el interior de S. Finalmente, si un vector $\nu \in \mathbb{R}^n \setminus \{0\}$ forma un ángulo $\theta < \pi/2$ con $\bar{\nu}$, entonces también apunta hacia el interior de S. Eso significa que existe $\delta > 0$ lo suficientemente pequeño tal que

$$x + t\nu \in S, \quad \forall t \in (0, \delta).$$

Lo anterior es equivalente a que ν sea una dirección de descenso: Para un $\alpha > 0$, $x + \alpha\nu$ está en el interior de S y por lo tanto $f(x + \alpha\nu) < f(x)$. La Figura 3.2 muestra esta interpretación.

Observación 3.7. *Cabe destacar que al momento de elegir una dirección de descenso* $\nu \in \mathbb{R}^n \setminus \{0\}$ *para* f *en* x, *es equivalente tomar* ν *directamente o tomar el vector normalizado* $\nu/\|\nu\|$. *Esto pues en la fórmula* (3.3), *el paso* $\alpha > 0$ *que elegimos para* ν *se reemplaza por el paso* $\alpha \cdot \|\nu\|$ *para el vector* $\nu/\|\nu\|$.

La Proposición 3.6 nos entrega una fórmula directa para seleccionar la dirección de descenso ν_k en la expresión (3.3):

- Si x_k no es punto crítico de f, entonces tomar $\nu_k = -\nabla f(x_k)$.
- Si x_k es punto crítico de f, entonces la búsqueda termina, pues x_k es un candidato a mínimo local (ver Proposición 3.2).

Esta estrategia se conoce como *Algoritmo de máximo descenso*, pues en cada iteración k, toma la dirección de máximo descenso de f en el iterando x_k. Sin embargo, incluso si consideramos la dirección de máximo descenso, aún debemos escoger el paso $\alpha_k > 0$. Si el paso escogido no es lo "suficientemente bueno", entonces el método de búsqueda lineal puede fallar en encontrar puntos críticos.

Universidad de O'Higgins

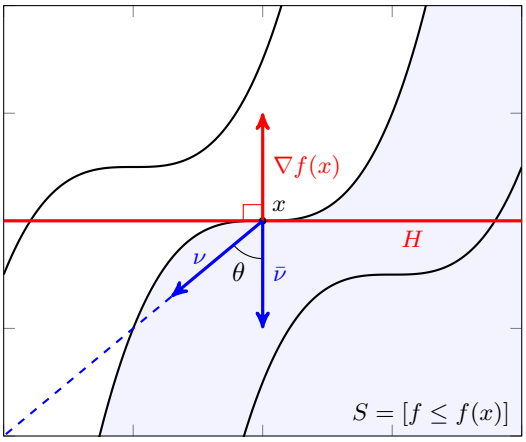


Figura 3.2: Interpretación geométrica de dirección de descenso. En negro, las curvas de nivel. En azul claro rellenado, el conjunto de subnivel $S = [f \leq f(x)]$. En rojo, el hiperplano H y el vector $\nabla f(x)$, ortogonal a H y exterior a S. En azul, los vectores de descenso $\bar{\nu} = -\nabla f(x)/\|\nabla f(x)\|$ y ν, que forman un ángulo $\theta < \pi/2$.

Ejemplo 3.8 Consideremos la función $f : \mathbb{R} \to \mathbb{R}$ dada por $f(t) = t^2$. Sabemos que f alcanza su único mínimo global en $t = 0$, que a su vez es el único punto crítico de f. Por otro lado, para $t \neq 0$ tenemos que la dirección de máximo descenso está dada por

$$-\frac{f'(t)}{|f'(t)|} = -\frac{2t}{|2t|} = \begin{cases} 1 & \text{si } t < 0, \\ -1 & \text{si } t > 0. \end{cases}$$

Tomemos $t_0 = 2$ y para cada $k \in \mathbb{N}$, fijemos $\nu_k = -\frac{f'(t_k)}{|f'(t_k)|}$ y $\alpha_k = 4$. En este caso, tenemos que

$$t_1 = t_0 + 4\nu_0 = 2 - 4 = -2$$
$$t_2 = t_1 + 4\nu_1 = -2 + 4 = 2$$
$$\vdots$$
$$t_{k+1} = t_k + 4\nu_k = \begin{cases} 2 & \text{si } k \text{ es par}, \\ -2 & \text{si } k \text{ es impar}. \end{cases}$$

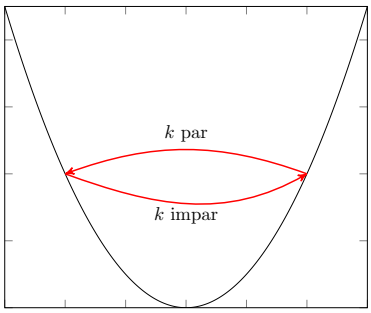


Figura 3.3: Iteración de máximo descenso para $f(t) = t^2$ con $\alpha_k = 4$, $\nu_k = -\frac{f'(t_)}{|f'(t_k)|}$ y $t_0 = 2$.

Es decir, en este caso la sucesión (t_k) oscila entre 2 y -2, y nunca converge a un punto crítico. El problema acá es que α_k es demasiado grande.

Por otro lado, tomemos $t_0 = 3$, $\nu_k = -\frac{f'(t_k)}{|f'(t_k)|}$ y $\alpha_k = \frac{1}{2^k}$. En este caso, tenemos que

En este caso, la sucesión (t_k) converge, pero el límite $\bar{t} = 1$ no es un punto crítico de f. El problema aquí es que α_k es muy pequeño. □

Universidad de O'Higgins

$$t_1 = t_0 + \nu_0 = 3 - 1 = 2$$
$$t_2 = t_1 + \frac{1}{2}\nu_1 = 2 - \frac{1}{2} = \frac{3}{2}$$
$$\vdots$$
$$t_{k+1} = t_0 - \sum_{j=0}^{k} \frac{1}{2^j} \xrightarrow{k\to\infty} 3 - 2 = 1.$$

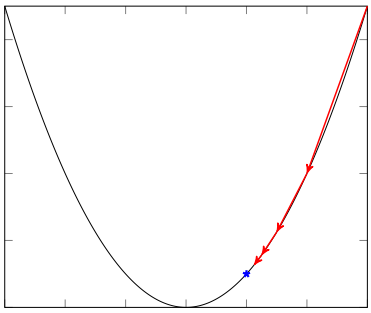


Figura 3.4: Iteración de máximo descenso para $f(t) = t^2$ con $\alpha_k = 1/2^k$, $\nu_k = -\frac{f'(t_k)}{|f'(t_k)|}$ y $t_0 = 3$. En azul, el límite $\bar{t} = 1$.

3.2.1. Selección de paso: Condiciones de Armijo y Wolfe

Supongamos que hemos diseñado un método de búsqueda lineal que en cada iteración x_k selecciona una dirección de descenso ν_k (puede ser $\nu_k = -\nabla f(x_k)/\|\nabla f(x_k)\|$ como en el caso del algoritmo de máximo descenso, u otra diferente). En esta sección estudiaremos cómo seleccionar $\alpha_k > 0$ de tal manera que no sea ni muy grande ni muy pequeño, y evitar comportamientos indeseados como en el Ejemplo 3.8.

Definiendo la función $\phi : [0, +\infty) \to \mathbb{R}$ dada por $\phi(\alpha) = f(x_k + \alpha\nu_k)$, nos gustaría seleccionar el paso α_k que resuelve el problema

$$\min_{\alpha>0} \phi(\alpha). \tag{3.4}$$

Sin embargo, salvo casos particulares, esto resulta demasiado costoso computacionalmente o simplemente impracticable. Por lo tanto, estudiaremos condiciones para seleccionar α_k de tal manera que garanticen dos cosas: 1) *Suficiente descenso*, o sea, que $f(x_k + \alpha_k\nu_k)$ sea lo "suficentemente mejor" que $f(x_k)$ (esto evita el comportamiento oscilante del primer caso del Ejemplo 3.8); y 2) *Suficiente variación*, o sea, que el punto $x_k + \alpha_k\nu_k$ esté los "suficientemente lejos" de x_k (esto evita la convergencia prematura del segundo caso del Ejemplo 3.8).

Definición 3.9 (Condiciones de Wolfe)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^1$, $x \in \mathbb{R}^n$ un punto regular de f y ν una dirección de descenso de f en x. Sea $c_1 \in (0, 1)$ y $c_2 \in (c_1, 1)$. Decimos que $\alpha > 0$ cumple*

1. *La condición de **descenso suficiente** (o condición de **Armijo**) si*

$$f(x + \alpha\nu) \leq f(x) + (c_1\alpha)\langle\nabla f(x), \nu\rangle. \tag{3.5}$$

2. *La condición de **curvatura** si*

$$\langle\nabla f(x + \alpha\nu), \nu\rangle \geq c_2\langle\nabla f(x), \nu\rangle. \tag{3.6}$$

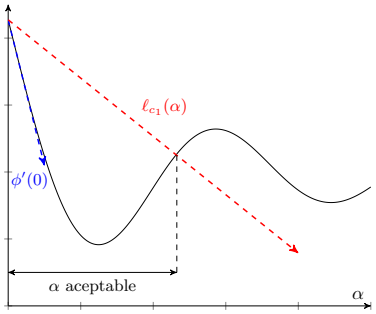
*Finalmente, decimos que α cumple las **condiciones de Wolfe** si satisface ambas condiciones, de descenso suficiente y de curvatura.*

La condición de Armijo (3.5) nos dice que $f(x + \alpha\nu)$ debe estar por debajo de la aproximación lineal afín $\ell_{c_1}(\alpha) := f(x) + (c_1\alpha)\langle\nabla f(x), \nu\rangle$. Como $c_1 < 1$, esto nos dice que ℓ_{c_1} está por sobre la aproximación de Taylor de primer orden de f en el rayo $x + \mathbb{R}_+\nu$, es decir,

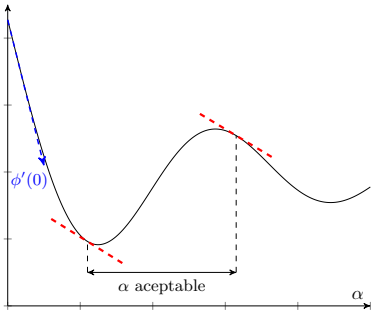
$$\ell_{c_1}(\alpha) > f(x) + \alpha\langle\nabla f(x), \nu\rangle,$$

Universidad de O'Higgins

y por lo tanto siempre es posible encontrar $\alpha > 0$ que satisfaga la condición de Armijo, como ilustra la Figura 3.5a.



(a) Ilustración de condición de Armijo. En negro, la función $\phi(\alpha) = f(x + \alpha\nu)$. En axul, la aproximación de primer orden de ϕ en 0. En rojo, la función $\ell_{c_1}(\alpha)$ que determina la condición de Armijo.



(b) Ilustración de condición de Curvatura. En negro, la función $\phi(\alpha) = f(x + \alpha\nu)$. En azul, la aproximación de primer orden de ϕ en 0. En rojo, las rectas asociadas a las derivadas de borde $\phi'(\alpha) = c_2\phi'(0)$.

La condición de curvatura, por otro lado, es un condición que fuerza a buscar α lejos de 0. En efecto, como $\langle \nabla f(x), \nu \rangle < 0$, la ecuación (3.6) fuerza a que la pendiente $\phi'(\alpha)$ de la función $\phi(\alpha) = f(x+\alpha\nu)$ crezca lo suficiente con respecto a la pendiente $\phi'(0)$. Dado que $\phi' : [0, +\infty) \to \mathbb{R}$ es continua, necesariamente α debe ser al menos suficientemente grande para acercarse al primer punto crítico α^* de ϕ, que es cuando $\phi'(\alpha^*) = 0$.

Cuando $\phi'(0) = \langle \nabla f(x), \nu \rangle$ es muy negativo, entonces la condición de curvatura fuerza a que α esté más lejos de 0. Por el contrario, si $\langle \nabla f(x), \nu \rangle$ es muy cercano a cero, la condición de curvatura comienza a cumplirse para valores de α más cercanos a 0. Esto está ilustrado en la Figura 3.5b.

Observación 3.10. *En implementaciones, debemos elegir los valores de c_1 y c_2 con los que trabajaremos. Normalmente, c_1 se elige como un valor pequeño, típicamente $c_1 = 10^{-4}$. Por otro lado, dependiendo del método que se ocupe, típicamente c_2 se fija en algún valor entre 0.1 y 0.9.*

Las condiciones de Wolfe pueden ser reemplazadas por una versión más fuerte, que considera reforzar la condición de curvatura.

Definición 3.11 (Condiciones fuertes de Wolfe)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^1$, $x \in \mathbb{R}^n$ un punto regular de f y ν una dirección de descenso de f en x. Sea $c_1 \in (0, 1)$ y $c_2 \in (c_1, 1)$. Decimos que $\alpha > 0$ cumple las **condiciones fuertes de Wolfe** si cumple la condición de Armijo (3.5) y la condición de **curvatura fuerte** dada por*

$$|\langle \nabla f(x + \alpha\nu), \nu \rangle| \leq c_2 |\langle \nabla f(x), \nu \rangle|. \tag{3.7}$$

Las condiciones fuertes de Wolfe aseguran decrecimiento suficiente pero además piden que α esté lo suficientemente cerca de un punto crítico de la función $\phi(\alpha) = f(x + \alpha\nu)$.

El siguiente teorema nos asegura que, para problemas acotados, siempre podemos encontrar un paso α que cumpla las condiciones de Wolfe, e incluso las condiciones fuertes de Wolfe.

Teorema 3.12. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^1$, $x \in \mathbb{R}^n$ un punto regular de f y ν una dirección de descenso de f en x. Sean $c_1 \in (0, 1)$ y $c_2 \in (c_1, 1)$. Supongamos además que f es acotada inferiormente en el rayo*

$$x + \mathbb{R}_+\nu = \{x + \alpha\nu \ : \ \alpha \geq 0\}.$$

Universidad de O'Higgins

Entonces

1. *existe $\alpha > 0$ que satisface las condiciones de Wolfe para c_1 y c_2.*
2. *existe $\alpha > 0$ que satisface las condiciones fuertes de Wolfe para c_1 y c_2.*

Demostración. Tomemos la función $\phi : [0, +\infty) \to \mathbb{R}$ dada por $\phi(\alpha) = f(x + \alpha\nu)$ y la función $\ell_{c_1} : [0, +\infty) \to \mathbb{R}$ dada por $\ell_{c_1}(\alpha) = f(x) + (c_1\langle\nabla f(x), \nu\rangle)\alpha$.

Veamos primero que existe $\delta > 0$ tal que $\phi(\alpha) < \ell_{c_1}(\alpha)$ para todo $\alpha \in (0, \delta)$. En efecto, si así no lo fuera, tendríamos que existe una sucesión $\alpha_k \to 0$ tal que $\phi(\alpha_k) \geq \ell_{c_1}(\alpha_k)$, para todo $k \in \mathbb{N}$. Luego, notando que $\phi(0) = \ell_{c_1}(0)$, tendríamos que

$$\begin{aligned}\langle\nabla f(x), \nu\rangle &= \phi'(0) \\ &= \lim_k \frac{\phi(\alpha_k) - \phi(0)}{\alpha_k} \\ &\geq \lim_k \frac{\ell_{c_1}(\alpha_k) - \ell_{c_1}(0)}{\alpha_k} = c_1\langle\nabla f(x), \nu\rangle,\end{aligned}$$

lo cual sería una contradicción, pues $\langle\nabla f(x), \nu\rangle < 0$ y $c_1 \in (0, 1)$.

Ahora, como f es acotada inferiormente a lo largo del rayo $x + \mathbb{R}_+\nu$, tenemos que $\inf_{\alpha\geq 0} \phi > -\infty$. Como $\langle\nabla f(x), \nu\rangle < 0$, tenemos que $\ell_{c_1}(\alpha) \xrightarrow{\alpha\to+\infty} -\infty$. Por lo tanto, notando que $\ell_{c_1}(\delta/2) - \phi(\delta/2) > 0$ y que existe un $\alpha > 0$ lo suficientemente grande tal que $\ell_{c_1}(\alpha) - \phi(\alpha) < 0$, podemos aplicar el Teorema del Valor Intermedio, concluyendo que existe al menos un valor $\bar{\alpha} > 0$ tal que $\phi(\bar{\alpha}) = \ell_{c_1}(\bar{\alpha})$. Definamos

$$\alpha' := \inf\{\alpha > 0 \ : \ \phi(\alpha) = \ell_{c_1}(\alpha)\}.$$

Sabemos que necesariamente $\alpha' > 0$, o de lo contrario llegaríamos a la misma contradicción que al principio de la demostración. Más aún, por el mismo razonamiento, la condición de Armijo (3.5) se cumple para todo $\alpha \in (0, \alpha')$. Luego, aplicando el Teorema del Valor Medio, necesariamente existe $\alpha'' \in (0, \alpha')$ tal que

$$\begin{aligned}\langle\nabla f(x + \alpha''\nu), \nu\rangle = \phi'(\alpha'') &= \frac{\phi(\alpha') - \phi(0)}{\alpha'} \\ &= \frac{\ell_{c_1}(\alpha') - \ell_{c_1}(0)}{\alpha'} = c_1\langle\nabla f(x), \nu\rangle > c_2\langle\nabla f(x), \nu\rangle.\end{aligned}$$

Por lo tanto, α'' cumple también la condición de curvatura (3.6). Más aún, tenemos que

$$|\langle\nabla f(x + \alpha''\nu), \nu\rangle| = c_1|\langle\nabla f(x), \nu\rangle| \leq c_2|\langle\nabla f(x), \nu\rangle|,$$

y por lo tanto α'' cumple también las condiciones fuertes de Wolfe. Esto concluye la demostración. □

Observación 3.13. *Cuando reemplazamos la dirección de descenso $\nu \in \mathbb{R}^n \setminus \{0\}$ por el vector normalizado $\nu/\|\nu\|$, se tiene la siguiente equivalencia:*

$$\begin{gathered}\alpha \text{ cumple las condiciones de Wolfe para } \tfrac{\nu}{\|\nu\|} \\ \Leftrightarrow \\ \tfrac{\alpha}{\|\nu\|} \text{ cumple las condiciones de Wolfe para } \nu.\end{gathered}$$

La misma equivalencia se tiene para las condiciones fuertes de Wolfe. Por lo tanto, es equivalente considerar ν o $\nu/\|\nu\|$ como dirección de descenso, como se mencionó previamente en la Observación 3.7.

Universidad de O'Higgins

El siguiente teorema, aunque a primera vista algo técnico, es la herramienta fundamental que nos permite asegurar que los métodos de búsqueda lineal son exitosos, en el sentido que producen secuencias de iterandos (x_k) que aproximan puntos críticos.

Teorema 3.14 (Zoutendijk). *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^1$, $x_0 \in \mathbb{R}^n$. Supongamos que f es acotada inferiormente y que existe un abierto A con $[f \leq f(x_0)] \subset A$ tal que ∇f es Lipschitz en A, es decir, tal que existe $L > 0$ que cumple que*

$$\|\nabla f(z_1) - \nabla f(z_2)\| \leq L\|z_1 - z_2\|, \quad \forall z_1, z_2 \in A.$$

Sean $c_1 \in (0,1)$, $c_2 \in (c_1, 1)$ y sean tres sucesiones $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$, $(\nu_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n \setminus \{0\}$ y $(\alpha_k)_{k\in\mathbb{N}} \subset (0, +\infty)$ tal que

(i) Para todo $k \in \mathbb{N}$, $x_{k+1} = x_k + \alpha_k \nu_k$ y $f(x_{k+1}) \leq f(x_k)$.

(ii) Para todo $k \in \mathbb{N}$, ν_k es una dirección de descenso de f en x_k.

(iii) Existe $k_0 \in \mathbb{N}$ lo suficientemente grande tal que para todo $k \geq k_0$ el paso α_k satisface las condiciones de Wolfe con c_1 y c_2, para el punto x_k y la dirección ν_k.

Entonces, se tiene que

$$\sum_{k=0}^{\infty} \cos^2(\theta_k)\|\nabla f(x_k)\|^2 < +\infty, \tag{3.8}$$

donde, para todo $k \in \mathbb{N}$, θ_k es el ángulo entre ν_k y $\nabla f(x_k)$.

Demostración. Sin perder generalidad, tomemos $k_0 = 0$. Usando la condición de curvatura (3.6), tenemos que para todo $k \in \mathbb{N}$

$$\langle \nabla f(x_{k+1}) - \nabla f(x_k), \nu_k \rangle = \langle \nabla f(x_k + \alpha_k \nu_k) - \nabla f(x_k), \nu_k \rangle \geq (c_2 - 1)\langle \nabla f(x_k), \nu_k \rangle.$$

Por otro lado, la condición de Armijo (3.5) implica que para todo $k \in \mathbb{N}$, $f(x_{k+1}) < f(x_k)$ y por lo tanto $x_k \in [f \leq f(x_0)] \subset A$. Luego, podemos ocupar la condición de Lipschitz de ∇f para escribir

$$\langle \nabla f(x_{k+1}) - \nabla f(x_k), \nu_k \rangle \leq L\|x_{k+1} - x_k\|\|\nu_k\| = L\alpha_k\|\nu_k\|^2.$$

Mezclando ambas desigualdades, concluimos que para todo $k \in \mathbb{N}$,

$$-\alpha_k \leq \frac{1 - c_2}{L} \frac{\langle \nabla f(x_k), \nu_k \rangle}{\|\nu_k\|^2}.$$

Ocupando nuevamente la condición de Armijo (3.5), tenemos que

$$f(x_{k+1}) \leq f(x_k) - (-\alpha_k)c_1\langle \nabla f(x_k), \nu_k \rangle \leq f(x_k) - \underbrace{c_1 \frac{1 - c_2}{L}}_{=c} \frac{\langle \nabla f(x_k), \nu_k \rangle^2}{\|\nu_k\|^2}.$$

Finalmente, recordando que $\cos(\theta_k) = \langle \frac{\nabla f(x_k), \nu_k \rangle}{\|\nabla f(x_k)\|\|\nu_k\|}$, concluimos que para todo $k \in \mathbb{N}$

$$f(x_{k+1}) \leq f(x_k) - c\cos(\theta_k)^2\|\nabla f(x_k)\|^2.$$

Razonando por inducción, tenemos que para todo $k \in \mathbb{N}$

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

Universidad de O'Higgins

Finalmente, como f es acotada inferiormente, tenemos que $(f(x_k))_k$ es una sucesión decreciente y acotada, y por lo tanto

$$\begin{aligned}\sum_{j=0}^{\infty} \cos(\theta_j)^2 \|\nabla f(x_j)\|^2 &= \lim_k \sum_{j=0}^{k} \cos(\theta_j)^2 \|\nabla f(x_j)\|^2 \\ &\leq \lim_k \frac{f(x_0) - f(x_{k+1})}{c} \\ &\leq \frac{f(x_0) - \inf f}{c} < +\infty.\end{aligned}$$

□

Bajo las hipótesis del teorema de Zoutendijk, podemos dar condiciones suficientes para que la sucesión $(x_k)_{k\in\mathbb{N}}$ aproxime un punto crítico, en el sentido que

$$\lim_k \|\nabla f(x_k)\| = 0. \tag{3.9}$$

Cuando una sucesión (x_k) producida por un algoritmo iterativo satisface (3.9), decimos que (x_k) es *globalmente convergente*. Esta es la convergencia más fuerte que podemos lograr, y nos dice que mientras más grande sea la iteración k, más se parece x_k a un punto crítico de f.

La condición suficiente es que la direcciones de descenso ν_k nunca estén muy cerca de ser ortogonales al gradiente $\nabla f(x_k)$, es decir, que θ_k esté siempre "lejos"de $\pi/2$.

Corolario 3.15. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^1$, $x_0 \in \mathbb{R}^n$, y $(x_k) \subset \mathbb{R}^n$ una sucesión de la forma*

$$x_{k+1} = x_k + \alpha_k \nu_k, \quad \forall k \in \mathbb{N}.$$

Supongamos que se tienen las condiciones del Teorema 3.14 y además supongamos que existen $k_0 \in \mathbb{N}$ y $\delta > 0$ tal que

$$\forall k \geq k_0, \quad \cos(\theta_k) \geq \delta.$$

Entonces, (x_k) es globalmente convergente, es decir, $\|\nabla f(x_k)\| \to 0$. En particular, si $\frac{\nu_k}{\|\nu_k\|} = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$, entonces (x_k) es globalmente convergente.

3.2.2. Algoritmo de Backtrack para encontrar pasos inexactos

En esta sección veremos cómo calcular un paso $\alpha > 0$ que cumpla las condiciones de Wolfe, dadas una función $f : \mathbb{R}^n \to \mathbb{R}$, un punto regular $x \in \mathbb{R}^n$ y una dirección de descenso ν.

La primera opción es encontrar $\alpha > 0$ como el mínimo global del problema (3.4). Cuando esto es posible, decimos que α es un *paso exacto* para f en x y con dirección ν.

La segunda opción es buscar computacionalmente un paso α a través del rayo $x + \mathbb{R}_+\nu$. Para esto, normalmente tomamos un parámetro $\alpha_{\text{máx}} > 0$ fijo (en algunos algoritmos se toma $\alpha_{\text{máx}} = 1$), asumimos que $\|\nu\| = 1$ y definimos

$$\alpha := \begin{cases} \alpha_{\text{máx}} & \text{si } \alpha_{\text{máx}} \text{ cumple la condición de Armijo} \\ \alpha' \in (0, \alpha_{\text{máx}}) & \text{si no,} \end{cases}$$

donde α' es un valor encontrado algorítmicamente y que cumple ambas condiciones de Wolfe. Cuando el paso α es construido de esta manera, se dice que es un *paso inexacto* para f en x y con dirección ν. Observe que fijar un valor máximo α_0 para buscar el paso inexacto aún preserva los resultados del Teorema 3.14 y del Corolario 3.15 (Ver Problema **P1.** de la Sección 3.6).

Universidad de O'Higgins

Para encontrar el paso $\alpha_k \in \mathbb{N}$, primero buscamos un un primer valor $\alpha^* \in (0, \alpha_{\text{máx}}]$ que cumpla la condición de Armijo, ocupando lo que se conoce como *algoritmo de Backtrack*.

Algoritmo 3.1: Algoritmo de Backtrack

Entrada: Función objetivo f; Punto x; Dirección de descenso ν con $\|\nu\| = 1$; $\alpha_{\text{máx}} > 0$; $c_1 \in (0,1)$; $\rho \in (0,1)$.
$\alpha \leftarrow \alpha_{\text{máx}}$;
while $f(x + \alpha\nu) > f(x) + (c_1\alpha)\langle \nabla f(x), \nu\rangle$ **do**
  $\alpha \leftarrow \rho\alpha$;
end
Return α; **Salida:** α que cumple condición de Armijo.

El Algoritmo 3.1 comienza probando si $\alpha = \alpha_{\text{máx}}$ satisface la condición de Armijo. En tal caso, devuelve $\alpha_{\text{máx}}$, pues nunca entra al ciclo-while. Si no, comienza a contraer α reemplazándolo por $\rho\alpha$, hasta alcanzar un valor de α lo suficiente pequeño tal que la condición de Armijo se satisfaga. El resultado es el primer valor en $\{\rho^j \alpha_{\text{máx}} \ : \ j \in \mathbb{N}\}$ que cumple dicha condición. Dado que ν es una dirección de descenso, se tiene necesariamente que el Algoritmo 3.1 termina en finitos pasos.

Ahora que sabemos encontrar valores que cumplan la condición de Armijo, podemos de hecho buscar α que cumpla las condiciones de Wolfe. El paso clave es el siguiente lema.

Lema 3.16. *Sea $(f, x, \alpha_{\text{máx}}, c_1, \rho)$ la entrada del Algoritmo 3.1 y $c_2 \in (c_1, 1)$. Sean $\alpha, \beta \in (0, \alpha_{\text{máx}})$ tal que $\alpha < \beta$, α cumple la condición de Armijo (3.5), y β no. Entonces, al menos una de las siguientes afirmaciones es verdadera:*

1. *α satisface las condiciones de Wolfe.*
2. *existe un subintervalo $(a, b) \subset (\alpha, \beta)$ que satisface las condiciones de Wolfe.*

Demostración. Supongamos que la primera afirmación no se tiene. Eso significa que α no satisface la condición de curvatura (3.6), es decir,

$$\langle \nabla f(x + \alpha\nu), \nu\rangle < c_2\langle \nabla f(x), \nu\rangle.$$

Definamos $\phi : [0, +\infty) \to \mathbb{R}$ como $\phi(t) = f(x+t\nu)$ y $\ell : [0, +\infty) \to \mathbb{R}$ como $\ell(t) = f(x)+c_1 t\langle f(x), \nu\rangle$. Como $c_2 > c_1$ y $\langle \nabla f(x), \nu\rangle < 0$, tenemos que de hecho

$$\phi'(\alpha) = \langle \nabla f(x + \alpha\nu), \nu\rangle < c_2\langle \nabla f(x), \nu\rangle < c_1\langle \nabla f(x), \nu\rangle = \ell'(\alpha).$$

Repitiendo el razonamiento del Teorema 3.12, sabemos que

$$r = \text{ínf}\{t > \alpha \ : \ \phi(t) = \ell(t)\} \in (\alpha, \beta),$$

y más aún, para todo $t \in (\alpha, r)$, tenemos que $\phi(t) < \ell(t)$, es decir, t satisface la condición de Armijo (3.5). Luego, ocupando el Teorema del Valor Medio para la función $\phi - \ell$ en el intervalo $[\alpha, r]$, sabemos que existe $b \in (\alpha, r)$ tal que

$$0 = \phi'(b) = \langle \nabla f(x + b\nu), \nu\rangle - c_1\langle \nabla f(x), \nu\rangle.$$

Luego,

$$\langle \nabla f(x + b\nu), \nu\rangle = c_1\langle \nabla f(x), \nu\rangle > c_2\langle \nabla f(x), \nu\rangle.$$

Como ∇f es continua, tenemos que existe δ lo suficientemente pequeño tal que $a = b - \delta > \alpha$ y además

$$\langle \nabla f(x + t\nu) > c_2\langle \nabla f(x), \nu\rangle, \quad \forall t \in (a, b),$$

lo que concluye la demostración. □

Universidad de O'Higgins

Algoritmo 3.2: Algoritmo de Búsqueda lineal de paso inexacto (LineSearch)

Entrada: Función objetivo f; Punto x; Dirección de descenso ν; $\alpha_{\text{máx}} > 0$; $c_1 \in (0,1)$; $c_2 \in (c_1, 1)$; $\rho \in (0,1)$.
Normalizar ν: $\nu' \leftarrow \nu/\|\nu\|$;
$\alpha \leftarrow \text{Bactrack}(f, x, \nu', \alpha_{\text{máx}}, c_1, \rho)$;
if *($\alpha = \alpha_{\text{máx}}$) or ($\alpha$ cumple curvatura)* **then**
    $\alpha^* = \alpha$;
    Return $\alpha^* = \|\nu\|^{-1}\alpha^*$;
end
$\beta \leftarrow \alpha/\rho$;
while *TRUE* **do**
    $\alpha^* \leftarrow \frac{\alpha+\beta}{2}$;
    if $f(x+\alpha^*\nu) > f(x) + (c_1\alpha^*)\langle\nabla f(x), \nu\rangle$ **then**
        $\beta \leftarrow \alpha^*$; `// Si falla Armijo, bajamos la cota superior`
    else if $\langle\nabla f(x+\alpha^*\nu), \nu\rangle < c_2\langle\nabla f(x), \nu\rangle$ **then**
        $\alpha \leftarrow \alpha^*$; `// Si falla Curvatura, subimos la cota inferior.`
    else
        Return $\alpha^* = \|\nu\|^{-1}\alpha^*$;
    end
end
Salida: α^*; que cumple las condiciones de Wolfe.

El Lema 3.16 nos asegura que podemos encontrar un paso que cumpla las condiciones de Wolfe en finitas iteraciones mediante el Algoritmo 3.2. Lo primero es normalizar la dirección de descenso y ocupar el Algoritmo de Backtrack para producir un primer candidato $\alpha \in (0, \alpha_{\text{máx}}]$. Hay dos casos de borde: 1) Si $\alpha = \alpha_{\text{máx}}$, entonces estamos en la cota superior y terminamos la búsqueda. 2) Si α satisface la condición de curvatura, entonces ya tenemos el paso que cumple lo requerido.

Si estas situaciones de borde no se tienen, siginifica que el Algoritmo 3.1 entregó $\alpha = \rho^j\alpha_{\text{máx}}$ con $j > 0$. Eso significa que $\beta = \alpha/\rho = \rho^{j-1}\alpha_{\text{máx}}$ no satisface la condición de Armijo. Por lo tanto, el lema 3.16 nos asegura que hay un intervalo abierto $(a, b) \subset (\alpha, \beta)$ donde cualquier paso $\alpha^* \in (a, b)$ cumple ambas condiciones de Wolfe. Para llegar a un punto en este intervalo, hacemos búsqueda binaria: Dividimos el intervalo (α, β) por su punto medio $\alpha^* = (\alpha + \beta)/2$. Si α^* no es el punto que buscamos, debemos quedarnos con uno de los subintervalos (α, α^*) o (α^*, β) y volver a dividir. Para esto, hay que asegurarse que el nuevo intervalo cumpla las condiciones del Lema 3.16: El extremo inferior debe satisfacer Armijo, y el superior no. En finitos pasos, llegaremos al paso α^* deseado.

En la práctica, existes varios métodos diferentes para fijar $\alpha_{\text{máx}}$ y también para seleccionar cómo actualizar α^* en la línea 10 del Algoritmo 3.2, diferentes a simplemente tomar el punto medio de los extremos α y β. Muchas veces, debido a errores de precisión numérica, es posible que la búsqueda del valor α^* no se logre con éxito, y por lo tanto, se incluyen criterios de parada auxiliares, como por ejemplo detener la búsqueda después de un número fijo (usualmente 10) iteraciones del while-loop de las líneas 9-18 del Algoritmo 3.2. Cuando la búsqueda se interrumpe, se entrega el valor α^* como el extremo inferior del intervalo $[\alpha, \beta]$, que al menos siempre satisface la condición de Armijo.

3.3 Método de Máximo Descenso

Utilizando el Algoritmo 3.2, es posible finalmente formalizar el método de máximo descenso. Si bien

Universidad de O'Higgins

el algoritmo iterativo (3.3) queda correctamente definido, es necesario incluir un criterio de Parada (como se discutió en la Sección 1.5), pues de lo contrario el proceso podría no terminar. En general se consideran tres criterios de parada:

1. La norma del gradiente es lo suficientemente pequeña, es decir, $\|\nabla f(x_k)\| \leq \varepsilon$, para una tolerancia $\varepsilon > 0$ dada.
2. La distancia entre el iterando x_k y el iterando anterior x_{k-1} es suficientemente pequeña, es decir, $\|x_k - x_{k-1}\| \leq \varepsilon$, para una tolerancia $\varepsilon > 0$ dada.
3. El número de iteraciones supera cierto umbral, es decir $k \geq \text{MaxIter}$, para algún $\text{MaxIter} \in \mathbb{N}$ dado. En tal caso, el algoritmo se detiene pues ya ha pasado demasiado tiempo sin obtener resultados de convergencia favorables.

La implementación del Método de Máximo Descenso con estos criterios se puede ver en el Algoritmo 3.3.

Algoritmo 3.3: Método de Máximo Descenso (SteepestDescent)

Entrada: Función objetivo f; Función gradiente ∇f; Punto inicial x_0; Tolerancia $\varepsilon > 0$.
$x_{\text{curr}} \leftarrow x_0$; $x_{\text{next}} \leftarrow x_0$;
$\nu \leftarrow -\nabla f(x_{\text{curr}})$;
if $\|\nu\| \leq \varepsilon$ **then**
    $x^* \leftarrow x_{\text{curr}}$;
    Return $(x^*, \nabla f(x^*))$;
end
for k = *1:MaxIter* **do**
    $\alpha \leftarrow$ `MetodoPaso`$(f, x_{\text{curr}}, \nu)$; `// Puede ser exacto, o inexacto como LineSeach`
    $x_{\text{next}} \leftarrow x_{\text{curr}} + \alpha\nu$;
    if $\|x_{\text{next}} - x_{\text{curr}}\| \leq \varepsilon$ ***or*** $\|\nabla f(x_{\text{next}})\| \leq \varepsilon$ **then**
        $x^* \leftarrow x_{\text{next}}$;
        Break;
    end
    $x_{\text{curr}} \leftarrow x_{\text{next}}$; $\nu \leftarrow -\nabla f(x_{\text{curr}})$;
end
Return $(x^*, \nabla f(x^*))$;
Salida: x^* punto crítico aproximado con $f(x^*) \leq f(x_0)$, y su gradiente $\nabla f(x^*)$.

Es común tomar $\varepsilon = 10^{-6}$ como tolerancia en implementaciones. El caso ideal es cuando tanto el primer criterio como el segundo criterio se cumplen: hemos encontrado un punto que aproxima lo suficientemente bien a un punto crítico. Cuando el algoritmo para únicamente con el primer o con el segundo criterio, también se considera correcto: si bien decimos que el algoritmo convergió, no tenemos (a priori) garantías de que este punto sea cercano a un punto crítico. Sin embargo, en muchos casos, la detención ocurre exactamente pues estamos cerca de un óptimo local. El último criterio se ocupa para asegurar que el algoritmo se detenga. Normalmente es el peor escenario, y nos dice que el método de optimización no fue exitoso.

La idea de devolver el gradiente $\nabla f(x^*)$ en la salida del Algoritmo 3.3 es para evaluar la condición de primer orden aproximada $\nabla f(x^*) \sim 0$. Con esto, y agregando un indicador del criterio de parada con el que el algoritmo terminó, es posible saber si `SteepestDescent` fue exitoso (criterios 1 y 2), parcialmente exitoso (solo criterio 1 o solo criterio 2), o falló (criterio 3). Además, podemos medir el nivel de éxito: Si $\nabla f(x^*) \sim 10^{-6}$, significa que x^* es más bien un punto estacionario aproximado. En cambio, si $\nabla f(x^*) \sim 10^{-12}$, computacionalmente hablando, el punto x^* es prácticamente un punto crítico (10^{-12} es prácticamente igual a 0).

Universidad de O'Higgins

3.3.1. Radio de convergencia del Método de Máximo Descenso

Gracias al teorema de Zoutendijk (Teorema 3.14), podemos asegurar condiciones suficientes sobre la función f de tal manera que el Algoritmo 3.3 entregue un punto crítico (aproximado) x^* como resultado. Sin embargo, nos gustaría saber qué tan rápido este algoritmo es capaz de encontrar dicho punto x^*.

Para responder esta pregunta, estudiaremos el radio de convergencia (ver Definición 1.5) de la sucesión (x_k) que genera Método de Máximo Descenso. Partamos por un caso ideal, donde la función f es una *forma cuadrática* pura, es decir,

$$f(x) = \frac{1}{2}x^T Q x, \tag{3.10}$$

con $Q \in \mathcal{M}_{n\times n}(\mathbb{R})$ simétrica definida positiva. Sabemos que en este caso f es una función convexa y que su único mínimo global el punto $x^* = 0$.

Veremos ahora cómo se comporta el Método de Máximo Descenso en este caso. Sabiendo que $\nabla f(x) = Qx$, podemos calcular el paso exacto para la función f.

Proposición 3.17. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ la forma cuadrática dada por la ecuación (3.10) y sea x un punto regular de f. Tomando $\nu = -\nabla f(x)$ tenemos que el paso exacto $\alpha > 0$ que resuelve (3.4) está dado por*

$$\alpha = \frac{\langle \nabla f(x), \nabla f(x)\rangle}{\nabla f(x)^T Q \nabla f(x)}. \tag{3.11}$$

Demostración. Tomemos $\phi : \mathbb{R} \to \mathbb{R}$ dada por $\phi(t) = f(x + t\nu)$. Como f es convexa y coercitiva, sabemos que ϕ también es convexa y coercitiva. Más aún, la derivada de ϕ está dada por

$$\begin{aligned}\phi'(t) &= \langle \nabla f(x + t\nu), \nu\rangle \\ &= \langle Q(x + t\nu), \nu\rangle \\ &= t\langle Q\nu, \nu\rangle - \|\nu\|^2 \\ &= t\nu^T Q\nu - \|\nu\|^2.\end{aligned}$$

Sea ahora α un mínimo global de ϕ. Ocupando el Teorema 2.29, sabemos que

$$0 = \phi'(\alpha) = \alpha\nu^T Q\nu - \|\nu\|^2,$$

lo que implica que $\alpha = \frac{\|\nu\|^2}{\nu^T Q\nu} = \frac{\langle \nabla f(x), \nabla f(x)\rangle}{\nabla f(x)^T Q \nabla f(x)} > 0$. □

Cuando ocupamos el paso exacto (3.11) en el Método de Máximo Descenso, es decir, tomamos

$$x_{k+1} = x_k - \frac{\langle \nabla f(x_k), \nabla f(x_k)\rangle}{\nabla f(x_k)^T Q \nabla f(x_k)} \nabla f(x_k), \tag{3.12}$$

podemos estimar la distancia de cada iterando x_k al punto óptimo $x^* = 0$ usando los valores propios de la matriz Q. Para esto, necesitamos el siguiente Lema.

Lema 3.18 (Desigualdad de Kantorovich)**.** *Sea $Q \in \mathcal{M}_{n\times n}(\mathbb{R})$ una matriz simétrica definida positiva, con $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ sus valores propios. Entonces, para cualquier vector $y \in \mathbb{R}^n \setminus \{0\}$, se tiene que*

$$\frac{\langle y, y\rangle^2}{(y^T Q y)(y^T Q^{-1} y)} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2}. \tag{3.13}$$

Universidad de O'Higgins

Demostración. Fijemos $y \in \mathbb{R}^n \setminus \{0\}$. Como Q es una matriz simétrica definida positiva, admite una descomposición

$$Q = P\Lambda P^T,$$

donde $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ y P es una matriz ortonormal (es decir, donde las columnas de P forman una base ortogonal unitaria de $\mathbb{R}^n$). Sea $x = P^T y$. Recordando que P es invertible con $P^{-1} = P^T$ y definiendo $\xi_i = x_i^2/\|x\|^2$ para cada $i = 1, \ldots, n$, tenemos que

$$\begin{aligned}
\frac{\langle y, y\rangle^2}{(y^TQy)(y^TQ^{-1}y)} &= \frac{\langle Px, Px\rangle^2}{((Px)^TQ(Px))((Px)^TQ^{-1}(Px))} \\
&= \frac{(x^TP^TPx)^2}{(x^TP^TP\Lambda P^TPx)(x^TP^TP\Lambda^{-1}P^TPx)} \\
&= \frac{(x^Tx)^2}{(x^TP^TQPx)(x^TP^TQ^{-1}Px)} \\
&= \frac{\langle x, x\rangle^2}{(x^T\Lambda x)(x^T\Lambda^{-1}x)} \\
&= \frac{1}{\left(\sum_{i=1}^n \lambda_i\xi_i\right)\left(\sum_{i=1}^n \lambda_i^{-1}\xi_i\right)}.
\end{aligned} \tag{3.14}$$

Sea $\bar{\lambda} = \sum_{i=1}^n \lambda_i\xi_i$. Como $\xi_i \geq 0$ para todo $i = 1, \ldots, n$, y $\sum_{i=1}^n \xi_i = 1$, tenemos que $\bar{\lambda} \in [\lambda_1, \lambda_n]$, al ser combinación convexa de elementos de dicho intervalo.

Sea $\phi : (0, +\infty) \to (0, +\infty)$ la función dada por $\phi(t) = t^{-1}$. Sabemos que ϕ es convexa (ver **P8.** de la Sección 2.6). Además que para todo $i = 1, \ldots, n$, podemos escribir,

$$\lambda_i = \frac{\lambda_n - \lambda_i}{\lambda_n - \lambda_1}\lambda_1 + \frac{\lambda_i - \lambda_1}{\lambda_n - \lambda_1}\lambda_n.$$

Luego, podemos escribir:

$$\begin{aligned}
\sum_{i=1}^n \lambda_i^{-1}\xi_i = \sum_{i=1}^n \phi(\lambda_i)\xi_i &\leq \sum_{i=1}^n \left(\frac{\lambda_n - \lambda_i}{\lambda_n - \lambda_1}\phi(\lambda_1) + \frac{\lambda_i - \lambda_1}{\lambda_n - \lambda_1}\phi(\lambda_n)\right)\xi_i \\
&= \sum_{i=1}^n \left(\frac{\lambda_n - \lambda_i}{\lambda_1(\lambda_n - \lambda_1)} + \frac{\lambda_i - \lambda_1}{\lambda_n(\lambda_n - \lambda_1)}\right)\xi_i \\
&= \sum_{i=1}^n \left(\frac{\lambda_n^2 - \lambda_n\lambda_i + \lambda_1\lambda_i - \lambda_1^2}{\lambda_1\lambda_n(\lambda_n - \lambda_1)}\right)\xi_i \\
&= \sum_{i=1}^n \left(\frac{\lambda_n + \lambda_1 - \lambda_i}{\lambda_1\lambda_n}\right)\xi_i = \frac{\lambda_n + \lambda_1 - \bar{\lambda}}{\lambda_1\lambda_n}.
\end{aligned}$$

Mezclando el desarrollo anterior con (3.14), concluimos que

$$\begin{aligned}
\frac{\langle y, y\rangle^2}{(y^TQy)(y^TQ^{-1}y)} &= \frac{1}{\bar{\lambda}\left(\sum_{i=1}^n \lambda_i^{-1}\xi_i\right)} \\
&\geq \frac{\lambda_1\lambda_n}{\bar{\lambda}(\lambda_1 + \lambda_n - \bar{\lambda})} \\
&\geq \frac{\lambda_1\lambda_n}{\operatorname{máx}_{\lambda\in[\lambda_1\lambda_n]}\lambda(\lambda_1 + \lambda_n - \lambda)} = \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2},
\end{aligned}$$

donde la última igualdad se sigue del hecho que el máximo de la función $\lambda \mapsto \lambda(\lambda_1 + \lambda_n - \lambda)$ se alcanza en $\lambda^* = \frac{\lambda_1+\lambda_n}{2}$. Esto concluye la demostración. □

Universidad de O'Higgins

Teorema 3.19. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una forma cuadrática dada por (3.10), y sean $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ los valores propios de Q. Para un punto inicial cualquiera $x_0 \in \mathbb{R}^n$, considere la sucesión $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ generada por el Método de Máximo Descenso con paso exacto, es decir, dada por la fórmula (3.12), se tiene que*

$$f(x_{k+1}) \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 f(x_k), \qquad \forall k \in \mathbb{N}.$$

Demostración. Sea $k \in \mathbb{N}$ fijo. Si $\nabla f(x_k) = 0$, entonces $x_k = 0$ y por lo tanto $x_{k+1} = x_k$. En tal caso, el resultado del teorema se deduce trivialmente. Supongamos entonces que $\nabla f(x_k) \neq 0$. Siguiendo la fórmula (3.12) y recordando que $\nabla f(x_k) = Qx_k$, podemos escribir

$$\begin{aligned}
f(x_k) - f(x_{k+1}) &= \frac{1}{2}x_k^T Q x_k - \frac{1}{2}(x_k + \alpha_k \nabla f(x_k))^T Q (x_k + \alpha_k \nabla f(x_k)) \\
&= \alpha_k \nabla f(x_k)^T Q x_k - \frac{\alpha_k^2}{2}\nabla f(x_k)^T Q \nabla f(x_k) \\
&= \alpha_k \langle \nabla f(x_k), \nabla f(x_k)\rangle - \frac{\alpha_k^2}{2}\nabla f(x_k)^T Q \nabla f(x_k) \\
&= \frac{\langle \nabla f(x_k), \nabla f(x_k)\rangle^2}{\nabla f(x_k)^T Q \nabla f(x_k)} - \frac{1}{2}\frac{\langle \nabla f(x_k), \nabla f(x_k)\rangle^2}{\nabla f(x_k)^T Q \nabla f(x_k)} \\
&= \frac{1}{2}\frac{\langle \nabla f(x_k), \nabla f(x_k)\rangle^2}{\nabla f(x_k)^T Q \nabla f(x_k)}
\end{aligned}$$

Usando la desigualdad de Kantorovich (3.13), tenemos que

$$f(x_k) - f(x_{k+1}) \geq \frac{2\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2}\nabla f(x_k)^T Q^{-1} \nabla f(x_k)$$

Ahora, notando nuevamente que $\nabla f(x_k) = Qx_k$ y que Q es simétrica, tenemos que

$$\nabla f(x_k)^T Q^{-1} \nabla f(x_k) = x_k^T Q Q^{-1} Q x_k = 2f(x_k).$$

Así, concluimos que

$$f(x_{k+1}) \leq \left(1 - \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2}\right) f(x_k) = \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 f(x_k).$$

□

El teorema anterior nos dice que el Método de Máximo Descenso para formas cuadráticas puras tiene radio de convergencia lineal. En efecto, basta notar que la función $f(x) = \frac{1}{2}x^T Q x$ induce una norma en $\mathbb{R}^n$, que podemos denotamos por $\|\cdot\|_Q$, y que está definida por

$$\|x\|_Q = \sqrt{x^T Q x}. \tag{3.15}$$

Usando esta norma, la desigualdad (3.12) se puede escribir como

$$\|x_{k+1}\|_Q \leq r\|x_k\|_Q, \quad \forall k \in \mathbb{N}, \tag{3.16}$$

donde $r = \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right) < 1$. Es decir, la sucesión $(x_k)_k$ tiene radio de convergencia lineal para la norma $\|\cdot\|_Q$.

Si bien la desigualdad (3.16) es sólo una cota superior, en general es un buen estimador del radio de convergencia del Método de Máximo Descenso. Dicho de otro modo, aunque en teoría el radio

Universidad de O'Higgins

de convergencia podría ser mejor que lineal, en la práctica resulta que este radio de convergencia (lineal) es el mejor que podemos esperar para el Método de Máximo Descenso en problemas no-lineales genéricos.

Finalmente, es importante destacar que, dependiendo del problema, el radio de convergencia lineal puede llegar a ser bastante insatisfactorio. De hecho, en el mismo problema cuadrático, si el menor valor propio λ_1 de Q es muy pequeño en comparación al mayor valor propio λ_n de Q, entonces el factor r en (3.16) puede llegar a ser muy cercano a 1, produciendo un decrecimiento extremadamente lento. Por ejemplo, si $\lambda_1 = 1$ y $\lambda_n = 999$, tendríamos que $r = 0.998$. Si partimos de un punto inicial $x_0 \in \mathbb{R}^n$ con $f(x_0) = 1$ y realizamos 1000 iteraciones, sólo podríamos asegurar que $f(x_{1000}) \leq 0.136$, lo cual en la práctica (para tolerancias del tipo $\varepsilon = 10^{-6}$) es aún demasiado lejos del óptimo $f(x^*) = 0$.

3.4 Método de Newton

El Método de Máximo Descenso se basa en considerar, en cada iteración $k \in \mathbb{N}$, la dirección de descenso como $\nu_k = -\nabla f(x_k)$. Las ventajas de este método son dos: 1) Es fácil de implementar y 2) Cada iteración es "barata" computacionalmente, en el sentido que calcular la dirección ν_k y el paso α_k consume pocos recursos (tiempo de cómputo y memoria). Sin embargo, como discutimos en la sección anterior, las garantías de convergencia no son muy buenas, pues el radio de convergencia de este método es, en general, solo lineal: El Método de Máximo Descenso se vuelve costoso pues puede requerir demasiadas iteraciones antes de entregar un iterando x_k que satisfaga los criterios de parada.

En esta sección estudiaremos un segundo tipo de métodos, adaptado para funciones f de clase $\mathcal{C}^2$, llamados Métodos de Newton. Este tipo de métodos consiste en ocupar información de segundo orden (el Hessiano $\nabla^2 f(x_k)$) para construir la dirección de descenso, y se basan en el siguiente Lema elemental:

Lema 3.20. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^1$ y $x \in \mathbb{R}^n$ un punto regular de f. Entonces, para toda matriz $B \in \mathcal{M}_{n\times n}(\mathbb{R})$ simétrica y definida positiva, se tiene que*

$$\nu = -B^{-1}\nabla f(x)$$

es una dirección de descenso de f en x.

Demostración. Sea B una matriz simétrica y definida positiva. Se tiene entonces que B^{-1} también es definida positiva y por lo tanto

$$f'(x;\nu) = \langle \nabla f(x), \nu\rangle = -\nabla f(x)^T B^{-1}\nabla f(x) < 0.$$

Por lo tanto, $\nu = B^{-1}\nabla f(x)$ es dirección de descenso de f en x. □

3.4.1. Método de Newton puro

Sea f una función de clase $\mathcal{C}^2$ y $x \in \mathbb{R}^n$ un punto regular de f tal que el Hessiano $\nabla^2 f(x)$ es una matriz definida positiva. El Método de Newton consiste en elegir la dirección de descenso de f en x como

$$\nu^N = -\nabla^2 f(x)^{-1}\nabla f(x). \tag{3.17}$$

Si ahora aplicamos las iteraciones de búsqueda lineal (3.3), para poder usar el Método de Newton necesitamos que la matriz Hessiana $\nabla^2 f(x_k)$ en cada iterando sea definida positiva. Esto claramente

Universidad de O'Higgins

es una limitante, pues en general podemos partir de un punto inicial x_0 donde dicha condición no se cumpla. Sin embargo, en los casos donde si podemos asegurar esta condición, utilizar la dirección de descenso (3.17) se vuelve extremadamente eficiente.

Teorema 3.21 (Radio de Convergencia del Método de Newton)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^2$ y $x^* \in \mathbb{R}^n$. Supongamos que*

1. *El punto x^* es un punto crítico de f y que $\nabla^2 f(x^*)$ es definida positiva.*
2. *Existe $\delta > 0$ tal que $\nabla^2 f : \mathbb{R}^n \to \mathcal{M}_{n\times n}(\mathbb{R})$ es Lipschitz en la bola $B(x^*, \delta)$.*

Sea ahora $x_0 \in \mathbb{R}^n$. Considere la sucesión (x_k) producida por el algoritmo de búsqueda lineal (3.3), con $\alpha_k = 1$ y con ν_k como la dirección de Newton (3.17) en el punto x_k, es decir,

$$x_{k+1} = x_k + \underbrace{(-\nabla^2 f(x_k)^{-1}\nabla f(x_k))}_{=\nu_k^N}, \quad \forall k \in \mathbb{N}.$$

Si x_0 está lo suficientemente cerca de x^*, se tiene entonces que

(i) La sucesión $(x_k)_k$ está bien definida, y converge a x^* con radio de convergencia cuadrático.

(ii) La sucesión $(\|\nabla f(x_k)\|)_k$ converge a 0 con radio de convergencia cuadrático.

Demostración. Tomemos $x_0 \in B(x^*, \varepsilon)$ donde $\varepsilon > 0$ es menor que δ y es tal que $\nabla^2 f(x)$ es definida positiva para todo $x \in B(x^*, \varepsilon)$. Mientras avancemos en la demostración, actualizaremos el valor de ε cuando sea necesario.

Supongamos primero que $x_k \in B(x^*, \varepsilon)$. Entonces podemos escribir

$$\begin{aligned} x_{k+1} - x^* &= x_k + \nu_k^N - x^* \\ &= x_k - x^* - \nabla^2 f(x_k)^{-1}\nabla f(x_k) \\ &= \nabla^2 f(x_k)^{-1}[\nabla^2 f(x)(x_k - x^*) - \nabla f(x_k)] \\ &= \nabla^2 f(x_k)^{-1}[\nabla^2 f(x_k)(x_k - x^*) - (\nabla f(x_k) - \nabla f(x^*))]. \end{aligned}$$

Ocupando la forma integral del teorema de Taylor para ∇f, podemos escribir

$$\begin{aligned} \nabla f(x_k) - \nabla f(x^*) &= -(\nabla f(x^*) - \nabla f(x_k)) \\ &= -\int_0^1 \nabla^2 f(x_k + t(x^* - x_k))(x^* - x_k)dt \\ &= \int_0^1 \nabla^2 f(x_k + t(x^* - x_k))(x_k - x^*)dt \end{aligned}$$

Luego, tenemos que

$$\begin{aligned} &\|\nabla^2 f(x_k)(x_k - x^*) - (\nabla f(x_k) - \nabla f(x^*))\| \\ &= \left\|\nabla^2 f(x_k)(x_k - x^*) - \int_0^1 \nabla^2 f(x_k + t(x^* - x_k))(x_k - x^*)dt\right\| \\ &\leq \int_0^1 \left\|\nabla^2 f(x_k) - \nabla^2 f(x_k + t(x^* - x_k))\right\| \|x_k - x^*\|dt \\ &\leq \int_0^1 tL\|x_k - x^*\|^2 dt = \frac{L}{2}\|x_k - x^*\|^2, \end{aligned}$$

Universidad de O'Higgins

donde $L > 0$ es la constante de Lipschitz de $\nabla^2 f$ en $B(x^*, \delta)$. Como la función $x \mapsto \nabla^2 f(x)^{-1}$ es continua en $B(x^*, \varepsilon)$, podemos achicar ε si es necesario de tal manera que

$$\|\nabla^2 f(x)^{-1}\| \leq 2\|\nabla^2 f(x^*)^{-1}\|, \quad \forall x \in B(x^*, \varepsilon).$$

Finalmente, concluimos que

$$\|x_{k+1} - x^*\| \leq \frac{L}{2}\|\nabla^2 f(x_k)^{-1}\|\|x_k - x^*\|^2 \leq \underbrace{L\|\nabla^2 f(x^*)^{-1}\|}_{=K}\|x_k - x^*\|^2$$

Finalmente, actualicemos nuevamente ε tomando $\varepsilon = \min\{\varepsilon, (2K)^{-1}\}$. Veamos ahora que para todo $k \in \mathbb{N}$, $x_k \in B(x^*, \varepsilon)$ y que $x_k \to x^*$. Para esto, mostremos por inducción que $\|x_k - x^*\| < \varepsilon/2^k$:

- **Caso base:** $x_0 \in B(x^*, \varepsilon)$ por hipótesis y por lo tanto $\|x_0 - x^*\| < \varepsilon/2^0$.
- **Paso inductivo:** Asumamos que $\|x_k - x^*\| < \varepsilon/2^k$. Entonces, el desarrollo de la primera parte de la demostración es válido (pues $x_k \in B(x^*, \varepsilon)$) y por lo tanto

$$\|x_{k+1} - x^*\| \leq K\|x_k - x^*\|^2 < K\varepsilon\|x_k - x^*\| < \frac{K}{2K}\frac{\varepsilon}{2^k} = \frac{\varepsilon}{2^{k+1}}.$$

Concluimos entonces que la sucesión (x_k) siempre se mantiene dentro de $B(x^*, \varepsilon)$ y además que $x_k \to x^*$ con radio de convergencia cuadrático.

Ahora, dado que ∇f es continua, es claro que $\nabla f(x_k) \to \nabla f(x^*) = 0$. Solo resta ver que esta convergencia también es con radio cuadrático. Notando que

$$x_{k+1} - x_k = \nu_k^N \quad \text{y } \nabla f(x_k) + \nabla^2 f(x_k)\nu_k^N = 0,$$

podemos escribir, usando nuevamente el Teorema de Taylor, que

$$\begin{aligned}
\|\nabla f(x_{k+1})\| &= \|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)\nu_k^N\| \\
&= \left\| \int_0^1 \nabla^2 f(x_k + t\nu_k^N)\nu_k^N dt - \nabla^2 f(x_k)\nu_k^N \right\| \\
&\leq \int_0^1 \|\nabla^2 f(x_k + t\nu_k^N) - \nabla^2 f(x_k)\|\|\nu_k^N\| dt \\
&\leq \int_0^1 Lt\|\nu_k^N\|^2 dt \\
&= \frac{L}{2}\|\nabla^2 f(x_k)^{-1}\nabla f(x_k)\|^2 \\
&\leq \underbrace{2L\|\nabla^2 f(x^*)^{-1}\|^2}_{=M}\|\nabla f(x_k)\|^2,
\end{aligned}$$

donde las últimas dos desigualdades siguen del hecho que $x_k, x_{k+1} \in B(x^*, \varepsilon)$ y por lo tanto $\nabla^2 f$ es L-Lipschitz en el segmento $[x_k, x_{k+1}]$ y $\|\nabla^2 f(x_k)^{-1}\| \leq 2\|\nabla^2 f(x^*)^{-1}\|$. Esto concluye la demostración. □

La conclusión del Teorema 3.21 es extremadamente positiva en términos de radio de convergencia. Sin embargo, sus hipótesis son muy restrictivas: es necesario que el punto inicial esté lo suficientemente cerca de un mínimo local estricto, que podría incluso no existir. Por lo tanto, este resultado hasta ahora solo tiene valor teórico: Nos dice que cuando podemos aplicar el método de Newton, la información que entrega el Hessiano acelera potentemente la convergencia del método de búsqueda lineal.

Universidad de O'Higgins

3.4.2. Método de Newton con modificación de la matriz Hessiana

Para poder usar el método de Newton desde cualquier punto inicial $x_0 \in \mathbb{R}^n$, es necesario "corregir", en cada iterando x_k, la matriz Hessiana $\nabla^2 f(x_k)$. La forma más simple de hacer esto (no la única) es sumar una ponderación los suficientemente grande de la matriz identidad I_n, de tal manera que la matriz modificada resultante sea lo suficientemente definida positiva. Concretamente, para cada punto x, queremos construir una matriz

$$B = \nabla^2 f(x) + \tau I_n,$$

donde $\tau > 0$ es un coeficiente lo suficientemente grande (que depende de x) de tal manera que

$$\lambda_{\text{mín}}(B) \geq \delta, \tag{3.18}$$

donde $\lambda_{\text{mín}}(B)$ es el menor valor propio de la matriz B y $\delta > 0$ es un parámetro de estabilidad numérica, definido por nosotros. Finalmente, la dirección de descenso en x se toma como $\nu = -B^{-1}\nabla f(x)$, en el mismo espíritu que (3.17).

La necesitad de tomar $\delta > 0$ como umbral mínimo para los valores propios de B es controlar la norma de la dirección ν. Esto es necesario pues implícitamente los métodos de Newton se basan en la apróximación de Taylor de segundo orden

$$f(x+\nu) \sim f(x) + \langle \nabla f(x), \nu \rangle + \frac{1}{2}\nu^T \nabla^2 f(x)\nu. \tag{3.19}$$

Esta aproximación es buena en la medida que ν no sea demasiado grande. El siguiente ejemplo ilustra que, si no imponemos la condición (3.18), entonces ν podría ser demasiado grande, violando la aproximación de segundo orden

Ejemplo 3.22 Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^2$ y $x \in \mathbb{R}^n$ tal que $\nabla f(x) = (1, -3, 2)^T$ y $\nabla^2 f(x) = \text{diag}(10, 3, -1)$. En este caso, la dirección de Newton está dada por

$$\nu^N = -\nabla^2 f(x)^{-1}\nabla f(x) = -\begin{pmatrix} -0.1 \\ 1 \\ 2 \end{pmatrix}.$$

Esta dirección no es una dirección de descenso para f en x, pues $\langle \nabla f(x), \nu \rangle = -0.1 - 3 + 4 = 0.9 > 0$. Tomemos ahora $\tau = 1 + 10^{-8}$. En este caso, tendremos que

$$B = \nabla^2 f(x) = \text{diag}(11 + 10^{-8}, 4 + 10^{-8}, 10^{-8}).$$

Luego, la dirección de descenso usando B en vez de la matriz Hessiana quedaría

$$\nu = -B^{-1}\nabla f(x) = \begin{pmatrix} \frac{-1}{11+10^{-8}} \\ \frac{3}{4+10^{-8}} \\ \frac{-2}{10^{-8}} \end{pmatrix} \approx -10^8 \nabla f(x).$$

La dirección ν por lo tanto tendría un tamaño del orden de 10^8, lo cual claramente es demasiado grande como para que (en general) la aproximación de Taylor (3.19) sea válida en $x + \nu$. Incluso si intentamos corregir la dirección anterior con un paso $\alpha \in (0, 1)$ que cumpla las condiciones de Wolfe, el factor 10^8 ya puede generar cierta inestabilidad numérica (dependiendo de la precisión de la máquina con la que estemos trabajando). □

Para evitar la situación patológica del Ejemplo 3.22, necesitamos asegurarnos que el coeficiente de modificación τ sea lo suficientemente grande. El siguiente lema nos entrega un valor mínimo para τ que garantiza el control sobre la dirección de descenso $\nu = -B\nabla f(x)$.

Universidad de O'Higgins

Lema 3.23. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^2$ y $x \in \mathbb{R}^n$ un punto regular de f. Sea $B = \nabla^2 f(x) + \tau I_n$ donde*

$$\tau \geq \max\{0, \delta - \lambda_{\min}(\nabla^2 f(x))\}. \tag{3.20}$$

Se tiene entonces que B es simétrica y definida positiva, y además $\lambda_{\min}(B) \geq \delta$. Más aún, tomando $\nu = -B^{-1}\nabla f(x)$ se tiene que $\|\nu\| \leq \delta^{-1}\|\nabla f(x)\|$.

Demostración. Como la matriz Hessiana $\nabla^2 f(x)$ es simétrica, entonces admite una descomposición de la forma

$$\nabla^2 f(x) = P\Lambda P^T,$$

donde P es una matriz ortonormal y λ es la matriz de valores propios de $\nabla^2 f(x)$. Luego, es fácil ver que

$$B = P(\Lambda + \tau I_n)P^T,$$

Concluimos entonces que B es una matriz simétrica y que

$$\begin{aligned}\lambda_{\min}(B) = \min_{i=1,\dots,n}(\lambda_i + \tau) &= \lambda_{\min}(\nabla^2 f(x)) + \tau \\ &\geq \lambda_{\min}(\nabla^2 f(x)) + \delta - \lambda_{\min}(\nabla^2 f(x)) = \delta.\end{aligned}$$

En particular, B es definida positiva. Finalmente, tenemos que

$$\|\nu\| = \|B^{-1}\nabla f(x)\| \leq \|P\| \underbrace{\|(\Lambda + \tau I_n)^{-1}\|}_{\leq \delta^{-1}} \|P^T\| \|\nabla f(x)\| = \delta^{-1}\|\nabla f(x)\|.$$

□

Basado en la proposición anterior, podemos establecer el siguiente algoritmo para encontrar $\tau > 0$ que satisfaga (3.20). Si bien podríamos buscar los valores propios de $\nabla^2 f(x)$ para definir τ, este procedimiento es costoso computacionalmente cuando las dimensiones de la matriz son muy grandes. En cambio, lo que podemos hacer es incrementar iterativamente τ e intentar la descomposición de Cholesky de la matriz $B = \nabla^2 f(x) + \tau I$, que actúa como certificado de que la matriz B es definida positiva: Una matriz admite descomposición de Cholesky $B = LL^T$ (donde L es una matriz triangular inferior con entradas estrictamente positivas en la diagonal) si y sólo si es simétrica definida positiva.

Algoritmo 3.4: Búsqueda Coeficiente de Modificación (IdentityModification)

Entrada: Matriz $A \in \mathcal{M}_{n\times n}(\mathbb{R})$ simétrica; Umbral $\delta > 0$ de positividad; $\beta > 0$.
Calcular $\bar{a} = \min\{a_{ii} : i = 1, \dots, n\}$ (menor elemento de la diagonal de A);
$\tau \leftarrow \max\{0, \beta - \bar{a}\}$:
while *TRUE* **do**
  Intentar descomposición de Cholesky $LL^T = A + \tau I_n$.
  if *Factorización exitosa y* $\lambda_{\min}(L) \geq \sqrt{\delta}$ **then**
    Return: L;
  end
  $\tau \leftarrow \max(2\tau, \beta)$;
end
Salida: Factor L de la descomposición de Cholesky $LL^T = A + \tau I_n$.

Cuando aplicamos el Algoritmo 3.4 con $A = \nabla^2 f(x)$, la dirección de descenso que obtenemos tomando $\nu = -(A + \tau I_n)^{-1}\nabla f(x)$ es en general (no siempre) una buena dirección de descenso. El problema con este algoritmo, es que la búsqueda puede tomar muchas iteraciones antes de producir un coeficiente τ aceptable. La decisión de duplicar τ en cada paso puede ser reemplazada por un

Universidad de O'Higgins

incremento mayor para tratar de acelerar el proceso. Por otro lado, los coeficientes $\delta > 0$ y $\beta > 0$ se deciden heurísticamente. Un valor usado normalmente es $\delta = \beta = 10^{-3}$.

Usando el algoritmo 3.4 como subrutina, podemos establecer un algoritmo que aplique el método de Newton, modificando la matriz Hessiana cada vez que sea necesario. Como no contamos en general con las garantías del Teorema 3.21, aplicamos también paso inexacto para al menos obtener garantías similares al Método de Máximo Descenso.

Algoritmo 3.5: Método de Newton con Modificación de Hessiano

Entrada: Función f; Gradiente ∇f; Hessiano $\nabla^2 f$; Tolerancia $\varepsilon > 0$; Umbral $\delta > 0$ de positividad; $\beta > 0$.
$x_{\text{curr}} \leftarrow x_0$; $x_{\text{next}} \leftarrow x_0$;
if $\|\nabla f(x_{\text{curr}})\| \leq \varepsilon$ **then**
  $x^* \leftarrow x_{\text{curr}}$;
  Return $(x^*, \nabla f(x^*))$;
end
for k = *1:MaxIter* **do**
  $L =$ `IdentityModification`$(\nabla^2 f(x_{\text{curr}}), \delta, \beta)$.
  Resolver $Lp = -\nabla f(x_{\text{curr}})$;
  Resolver $L^T \nu = p$; `// ν resuelve` $B\nu = -\nabla f(x_{\text{curr}})$
  $\alpha \leftarrow$ `MetodoPaso`$(f, x_{\text{curr}}, \nu)$; `// Puede ser exacto, o inexacto como LineSeach`
  $x_{\text{next}} \leftarrow x_{\text{curr}} + \alpha\nu$;
  if $\|x_{\text{next}} - x_{\text{curr}}\| \leq \varepsilon$ ***or*** $\|\nabla f(x_{\text{next}})\| \leq \varepsilon$ **then**
    $x^* \leftarrow x_{\text{next}}$;
    Break;
  end
  $x_{\text{curr}} \leftarrow x_{\text{next}}$;
end
Return $(x^*, \nabla f(x^*))$;
Salida: x^* punto crítico aproximado con $f(x^*) \leq f(x_0)$, y su gradiente $\nabla f(x^*)$.

El Algoritmo 3.5 produce una sucesión de iterandos $(x_k)_k$ siguiendo la forma general (3.3) de métodos de búsqueda lineal, donde para cada iteración $k \in \mathbb{N}$, la dirección de descenso es de la forma $\nu_k = -B_k^{-1}\nabla f(x_k)$ con

$$B_k = \nabla^2 f(x_k) + \tau_k I_n.$$

Decimos que una sucesión de matrices (B_k) satisface la propiedad de *factorización modificada acotada* si existe una constante $C > 0$ tal que

$$\forall k \in \mathbb{N}, \quad \|B_k\|\|B_k^{-1}\| \leq C. \tag{3.21}$$

donde $\|\cdot\|$ es la norma de operador lineal en $\mathcal{M}_{n\times n}(\mathbb{R})$, es decir, para $M \in \mathcal{M}_{n\times n}(\mathbb{R})$

$$\|M\| = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Mx\|}{\|x\|} = \sup_{x\in\mathbb{S}_n} \|Mx\|. \tag{3.22}$$

Cuando la sucesión de matrices (B_k) producida en el Algoritmo 3.5 satisface esta propiedad, podemos dar garantías de convergencia utilizando el Teorema de Zoutendijk 3.14. Para esto, necesitamos el siguiente lema de Álgebra lineal.

Lema 3.24. *Sea $B \in \mathcal{M}_{n\times n}(\mathbb{R})$ una matriz simétrica semidefinida positiva. Se tiene que B admite una única raíz cuadrada, es decir, existe una única matriz $A \in \mathcal{M}_{n\times n}(\mathbb{R})$ simétrica semidefinida positiva tal que $B = A^2$. Más aún, se tiene que*

Universidad de O'Higgins

(a) $\|B\| = \|A\|^2$.

(b) Si B es definida positiva, entonces A también es definida positiva.

Demostración. Como B es simétrica, admite una descomposición de la forma $B = P\Lambda P^T$, donde P es una matriz ortonormal, y $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, donde $\lambda_1, \ldots, \lambda_n$ son los valores propios de B. Como los valores propios de B son no-negativos pues B es semidefinida positiva, podemos definir la matriz A dada por

$$A = P\sqrt{\Lambda}P^T \tag{3.23}$$

donde $\sqrt{\Lambda} = \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$. No es difícil confirmar que con esta definición, $B = A^2$ y además que se verifica (b).

Para demostrar la unicidad de la raíz, razonemos por contradicción, y supongamos que existen dos matrices $A_1, A_2 \in \mathcal{M}_{n\times n}(\mathbb{R})$ con $A_1 \neq A_2$, ambas simétricas y semidefinidas positivas, con $A_1^2 = B = A_2^2$. Consideremos la matriz $C = A_1 - A_2$.

Sabemos que C es simétrica y que $C \neq 0$. Por lo tanto, todos los valores propios de C son reales y existe al menos un valor propio λ de C con $\lambda \neq 0$. Nuevamente por simetría de C, existe al menos un vector propio $x \neq 0$ asociado al valor propio λ. Luego, podemos escribir

$$\begin{aligned} 0 = \langle x, (A_1^2 - A_2^2)x\rangle &= \langle x, (A_1 + A_2)Cx\rangle \\ &= \langle (A_1 + A_2)x, \lambda x\rangle = \lambda(\langle A_1 x, x\rangle + \langle A_2 x, x\rangle). \end{aligned}$$

Como $\lambda \neq 0$, el desarrollo anterior implica que $\langle A_1 x, x\rangle + \langle A_2 x, x\rangle = 0$. Más aún, como A_1 y A_2 son semidefinidas positivas, esto sólo se tiene si $\langle A_1 x, x\rangle = \langle A_1 x, x\rangle = 0$. Luego, podemos escribir

$$0 = \langle A_1 x, x\rangle - \langle A_2 x, x\rangle = \langle Cx, x\rangle = \lambda\langle x, x\rangle \neq 0,$$

lo cual es una contradicción. Concluimos entonces que la raíz de B es única.

Solo nos queda demostrar (a). Primero, recordando que para todo par de matrices $M_1, M_2 \in \mathcal{M}_{n\times n}(\mathbb{R})$ siempre se tiene que $\|M_1 M_2\| \leq \|M_1\|\|M_2\|$, podemos escribir directamente que

$$\|B\| = \|A^2\| \leq \|A\|^2.$$

Para la otra desigualdad, recordando que A es simétrica, podemos escribir

$$\begin{aligned} \|A\|^2 &= \sup_{x\in\mathbb{S}_n} \|Ax\|^2 \\ &= \sup_{x\in\mathbb{S}_n} \langle Ax, Ax\rangle \\ &= \sup_{x\in\mathbb{S}_n} \langle x, AAx\rangle \\ &= \sup_{x\in\mathbb{S}_n} \langle x, Bx\rangle \\ &\leq \sup_{x\in\mathbb{S}_n} \|x\|\|Bx\| = \sup_{x\in\mathbb{S}_n} \|Bx\| = \|B\|. \end{aligned}$$

Esto concluye la demostración. □

Usando el lema anterior, podemos establecer el siguiente resultado de convergencia que nos sirve de garantía para el Algoritmo 3.5.

Teorema 3.25. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^2$ y sea $x_0 \in \mathbb{R}^n$. Sea (x_k) la sucesión de iterandos producida por el Algoritmo 3.5, es decir, tal que*

$$x_{k+1} = x_k - \alpha_k B_k^{-1}\nabla f(x_k), \quad \forall k \in \mathbb{N},$$

donde $B_k = \nabla^2 f(x_k) + \tau_k I_n$ es la matriz Hessiana modificada dada por el Algoritmo 3.4. Supongamos que

Universidad de O'Higgins

(i) f *es acotada inferiormente.*

(ii) ∇f *es Lipschitz-continua en un abierto* A *que contiene al subnivel* $[f \leq f(x_0)]$.

(iii) La sucesión de matrices (B_k) *satisface la propiedad de factorización modificada acotada* (3.21).

Entonces (x_k) *es globalmente convergente, es decir,* $\nabla f(x_k) \to 0$.

Demostración. Sin perder generalidad, supongamos que para todo $k \in \mathbb{N}$, el iterando x_k es un punto regular de f (de lo contrario, el resultado es directo). Dadas las condiciones (i) y (ii), y el hecho que la dirección $\nu_k = -B_k \nabla f(x_k)$ es dirección de descenso de f en x_k, tenemos que se cumplen todas las hipótesis del Teorema de Zoutendijk 3.14. Por lo tanto,

$$\sum_{k=1}^{\infty} \cos(\theta_k)^2 \|\nabla f(x_k)\|^2 < +\infty.$$

En particular, tenemos que

$$|\cos(\theta_k)| \|\nabla f(x_k)\| = \frac{|\langle \nabla f(x_k), \nu_k \rangle|}{\|\nabla f(x_k)\| \|\nu_k\|} \|\nabla f(x_k)\| = \frac{\langle \nabla f(x_k), B_k^{-1} \nabla f(x_k) \rangle}{\|B_k^{-1} \nabla f(x_k)\|} \to 0$$

Sea A_k la raíz cuadrada de B_k, dada por el lema 3.24. Usando que B_k es definida positiva y recordando la construcción de A_k dada por (3.23), es fácil verificar que A_k^{-1} es la raíz cuadrada de B_k^{-1}. Luego, tenemos que

$$\begin{aligned} \|\nabla f(x_k)\|^2 &= \|A_k A_k^{-1} \nabla f(x_k)\|^2 \\ &\leq \|A_k\|^2 \|A_k^{-1} \nabla f(x_k)\|^2 \\ &= \|B_k\| \langle A_k^{-1} \nabla f(x_k), A_k^{-1} \nabla f(x_k) \rangle = \|B_k\| \langle \nabla f(x_k), B_k^{-1} \nabla f(x_k) \rangle \end{aligned}$$

Finalmente, tomando $C > 0$ como la constante dada por (3.21) y recordando que $\|B_k^{-1}\| \|\nabla f(x_k)\| \geq \|B_k^{-1} \nabla f(x_k)\| > 0$, podemos escribir

$$\begin{aligned} \frac{1}{C} \|\nabla f(x_k)\| &\leq \frac{1}{\|B_k\| \|B_k^{-1}\|} \|\nabla f(x_k)\| \\ &= \frac{\|\nabla f(x_k)\|^2}{\|B_k\| \|B_k^{-1}\| \|\nabla f(x_k)\|} \\ &\leq \frac{\|\nabla f(x_k)\|^2}{\|B_k\| \|B_k^{-1} \nabla f(x_k)\|} \leq \frac{\langle \nabla f(x_k), B_k^{-1} \nabla f(x_k) \rangle}{\|B_k^{-1} \nabla f(x_k)\|} \to 0. \end{aligned}$$

Concluimos entonces que $\|\nabla f(x_k)\| \to 0$, finalizando la demostración. □

La demostración del Teorema 3.25 está basada principalmente en la propiedad de Factorización Modificada Acotada (3.21) que debe cumplirse para las matrices producidas por el Algoritmo 3.4. Sin embargo, no es claro si esta condición es exigente o no. La verdad, es que bajo condiciones razonables sobre la función objetivo f, es posible verificar que la propiedad (3.21) se cumplirá durante la ejecución del Algoritmo 3.5. Para esto, es necesario utilizar un resultado de teoría espectral que no estamos en condiciones de demostrar, y es que los valores propios de una matriz cuadrada dependen "continuamente"de las entradas de la matriz. De hecho, es posible demostrar que la función

$$\begin{aligned} \rho : \mathcal{M}_{n\times n}(\mathbb{R}) &\to \mathbb{R} \\ A &\mapsto \rho(A) := \text{máx}\{|\lambda_i| \ : \ \lambda_i \text{ valor propio de } A\} \end{aligned} \tag{3.24}$$

es continua. El valor $\rho(A)$ se conoce como el **radio espectral** de la matriz A.

Universidad de O'Higgins

Como ya lo hemos ocupado varias veces en este capítulo, cuando tenemos una matriz B simétrica definida positiva, podemos escribir $B = P\Lambda P^T$ con P una matriz ortonormal y Λ la matriz diagonal de valores propios de B. Como P es ortonormal, tenemos que $\|P\| = 1$. Luego, tenemos que

$$\|B\| = \|P\Lambda P^T\| \leq \|\Lambda\| \leq \rho(B).$$

En particular, tenemos que para toda matriz simétrica definida positiva,

$$\|B\|\|B^{-1}\| \leq \rho(B)\rho(B^{-1}).$$

Ahora, supongamos que $f : \mathbb{R}^n \to \mathbb{R}$ es una función de clase $\mathcal{C}^2$ y $x_0 \in \mathbb{R}^n$ un punto regular de f tal que el conjunto de subnivel $S = [f \leq f(x_0)]$ es compacto. Para cualquier punto $x \in S$, tendremos por continuidad de $x \mapsto \nabla^2 f(x)$, que

$$\rho(\nabla^2 f(x)) \leq M,$$

para alguna constante $M > 0$ lo suficientemente grande. Esta cota uniforme nos asegura que el coeficiente $\tau > 0$ producido por el Algoritmo 3.4 es también uniformemente acotado: el valor máximo que puede tomar τ está dado por el primer valor $\bar{\tau}$ que es más grande que $M + \delta$ (para δ el umbral de positividad), pues $\lambda_{\text{mín}}(\nabla^2 f(x)) \geq -M$. Este valor depende de cómo estemos aumentando τ en la línea 9 de cada iteración, pero es constante una vez fijamos este criterio y el valor de β. Luego, notando que el Algoritmo 3.4 asegura que todos los valores propios de $B(x) := \nabla^2 f(x) + \tau I_n$ son mayores que δ, tendremos que para todo $x \in S$

$$\|B(x)\|\|B(x)^{-1}\| \leq \rho(B(x))\rho(B(x)^{-1}) \leq \underbrace{(M + \bar{\tau})\frac{1}{\delta}}_{=C}.$$

Dado que las condiciones de compacidad de S y continuidad de la matriz hessiana $\nabla^2 f$ aseguran también que el gradiente ∇f es Lipschitz en un abierto que contiene a S, podemos establecer el siguiente corolario, que resume cuándo podemos asegurar convergencia global del Algoritmo 3.5.

Corolario 3.26. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^2$ y $x_0 \in \mathbb{R}^n$ un punto tal que el subnivel $[f \leq f(x_0)]$ es compacto. Entonces, la sucesión (x_k) producida por el Algoritmo 3.5 con punto inicial x_0 es globalmente convergente, es decir, $\|\nabla f(x_k)\| \to 0$.*

Si durante la ejecución del Algoritmo 3.5, la sucesión (x_k) pasa lo suficientemente cerca de un mínimo local estricto con Hessiano suficientemente definido positivo, entonces el Algoritmo 3.4 comenzará a entregar la matriz Hessiana sin modificación. Más aún, el paso de descenso se volverá eventualmente constante igual a 1, y por lo tanto estaremos situados en el contexto del Método de Newton puro, obteniendo el radio de convergencia cuadrático dado por el Teorema 3.21. Este deseable resultado dependerá de cuán rápido ocurra dicho encuentro, lo cual dependerá fuertemente de cómo es la función f y el punto inicial x_0. Incluso podría pasar que la sucesión (x_k) nunca cumpla esta condición y por lo tanto, en general, el radio de convergencia cuadrático es "cosa de suerte".

3.5 Métodos de Quasi-Newton

Similarmente al Método de Newton con modificación del Hessiano, los métodos de Quasi-Newton consideran, en cada iteración x_k, direcciones de descenso de la forma

$$\nu_k = -B_k^{-1}\nabla f(x_k),$$

donde B_k es una matriz simétrica y definida positiva. La diferencia fundamental con los Métodos de Newton es que la matriz B_k no se construye a partir del Hessiano $\nabla^2 f(x_k)$, si no que se utilizan

Universidad de O'Higgins

métodos indirectos para seleccionar "la mejor matriz B_k". Luego, las iteraciones siguen la forma estándar de métodos de búsqueda lineal, es decir, $x_{k+1} = x_k + \alpha_k \nu_k$, donde α_k es un paso que cumple las condiciones de Wolfe.

Para ver cómo se construye la matrices $\{B_k \ : \ k \in \mathbb{N}\}$, supongamos que ya hemos construido los iterando $x_0, \ldots, x_k, x_{k+1}$, y que queremos construir la matriz B_{k+1}. En el iterando x_{k+1}, los métodos de Quasi-Newton usan la información tanto del punto x_{k+1} como la del punto anterior x_k para construir la matriz B_{k+1}. La idea es que B_{k+1} emule el Hessiano de f, capturando la evolución del gradiente ∇f desde x_k hasta x_{k+1}.

Suponiendo que conocemos B_{k+1} y motivados por la expansión de Taylor, se define la función $m_{k+1} : \mathbb{R}^n \to \mathbb{R}$ dada por

$$m_{k+1}(\nu) = f(x_{k+1}) + \langle \nabla f(x_{k+1}), \nu \rangle + \frac{1}{2} \nu^T B_{k+1} \nu. \tag{3.25}$$

El Teorema de Taylor aplicado al gradiente ∇f nos dice que

$$\nabla f(x_{k+1}) - \nabla f(x_k) \approx \nabla^2 f(x_{k+1)}(x_{k+1} - x_k).$$

Si B_{k+1} emula el Hessiano desde x_k a x_{k+1}, entonces m_{k+1} debiera ser una buena aproximación de segundo orden de f en el segmento $[x_{k+1}, x_k]$. Por lo tanto, nos gustaría que m_{k+1} cumpliera que

1. $\nabla m_{k+1}(0) = \nabla f(x_{k+1})$ (esto siempre se tiene).
2. $\nabla m_{k+1}(x_k - x_{k+1}) = \nabla f(x_k)$. Esta segunda condición nos dice que si avanzamos desde x_{k+1} a x_k, entonces el gradiente de m_{k+1} pasa de ser $\nabla f(x_{k+1})$ a $\nabla f(x_k)$.

El segundo requerimiento sobre m_{k+1} nos dice que

$$\nabla f(x_k) = \nabla m_{k+1}(x_k - x_{k+1}) = \nabla f(x_{k+1}) + B_{k+1}(x_k - x_{k+1}),$$

lo que se traduce en la ecuación

$$B_{k+1} \underbrace{(x_{k+1} - x_k)}_{=s_k} = \underbrace{\nabla f(x_{k+1}) - \nabla f(x_k)}_{y_k}. \tag{3.26}$$

Esta ecuación se conoce como **ecuación de la secante**, y se puede interpretar de la siguiente manera: La matriz B_{k+1} no tiene que ser el Hessiano $\nabla^2 f(x_{k+1})$, pero si debe aproximarlo en la dirección $x_{k+1} - x_k$, es decir,

$$B_{k+1} s_k \approx \nabla^2 f(x_{k+1}) s_k.$$

Suponiendo que existen matrices que satisfagan la ecuación de la secante, priori habrían muchas formas de seleccionar una matriz B_{k+1}. Diferentes formas de seleccionar esta matriz inducen diferentes algoritmos de búsqueda lineal de Quasi-Newton. En general, todo método de Quasi-Newton se puede expresar por el Algoritmo 3.6.

Para ver la correctitud del algoritmo 3.6, necesitamos asegurar en cada iteración $k+1$, la existencia de matrices definidas positivas que satisfagan la ecuación de la secante en los puntos x_k y x_{k+1}. Si existiera una matriz B_{k+1} definida positiva que cumple la ecuación de la secante, entonces se debe cumplir también que

$$0 < s_k^T B_{k+1} s_k = \langle s_k, y_k \rangle.$$

Por lo tanto, para que exista B_{k+1} con las propiedades deseadas, es condición necesaria que

$$0 < \langle s_k, y_k \rangle. \tag{3.27}$$

Esta condición necesaria no depende de B_{k+1} y se conoce como la **condición de curvatura de Quasi-Newton**. El siguiente lema nos muestra que para que exista al menos una matriz B_{k+1} con las propiedades deseadas, la condición de curvatura de Quasi-Newton es también suficiente.

Universidad de O'Higgins

Algoritmo 3.6: Método general de Quasi-Newton

Entrada: Función f; Gradiente ∇f; Punto inicial x_0; Matriz inicial B_0 simétrica definida positiva; Tolerancia $\varepsilon > 0$.
$x_{\text{curr}} \leftarrow x_0$;
if $\|\nabla f(x_{\text{curr}})\| \leq \varepsilon$ **then**
  $x^* \leftarrow x_{\text{curr}}$;
  Return $(x^*, \nabla f(x^*))$;
end
$\nu \leftarrow -B_0^{-1}\nabla f(x_{\text{curr}})$;
$\alpha \leftarrow$ `MetodoPaso`$(f, x_{\text{curr}}, \nu)$; `// Puede ser exacto, o inexacto como LineSeach`
$x_{\text{next}} \leftarrow x_0 + \alpha\nu$;
for k = *1:MaxIter* **do**
  Seleccionar B satisfaciendo $B(x_{\text{next}} - x_{\text{curr}}) = \nabla f(x_{\text{next}}) - \nabla f(x_{\text{curr}})$;
  $x_{\text{curr}} \leftarrow x_{\text{next}}$;
  $\nu \leftarrow -B^{-1}\nabla f(x_{\text{curr}})$;
  $\alpha \leftarrow$ `MetodoPaso`$(f, x_{\text{curr}}, \nu)$; `// Puede ser exacto, o inexacto como LineSeach`
  $x_{\text{next}} \leftarrow x_{\text{curr}} + \alpha\nu$;
  if $\|x_{\text{next}} - x_{\text{curr}}\| \leq \varepsilon$ ***or*** $\|\nabla f(x_{\text{next}})\| \leq \varepsilon$ **then**
    $x^* \leftarrow x_{\text{next}}$;
    Break;
  end
end
Return $(x^*, \nabla f(x^*))$;
Salida: x^* punto crítico aproximado con $f(x^*) \leq f(x_0)$, y su gradiente $\nabla f(x^*)$.

Lema 3.27. *Sean $s_k, y_k \in \mathbb{R}^n$ dos vectores tal que $\langle s_k, y_k\rangle > 0$. Entonces existe al menos una matriz B simétrica y definida positiva tal que $Bs_k = y_k$.*

Demostración. Consideremos la matriz

$$B = \left(I_n - \frac{1}{\langle s_k, y_k\rangle} y_k s_k^T\right)\left(I_n - \frac{1}{\langle s_k, y_k\rangle} s_k y_k^T\right) + \frac{1}{\langle s_k, y_k\rangle} y_k y_k^T. \tag{3.28}$$

Si definimos $A = I - \frac{1}{\langle s_k, y_k\rangle} s_k y_k^T$, entonces tenemos que $B = A^T A + \frac{1}{\langle s_k, y_k\rangle} y_k y_k^T$, lo cual implica inmediatamente que B es simétrica y semidefinida positiva.

Para ver que B es de hecho definida positiva, razonemos por contradicción, y supongamos que no lo es. Entonces, dado que si es semidefinida positiva, existe $x \neq 0$ tal que

$$x^T B x = x^T A^T A x + x^T \left(\frac{1}{\langle s_k, y_k\rangle} y_k y_k^T\right) x = 0.$$

Esto implica que cada uno de los sumandos tiene que ser cero (pues $A^T A$ e $y_k y_k^T$ son semidefinidas positivas). Por lo tanto, $Ax = 0$ y $\langle y_k, x\rangle = y_k^T x = 0$. La segunda igualdad implica que x es ortogonal a y_k. Luego, podemos escribir

$$Ax = x - \frac{\langle y_k, x\rangle}{\langle s_k, y_k\rangle} s_k = x.$$

Finalmente, concluimos que $Ax = 0 \implies x = 0$, lo cual es una contradicción. Hemos demostrado

Universidad de O'Higgins

que B es definida positiva. Finalmente,

$$\begin{aligned} Bs_k &= \left(I - \frac{1}{\langle s_k, y_k\rangle} y_k s_k^T\right)\left(I - \frac{1}{\langle s_k, y_k\rangle} s_k y_k^T\right) s_k + \frac{1}{\langle s_k, y_k\rangle} y_k y_k^T s_k \\ &= \left(I - \frac{1}{\langle s_k, y_k\rangle} y_k s_k^T\right)\left(s_k - \frac{1}{\langle s_k, y_k\rangle} s_k y_k^T s_k\right) + y_k \\ &= \left(I - \frac{1}{\langle s_k, y_k\rangle} y_k s_k^T\right)(s_k - s_k) + y_k = y_k. \end{aligned}$$

Por lo tanto, B satisface la ecuación de la secante (3.26). □

Finalmente, la siguiente proposición nos da un método para garantizar que la condición de curvatura de Quasi-Newton se mantenga en cada iteración $k+1$: necesitamos que el paso α_k cumpla las condiciones de Wolfe (ver Definición 3.9).

Proposición 3.28. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^2$ y sean x_k, x_{k+1} dos iterandos consecutivos del Algoritmo 3.6 con $x_{k+1} \neq x_k$. Si el paso α_k cumple las condición de curvatura (3.6) para $c_2 \in (0,1)$, entonces x_k y x_{k+1} satisfacen la condición de curvatura de Quasi-Newton*

$$\langle x_{k+1} - x_k, \nabla f(x_{k+1}) - \nabla f(x_k)\rangle > 0.$$

Demostración. Sea $c_2 \in (0,1)$. Recordemos que como el Algoritmo 3.6 produjo el iterando x_{k+1}, entonces $x_{k+1} = x_k + \alpha_k \nu_k$, $\nabla f(x_k) \neq 0$ y $\nu_k = -B_k^{-1}\nabla f(x_k)$, donde B_k es una matriz simétrica semidefinida positiva. De la condición (3.6), sabemos que

$$\langle \nabla f(x_{k+1}), \nu_k\rangle \geq c_2 \langle \nabla f(x_k), \nu_k\rangle.$$

Por lo tanto, podemos escribir

$$\begin{aligned} \langle \nabla f(x_{k+1}) - \nabla f(x_k), x_{k+1} - x_k\rangle &= \alpha_k \langle \nabla f(x_{k+1}) - \nabla f(x_k), \nu_k\rangle \\ &\geq \alpha_k (c_2 - 1)\langle \nabla f(x_k), \nu_k\rangle \\ &= \alpha_k (c_2 - 1)(\nabla f(x_k)^T B_k^{-1}(-\nabla f(x_k))) \\ &= \alpha_k (1 - c_2)(\nabla f(x_k)^T B_k^{-1}\nabla f(x_k)) > 0. \end{aligned}$$

□

Con esta última proposición, podemos estar seguros que los métodos de Quasi-Newton son globalmente convergentes, en la medida que el paso que elijamos siempre cumpla las condiciones de Wolfe. En la práctica, no siempre es posible garantizar la condición de curvatura para el paso α_k, pues el agloritmo de LineSearch 3.2 puede devolver $\alpha_k = \alpha_{\text{máx}}$, satisfaciendo solo la condición de Armijo (3.5). Para evitar problemas, las implementaciones de métodos de Quasi-Newton realizan iteraciones de máximo descenso cada vez que α_k no cumple la condición de curvatura, es decir:

1. Se toma $B_0 = I_n$ y se fija $\alpha_{\text{máx}} = 1$.
2. Si el paso α en la línea 14 de 3.2 no cumple la condición de curvatura (3.6), entonces se reemplaza $B = I_n$ y $\nu = -\nabla f(x_{\text{curr}})$ y se recalcula α.

Esta subrutina de seguridad se puede interpretar de la siguiente manera: Cada vez que α no cumple la condición de curvatura (3.6), el algoritmo 3.6 se reinicia con $x_0 = x_{\text{curr}}$.

Universidad de O'Higgins

3.5.1. Método de Davidon-Fletcher-Powell (DFP)

Para elegir B_{k+1}, el método de Davidon-Fletcher-Powell (DFP) consiste en buscar la matriz más cercana a la matriz de la iteración anterior B_k que sea simétrica y que cumpla la ecuación (3.26). Para esto, se resuelve el siguiente problema de optimización:

$$\begin{cases} \min\limits_{B\in\mathcal{M}_{n\times n}(\mathbb{R})} & \|B-B_k\|_W \\ s.a. & \begin{cases} B=B^T \\ Bs_k=y_k, \end{cases} \end{cases} \tag{3.29}$$

donde $W\in\mathcal{M}_{n\times n}(\mathbb{R})$ es una matriz simétrica definida positiva que satisface que $Ws_k=y_k$, donde

$$\|A\|_W=\|W^{1/2}AW^{1/2}\|_F. \tag{3.30}$$

y donde $\|\cdot\|_F$ denota la norma de Frobenius para matrices, es decir, para $M\in\mathcal{M}_{n\times n}(\mathbb{R})$,

$$\|M\|_F=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}M_{ij}^2}. \tag{3.31}$$

El siguiente teorema, nos muestra que, independiente de la matriz de pesos W que elijamos (que siempre existe gracias al Lema 3.27 y la proposición 3.28), la solución del problema (3.29) es única.

Teorema 3.29. *Sea B_k es simétrica definida positiva y $s_k,y_k\in\mathbb{R}^n$ con $\langle s_k,y_k\rangle>0$. Entonces para cualquier matriz W simétrica definida positiva satisfaciendo que $Ws_k=y_k$, el problema (3.29) tiene solución única, y está dada por*

$$B_{k+1}=\left(I-\frac{1}{\langle s_k,y_k\rangle}y_ks_k^T\right)B_k\left(I-\frac{1}{\langle s_k,y_k\rangle}s_ky_k^T\right)+\frac{1}{\langle s_k,y_k\rangle}y_ky_k^T. \tag{3.32}$$

Más aún, B_{k+1} es definida positiva.

la gran ventaja de la fórmula (3.32), conocida como la **fórmula DFP**, es que la inversa de B_{k+1} tiene expresión analítica, por lo que podemos evitar resolver el sistema $B_{k+1}\nu_{k+1}=-\nabla f(x_{k+1})$ y calcular directamente ν_{k+1} como $-B_{k+1}^{-1}\nabla f(x_{k+1})$.

Proposición 3.30. *Sea B_k una simétrica definida positiva, $s_k,y_k\in\mathbb{R}^n$ satisfaciendo la condición de curvatura de Quasi-Newton $\langle s_k,y_k\rangle>0$, y B_{k+1} la matriz dada por (3.32). Entonces, denotando $H_k=B_k^{-1}$ y $H_{k+1}=B_{k+1}^{-1}$, se tiene que*

$$H_{k+1}=H_k-\frac{H_ky_ky_k^TH_k}{y_k^TH_ky_k}+\frac{1}{\langle y_k,s_k\rangle}s_ks_k^T. \tag{3.33}$$

Ocupando esta última proposición, el método de Quasi-Newton que usa la fórmula DFP no selecciona B_{k+1}, si no que actualiza directamente la inversa H_{k+1}. Es decir, denotando por $\texttt{DFP}(H,s,y)$ la función que entrega H_{k+1} usando (3.33) con $H_k=H$, $s_k=s$ e $y_k=y$, el Algoritmo 3.6 se adapta de la siguiente manera:

1. Al inicio, antes de entrar al `loop-for`, el algoritmo recibe como entrada H_0 (en vez de B_0), inicializa la matriz $H\leftarrow H_0$ y calcula $\nu\leftarrow -H\nabla f(x_{\text{curr}})$.
2. En cada iteración del `loop-for`, se actualiza la matriz H usando la fórmula (3.33), es decir,

$$H\leftarrow\texttt{DFP}(H,x_{\text{next}}-x_{\text{curr}},\nabla f(x_{\text{next}})-\nabla f(x_{\text{curr}})).$$

Universidad de O'Higgins

3. Finalmente se actualiza $x_{\text{curr}} \leftarrow x_{\text{next}}$ y se calcula $\nu \leftarrow -H\nabla f(x_{\text{curr}})$. El resto del Algoritmo 3.6 se mantiene.

Cabe destacar que esta modificación requiere no solo guardar registro de la iteración anterior x_k (que se hace a través de las dos variables x_{curr} y x_{next}), si no que se debe mantener registro de la matriz H_k usada en la iteración anterior. Esto se logra con la nueva variable H, que se actualiza en base a su valor anterior.

3.5.2. Método de Broyden-Fletcher-Glodfarb-Shanno (BFGS)

Dado que estamos interesados en $H_{k+1} = B_{k+1}^{-1}$, podríamos olvidarnos de actualizar B_{k+1} a partir de B_k y concentrarnos en buscar H_{k+1} a partir directamente de H_k. El método de Broyden-Fletcher-Glodfarb-Shanno (BFGS) consiste precisamente en buscar la matriz más cercana a la matriz de la iteración anterior H_k que sea simétrica y que cuya inversa cumpla la ecuación (3.26), es decir, que cumpla que

$$H_{k+1}y_k = s_k. \tag{3.34}$$

Para esto, se resuelve el siguiente problema de optimización:

$$\begin{cases} \min\limits_{H \in \mathcal{M}_{n\times n}(\mathbb{R})} & \|H - H_k\|_W \\ s.a. & \begin{cases} H = H^T \\ s_k = Hy_k, \end{cases} \end{cases} \tag{3.35}$$

donde nuevamente $W \in \mathcal{M}_{n\times n}(\mathbb{R})$ es una matriz simétrica definida positiva que satisface que $Ws_k = y_k$, y donde $\|\cdot\|_W$ es la norma dada por (3.30). Al igual que en el método DFP, se puede demostrar que, independiente de la matriz de pesos W que elijamos, la solución del problema (3.35) es única.

Teorema 3.31. *Sea H_k es simétrica definida positiva y $s_k, y_k \in \mathbb{R}^n$ con $\langle s_k, y_k\rangle > 0$. Entonces para cualquier matriz W simétrica definida positiva satisfaciendo que $Ws_k = y_k$, el problema (3.35) tiene solución única, y está dada por*

$$H_{k+1} = \left(I - \frac{1}{\langle s_k, y_k\rangle} s_k y_k^T\right) H_k \left(I - \frac{1}{\langle s_k, y_k\rangle} y_k s_k^T\right) + \frac{1}{\langle s_k, y_k\rangle} s_k s_k^T. \tag{3.36}$$

Más aún, H_{k+1} es definida positiva.

El método de Quasi-Newton que utiliza la fórmula (3.36), conocida como **fórmula BFGS**, es reconocido en la práctica como el mejor método de Quasi-Newton, hasta el día de hoy. Las implementaciones profesionales de programas de optimización no-lineal como `Matlab` utilizan esta fórmula (o variaciones de ella) cómo método base de actualización. El método de Quasi-Newton usando la fórmula BFGS está dado por el Algoritmo 3.7, donde `BFGS`(H, s, y) es la fórmula BFGS (3.36) con $H_k = H$, $s_k = s$ e $y_k = y$.

3.5.3. Radio de convergencia para métodos de Quasi-Newton

En la sección 3.3 vimos que el Método de Máximo Descenso tiene, en general, radio de convergencia lineal, lo cual puede llegar a ser demasiado lento en aplicaciones. Por otro lado, en la sección 3.4 vimos que bajo condiciones ideales, el método de Newton tiene radio de convergencia cuadrático.

El método de Quasi-Newton es un punto entre medio del método de Máximo Descenso (lento pero seguro) y el método de Newton (rápido en condiciones ideales, pero que en muchos casos falla). Y

Universidad de O'Higgins

Algoritmo 3.7: Método de Quasi-Newton con actualización BFGS

Entrada: Función f; Gradiente ∇f; Punto inicial x_0; Matriz inicial H_0 simétrica definida positiva (por defecto, $H_0 = I_n$); Tolerancia $\varepsilon > 0$.
$x_{\text{curr}} \leftarrow x_0$;
if $\|\nabla f(x_{\text{curr}})\| \leq \varepsilon$ **then**
  $x^* \leftarrow x_{\text{curr}}$;
  Return $(x^*, \nabla f(x^*))$;
end
$H \leftarrow H_0$;
$\nu \leftarrow -H\nabla f(x_{\text{curr}})$;
$\alpha \leftarrow$ `MetodoPaso`$(f, x_{\text{curr}}, \nu)$;
$x_{\text{next}} \leftarrow x_0 + \alpha\nu$;
for k = *1:MaxIter* **do**
  $H \leftarrow$ `BFGS`$(H, x_{\text{next}} - x_{\text{curr}}, \nabla f(x_{\text{next}}) - \nabla f(x_{\text{curr}}))$; `// Actualizar` H `usando` (3.36)
  $x_{\text{curr}} \leftarrow x_{\text{next}}$;
  $\nu \leftarrow -H\nabla f(x_{\text{curr}})$;
  $\alpha \leftarrow$ `MetodoPaso`$(f, x_{\text{curr}}, \nu)$;
  if α *no satisface la condición de curvatura* (3.6) **then**
    Reiniciar el algoritmo con $x_0 = x_{\text{curr}}$;
  end
  $x_{\text{next}} \leftarrow x_{\text{curr}} + \alpha\nu$;
  if $\|x_{\text{next}} - x_{\text{curr}}\| \leq \varepsilon$ ***or*** $\|\nabla f(x_{\text{next}})\| \leq \varepsilon$ **then**
    $x^* \leftarrow x_{\text{next}}$;
    Break;
  end
end
Return $(x^*, \nabla f(x^*))$;
Salida: x^* punto crítico aproximado con $f(x^*) \leq f(x_0)$, y su gradiente $\nabla f(x^*)$.

como es de esperar, su radio de convergencia también está entre medio: en general, alcanza un radio de convergencia **superlineal**.

Para garantizar el radio de convergencia superlineal en los métodos de Quasi-Newton, es necesario que la búsqueda del paso α_k en cada iteración comience con $\alpha_{\text{máx}} = 1$ (ver algoritmos 3.1 y 3.2). Cuando se implementa de esta manera, los métodos de Quasi-Newton comienzan a usar $\alpha_k = 1$ a partir de cierto punto, cuando se encuentran lo suficientemente cerca de un mínimo local estricto. El siguiente teorema nos da condiciones sobre las cuales se alcanza el redio de convergencia superlineal

Teorema 3.32. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función de clase $\mathcal{C}^2$ y $x_0 \in \mathbb{R}^n$. Sea $(x_k)_{k\in\mathbb{N}}$ producida por un método de Quasi-Newton (ver algoritmo 3.6), donde para todo $k \in \mathbb{N}$,*

(i) La dirección de descenso en la iteración $k \in \mathbb{N}$ está dada por $\nu_k = -B_k^{-1}\nabla f(x_k)$ con B_k matriz simétrica definida positiva.

(ii) α_k satisface las condiciones de Wolfe.

(iii) Existe $k_0 \in \mathbb{N}$ lo suficientemente grande tal que $\alpha_k = 1$ para todo $k \geq k_0$.

Supongamos que la sucesión (x_k) converge a un punto $x^* \in \mathbb{R}^n$ crítico de f, con $\nabla^2 f(x^*)$ definida positiva y con $\nabla^2 f$ Lipschitz-continua cerca de x^*. Si la secuencia de matrices (B_k) cumple que

$$\lim_{k\to\infty} \frac{\|(B_k - \nabla^2 f(x^*))\nu_k\|}{\|\nu_k\|} = 0, \tag{3.37}$$

Universidad de O'Higgins

entonces (x_k) converge a x^* con radio de convergencia superlineal.

Demostración. Como $x^* \in \mathbb{R}^n$ es un punto crítico de f con $\nabla^2 f(x^*)$ definida positiva, entonces existen $\delta > 0$ y $M > 0$ tal que $\nabla^2 f(x)$ es definida positiva y $\|\nabla^2 f(x)^{-1}\| \leq M$ para todo $x \in B(x^*, \delta)$, y además $\nabla^2 f$ es Lischitz-continua en $B(x^*, \delta)$.

Sin perder generalidad, supongamos que $\alpha_k = 1$ para todo $k \in \mathbb{N}$ y que $x_0 \in B(x^*, \delta)$ (de lo contrario, basta tomar $k_0 \in \mathbb{N}$ lo suficientemente grande tal que $x_{k_0} \in \mathbb{N}$ y que verifica la condición (iii), y reemplazar x_0 por x_{k_0}).

Sea $\nu_k^N = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$, la dirección de Newton en el punto x_k. Tenemos que

$$\begin{aligned} \|\nu_k - \nu_k^N\| &= \|\nabla^2 f(x_k)^{-1}(\nabla^2 f(x_k)\nu_k + \nabla f(x_k))\| \\ &= \|\nabla^2 f(x_k)^{-1}(\nabla^2 f(x_k)\nu_k - B_k\nu_k)\| \\ &\leq M\|(B_k - \nabla^2 f(x_k))\nu_k\|, \end{aligned}$$

donde $\|\nabla^2 f(x_k)^{-1}\| \leq M$ pues $x_k \in B(x^*, \delta)$. Como $\nabla^2 f(x_k) \to \nabla^2 f(x^*)$, es fácil ver que la ecuación (**??**) implica que

$$\frac{\|(B_k - \nabla^2 f(x_k))\nu_k\|}{\|\nu_k\|} \xrightarrow{k\to\infty} 0.$$

Ocupando los argumentos de la demostración del Teorema 3.21, tenemos que para todo $x \in B(x, \delta)$, se cumple que

$$\|(x - \nabla^2 f(x)^{-1}\nabla f(x)) - x^*\| \leq K\|x - x^*\|^2,$$

para alguna constante $K > 0$ lo suficientemente grande. Más aún, la continuidad del Hessiano $\nabla^2 f$ nos asegura que ∇f es Lipschitz-continua en $B(x^*, \delta)$. Por lo tanto, como $\nabla f(x^*) = 0$, tenemos que

$$\begin{aligned} \|x_{k+1} - x^*\| &= \|x_k - \nu_k - x^*\| \\ &\leq \|(x_k - \nu_k^N) - x^*\| + \|\nu_k - \nu_k^N\| \\ &\leq K\|x_k - x^*\|^2 + M\|(B_k - \nabla^2 f(x_k))\nu_k\| \\ &= K\|x_k - x^*\|^2 + \underbrace{\left(M\frac{\|(B_k - \nabla^2 f(x_k))\nu_k\|}{\|\nu_k\|}\right)}_{a_k}\|\nu_k\| \end{aligned}$$

Como $a_k \to 0$, existe $k_0 \in \mathbb{N}$ lo suficientemente grande tal que $a_k < 1$ para todo $k \geq k_0$. Luego, basados en el desarrollo anterior y notando que $\|x_k - \nu_k - x^*\| \geq \|\nu_k\| - \|x_k - x^*\|$, podemos escribir que

$$\|\nu_k\| \leq \frac{K\|x_k - x^*\| + 1}{1 - a_k}\|x_k - x^*\|, \quad \forall k \geq k_0.$$

Finalmente, concluimos que para todo $k \geq k_0$

$$\begin{aligned} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} &\leq \frac{K\|x_k - x^*\|^2 + a_k\|\nu_k\|}{\|x_k - x^*\|} \\ &\leq \frac{K\|x_k - x^*\|^2 + a_k\frac{K\|x_k - x^*\|+1}{1-a_k}}{\|x_k - x^*\|} \\ &= K\|x_k - x^*\| + a_k\frac{K}{1 - a_k} \xrightarrow{k\to\infty} 0. \end{aligned}$$

Por lo tanto, (x_k) converge a x^* con radio superlineal. □

Universidad de O'Higgins

3.6 Ejercicios Capítulo 3

P1. (*Paso inexacto truncado*) Sean $f : \mathbb{R}^n \to \mathbb{R}$ y $x_0 \in \mathbb{R}^n$ satisfaciendo las hipótesis del Teorema 3.14. Sean además $c_1 \in (0,1)$, $c_2 \in (c_1,1)$ y $\bar{\alpha} > 0$, y sean tres sucesiones $(x_k)_k \subset \mathbb{R}^n$, $(\nu_k)_k \subset \mathbb{R}^n \setminus \{0\}$ y $(\alpha_k)_k \subset (0,+\infty)$ tal que

(i) Para todo $k \in \mathbb{N}$, $x_{k+1} = x_k + \alpha_k \nu_k$.

(ii) Para todo $k \in \mathbb{N}$, ν_k es dirección de descenso de f en x_k y $\|\nu_k\| = 1$.

(iii) Para todo $k \in \mathbb{N}$, se tiene que: 1) o bien $\alpha_k = \alpha_0$ si α_0 satisface la condición de Armijo para f en x_k y con dirección ν_k, 2) o bien $\alpha_k \in (0, \alpha_0)$ y satisface las condiciones de Wolfe para f en x_k con dirección ν_k.

El objetivo de este problema, es mostrar que en este caso también se verifica la condición de Zoutendijk (3.8). Para esto, proceda como sigue:

a) Defina los conjuntos $A = \{k \in \mathbb{N} \ : \ \alpha_k = \alpha_0\}$ y $B = \{k \in \mathbb{N} \ : \ \alpha_k < \alpha_0\}$. Muestre que existen $a, b > 0$ tal que para todo $k \in \mathbb{N}$ se tiene que

$$f(x_{k+1}) \leq f(x_0) - a \sum_{j\{0,\dots,k\}\cap A} \cos(\theta_k)\|\nabla f(x_j)\| - b \sum_{j\in\{0,\dots,k\}\cap B} \cos(\theta_j)^2 \|\nabla f(x_j)\|^2$$

b) Muestre que si $|A| = \infty$, entonces existe $k_0 \in \mathbb{N}$ tal que

$$\forall j \in A \cap \{k \ : \ k \geq k_0\}, \quad \|\nabla f(x_j)\| < 1.$$

Concluya entonces que, tanto si $|A| = \infty$ como si $|A| < \infty$, siempre existen $k_0 \in \mathbb{N}$, $c > 0$ y $M > 0$ tal que

$$\forall k \geq k_0, \quad f(x_{k+1}) \leq f(x_0) + M - c \sum_{j=k_0}^{k} \cos(\theta_j)^2 \|\nabla f(x_j)\|^2.$$

c) Concluya entonces que en este caso también se tiene la condición de Zoutendijk, es decir,

$$\sum_{j=0}^{\infty} \cos(\theta_j)^2 \|\nabla f(x_j)\|^2 < +\infty.$$

P2. Implemente en `Matlab` los Algoritmos 3.2 y 3.3.

P3. Implemente en `Matlab` los Algoritmos 3.4 y 3.5.

Universidad de O'Higgins

Capítulo 4

Optimización con Restricciones

En esta sección estudiaremos el problema de optimización con restricciones que presentamos en el Capítulo 1, es decir,

$$(\mathcal{P}) := \begin{cases} \min\limits_{x\in\mathbb{R}^n} & f(x) \\ s.a. & x \in X, \end{cases} \tag{4.1}$$

donde X será es el conjunto factible, determinado por restricciones de igualdad y desigualdad. Es decir, X tiene la forma

$$X = \left\{ x \in \mathbb{R}^n \ : \ \begin{array}{l} h_i(x) = 0, \forall i \in I \\ g_j(x) \leq 0, \forall j \in J \end{array} \right\}, \tag{4.2}$$

donde I y J son dos conjuntos de índices finitos. Para seguir con la línea que estudiamos en el caso de optimización sin restricciones, vamos a asumir que los datos del problema 4.1 son al menos de clase $\mathcal{C}^1$, es decir, que la función objetivo $f : \mathbb{R}^n \to \mathbb{R}$ y las funciones de restricción $h_i : \mathbb{R}^n \to \mathbb{R}$ (para todo $i \in I$) y $g_j : \mathbb{R}^n \to \mathbb{R}$ (para todo $j \in J$) son continuamente diferenciables. Recordemos que en este contexto, un punto $x^* \in \mathbb{R}^n$ se dice

- **mínimo global** del problema (4.1) si $x^* \in X$ y
$$\forall x \in X, \quad f(x^*) \leq f(x).$$
- **mínimo local** del problema (4.1) si $x^* \in X$ y existe $\delta > 0$ tal que
$$\forall x \in X \cap B(x^*, \delta), \quad f(x^*) \leq f(x).$$

Por supuesto, el problema (4.1) es un problema más difícil (en general) que el problema de optimización sin restricciones (3.1) estudiado en el Capítulo 3. Por lo mismo, es natural pensar que los algoritmos que podamos desarrollar para este problema apunten, a partir de un punto inicial $x_0 \in X$, solo a encontrar un punto $x^* \in X$ que sea "punto crítico" y tal que $f(x^*) \leq f(x_0)$. Con suerte, este punto crítico será un mínimo local e incluso en algunos casos (como problemas convexos), será un mínimo global.

Pero, ¿Qué significa punto crítico en el contexto de optimización con restricciones? Evidentemente, queremos que todo óptimo local sea un punto crítico. Esto requerimiento nos dice que la condición $\nabla f(x) = 0$ no es una buena definición de punto crítico, como se muestra en el Ejemplo 4.1.

Ejemplo 4.1 Consideremos el problema de optimización

$$\begin{cases} \min\limits_{x\in\mathbb{R}^2} & f(x) = (x_2 + 100)^2 + \frac{1}{100}x_1^2 \\ s.a. & \cos(x_1) - x_2 \leq 0. \end{cases}$$

Universidad de O'Higgins

Si calculamos el gradiente y el Hessiano de la función objetivo, tendremos que

$$\nabla f(x) = \begin{bmatrix} \frac{1}{50}x_1 \\ 2(x_2 + 100) \end{bmatrix} \qquad \text{y} \qquad \nabla^2 f(x) = \begin{bmatrix} \frac{1}{50} & 0 \\ 0 & 2 \end{bmatrix}.$$

Esto nos dice que la función es convexa, y que su único punto crítico (que a la vez es el mínimo global de f) está en $x^* = (0, -100)$. Sin embargo, este punto no está en el conjunto factible pues

$$\cos(x_1^*) - x_2^* = \cos(0) + 100 = 101 > 0.$$

Por otro lado, f tiene muchos mínimos locales cerca de los puntos de la forma $x = (k\pi, -1)$ con $k \in \mathbb{Z}$ impar, como muestra la Figura 4.1.

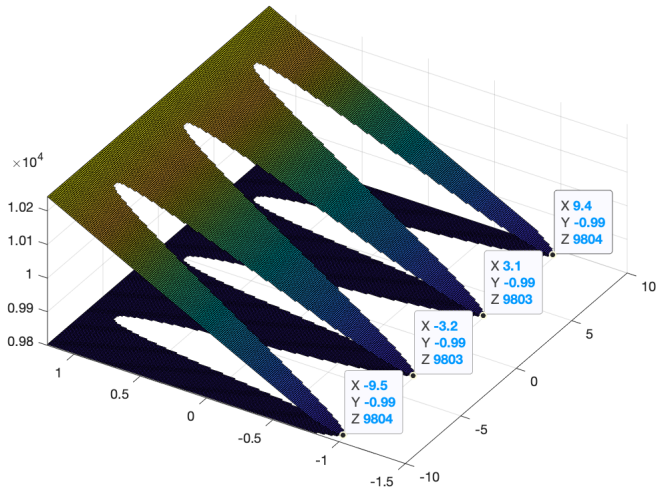


Figura 4.1: Función $f(x) = (x_2 + 100)^2 + \frac{1}{100}x_1^2$ (en degradé), cuyos valores en el eje Z están graficados sobre en dominio $\{x \ : \ \cos(x_1) - x_2 \leq 0\}$, El dominio está graficado como sombra en azul oscuro. Los puntos marcados son mínimos locales (aproximados).

□

En este capítulo estudiaremos cómo definir puntos críticos para problemas de optimización con restricciones. Derivaremos condiciones de optimalidad a partir de estas definiciones, conocidas como las condiciones de Karush-Kuhn-Tucker. Finalmente, estudiaremos el algoritmo más utilizado para problemas de optimización no-lineal con restricciones: El algoritmo de punto interior.

4.1 Condiciones de optimalidad

Antes de desarrollar la teoría de condiciones de optimalidad para problemas de optimización no-lineales, analizaremos dos ejemplos, de los cuales intentaremos deducir qué cosas son necesarias para definir el concepto de punto crítico.

Universidad de O'Higgins

Ejemplo 4.2 Partamos primero por el problema

$$\begin{cases} \min\limits_{x\in\mathbb{R}^2} & x_1 + x_2 \\ s.a. & x_1^2 + x_2^2 - 2 = 0, \end{cases} \tag{4.3}$$

que es un problema con solo una restricción de igualdad dada por la función $h(x) = x_1^2 + x_2^2 - 2$. Es fácil verificar que el único mínimo local de este problema se alcanza en $x^* = (-1, -1)$.

Intuitivamente, si $x \in X$ no es un mínimo local, deberíamos ser capaces de encontrar $y \in \mathbb{R}^2$ tal que

1. y está cerca de x;
2. $f(y) < f(x)$; y
3. $h(y) = 0$.

Tomando $s = y - x$, que debe ser un vector "pequeño", podríamos aplicar el Teorema de Taylor, deduciendo que

$$f(y) \sim f(x) + \langle \nabla f(x), s\rangle \quad \text{y} \quad h(y) \sim h(x) + \langle \nabla h(x), s\rangle.$$

Ahora, tomando $\nu = \|s\|^{-1}s$, tendremos que

$$\begin{cases} 0 > \frac{f(y)-f(x)}{\|s\|} & \sim \langle \nabla f(x), \nu\rangle \\ 0 = \frac{h(y)-h(x)}{\|s\|} & \sim \langle \nabla h(x), \nu\rangle. \end{cases}$$

Concluimos entonces que si x no es un mínimo local del problema (4.3), deberíamos poder construir un vector unitario ν tal que $\langle \nabla f(x), \nu\rangle < 0$ y $\langle \nabla h(x), \nu\rangle = 0$. Sin embargo, esto es imposible cuando los vectores $\nabla f(x)$ y $\nabla h(x)$ son **paralelos**, es decir, cuando

$$\exists \lambda \in \mathbb{R} \quad \text{tal que } \nabla f(x) + \lambda \nabla h(x) = 0. \tag{4.4}$$

En este caso, podríamos decir que x es punto crítico del problema (4.3) si existe $\lambda \in \mathbb{R}$ que verifica la ecuación (4.4). En nuestro ejemplo, esto ocurre en dos puntos: $x = (-1, -1)$ y $x = (1, 1)$. □

Ejemplo 4.3 Consideremos ahora el problema

$$\begin{cases} \min\limits_{x\in\mathbb{R}^2} & x_1 + x_2 \\ s.a. & x_1^2 + x_2^2 - 2 \leq 0, \end{cases} \tag{4.5}$$

que es un problema con solo una restricción de desigualdad dada por la función $g(x) = x_1^2 + x_2^2 - 2$. Nuevamente, el único mínimo local de este problema se alcanza en $x^* = (-1, -1)$. Ahora, si un punto $x \in X$ no es mínimo local, deberíamos poder encontrar $y \in \mathbb{R}^2$ tal que

1. y está cerca de x;
2. $f(y) < f(x)$; y
3. $g(y) \leq 0$.

Nuevamente, definiendo $s = y - x$ y usando el teorema de Taylor, podemos escribir

$$f(y) \sim f(x) + \langle \nabla f(x), s\rangle \quad \text{y} \quad g(y) \sim g(x) + \langle \nabla g(x), s\rangle.$$

Aquí, distinguimos dos casos:

1. **Caso $g(x) < 0$:** En este caso, para todo s lo suficientemente pequeño, digamos con $s \in B(0, \delta)$ para algún $\delta > 0$, tendríamos que
$$g(x + s) < 0.$$

Universidad de O'Higgins

Por lo tanto, la condición $g(y) \leq 0$ se cumple trivialmente si y está lo suficientemente cerca de x. Luego, la única condición que nos queda es verificar que $f(y) < f(x)$, lo cual implicaría que existe un vector unitario ν tal que

$$\langle \nabla f(x), \nu \rangle < 0.$$

La existencia de dicho vector falla sólo cuando $\nabla f(x) = 0$ (de lo contrario, podríamos tomar $\nu = -\nabla f(x)/\|\nabla f(x)\|$). Así, podríamos definir que x es punto crítico del problema (4.5) cuando $g(x) < 0$ y $\nabla f(x) = 0$.

2. **Caso $g(x) = 0$:** En este caso, tomando $\nu = \|s\|^{-1}s$ y recordando que $g(y) - g(x) = g(y) \leq 0$, tenemos que el desarrollo de Taylor nos dice que

$$\begin{cases} 0 > \frac{f(y)-f(x)}{\|s\|} & \sim \langle \nabla f(x), \nu \rangle \\ 0 \geq \frac{g(y)-g(x)}{\|s\|} & \sim \langle \nabla g(x), \nu \rangle. \end{cases}$$

Es decir, si x no es un mínimo local y $g(x) = 0$, entonces deberíamos poder encontrar un vector unitario ν tal que $\langle \nabla f(x), \nu \rangle < 0$ y $\langle \nabla g(x), \nu \rangle \leq 0$. Para que esto falle, necesitamos que $\nabla g(x)$ sea **paralelo** a $\nabla f(x)$ y además que ambos vectores **apunten en sentidos opuestos**. Es decir, necesitamos que

$$\exists \mu \geq 0, \quad \text{tal que } \nabla f(x) + \mu \nabla g(x) = 0, \tag{4.6}$$

o sea, que $\nabla g(x)$ sea una homotecia de $-\nabla f(x)$. En este caso, podríamos decir que x es punto crítico del problema (4.5) cuando $g(x) = 0$ y existe $\mu \geq 0$ que verifica la ecuación (4.6).

¿Cómo juntar ambos casos? Bueno, lo que podemos pedir para decir que x es un punto crítico es que verifique que

$$\exists \mu \geq 0, \quad \text{tal que } \begin{cases} \nabla f(x) + \mu \nabla g(x) = 0, \\ \mu \cdot g(x) = 0. \end{cases} \tag{4.7}$$

La condición $\mu \cdot g(x) = 0$ se conoce como **holgura complementaria**, y nos permite incluir ambos casos: Si $g(x) = 0$, entonces μ queda libre y por lo tanto tenemos que μ debe resolver la ecuación (4.6). Por el contrario, si $g(x) < 0$, entonces necesariamente $\mu = 0$, y por lo tanto, la única forma de verificar (4.7) es que $\nabla f(x) = 0$. Entonces, podemos decir que x es punto crítico del problema (4.5), si existe $\mu \geq 0$ que verifica (4.7). En nuestro ejemplo, esto ocurre solo en el punto $x = (-1, -1)$. □

Basado en los dos ejemplos anteriores, podríamos intentar atacar el caso general, es decir, el problema

$$\begin{cases} \min\limits_{x \in \mathbb{R}^n} & f(x) \\ s.a. & \begin{cases} h_i(x) = 0, \, \forall i \in I \\ g_j(x) \leq 0, \, \forall j \in J \end{cases} \end{cases} \tag{4.8}$$

Siguiendo la misma lógica de los ejemplos, si un punto factible x no es mínimo local, deberíamos poder encontrar un vector unitario ν tal que

$$\begin{cases} \langle \nabla f(x), \nu \rangle < 0, \\ \langle \nabla h_i(x), \nu \rangle = 0 & \forall i \in I, \\ \langle \nabla g_j(x), \nu \rangle \leq 0 & \forall j \in J \text{ tal que } g_j(x) = 0. \end{cases} \tag{4.9}$$

El conjunto $\{j \in J \, : \, g_j(x) = 0\}$ recibe el nombre de **conjunto de restricciones activas** de x, y lo denotamos $\mathcal{A}(x)$. Los Ejemplos 4.2 y 4.3 sugieren que la condición (4.9) falla cuando $-\nabla f(x)$ es combinación lineal de los vectores $\nabla h_i(x)$ (con $i \in I$) y $\nabla g_j(x)$ (con $j \in \mathcal{A}(x)$), con la salvedad de

Universidad de O'Higgins

que los ponderadores de los vectores $\nabla g_j(x)$ deben ser positivos. En efecto, supongamos que existen multiplicadores $(\lambda_i \ : \ i \in I) \in \mathbb{R}^{|I|}$ y $(\mu_j \ : \ j \in \mathcal{A}(x)) \in \mathbb{R}_+^{|\mathcal{A}(x)|}$ tal que

$$\nabla f(x) + \sum_{i \in I} \lambda_i \nabla h_i(x) + \sum_{j \in \mathcal{A}(x)} \mu_j \nabla g_j(x) = 0,$$

y que además existe $\nu \in \mathbb{S}_n$ que verifica (4.9). Entonces, podríamos escribir,

$$0 < \langle -\nabla f(x), \nu \rangle = \sum_{i \in I} \underbrace{\langle \lambda_i \nabla h_i(x), \nu \rangle}_{=0} + \sum_{j \, : \, g_j(x) = 0} \underbrace{\langle \mu_j \nabla g_j(x), \nu \rangle}_{\leq 0} \leq 0,$$

lo cual es una contradicción. Esto motiva la Definición 4.4 de punto crítico.

Definición 4.4 (Punto crítico)**.** *Un punto $x \in \mathbb{R}^n$ se dice punto crítico del problema (4.8) si existen multiplicadores $(\lambda_i \ : \ i \in I)$ y $(\mu_j \ : \ j \in J)$ tal que*

$$\begin{cases} \nabla f(x) + \sum_{i \in I} \lambda_i h_i(x) + \sum_{j \in J} \mu_j g_j(x) = 0, \\ \mu_j g_j(x) = 0, \quad \forall j \in J, \\ \mu_j \geq 0, \quad \forall j \in J, \\ h_i(x) = 0, \quad \forall i \in I, \\ g_j(x) \leq 0, \quad \forall j \in J. \end{cases} \tag{4.10}$$

Las condiciones (4.10) se conocen como **condiciones de Karush-Kuhn-Tucker** y son reconocidas como la mejor forma de definir puntos críticos para problemas de optimización no-lineal con restricciones. En este set de condiciones, la primera ecuación nos dice que $-\nabla f(x)$ debe ser una combinación lineal de los gradientes de las restricciones. La segunda ecuación es la ecuación de holgura complementaria, que nos permite descartar las restricciones de desigualdad inactivas, forzando al multiplicador μ_j correspondiente a ser cero. La tercera ecuación es la positividad del multiplicador para las restricciones de desigualdad: como vimos en el Ejemplo 4.3, para los gradientes de las restricciones de desigualdad activas, es necesario que $-\nabla f(x)$ apunte "en la misma dirección". Finalmente, las últimas dos restricciones nos dicen que el punto x es un punto factible del problema. En lo que sigue de esta sección, intentaremos responder dos preguntas:

1. ¿Cuál es la interpretación geométrica de los puntos críticos?
2. ¿Bajo que condiciones podemos asegurar que todo mínimo local es punto crítico?

Por un lado, si bien la noción de punto crítico la derivamos de un estudio razonable sobre las aproximaciones de Taylor, esta deducción tiene una interpretación geométrica: Un punto crítico debe ser aquel desde el cual la función debe ser no-decreciente a través de toda dirección que apunte "hacia dentro"del conjunto factible. Por otro lado, en este punto aún no sabemos si los mínimos locales efectivamente verifican (4.10): En efecto, asumimos que negación de (4.9) implican las condiciones (4.10). Sin embargo, como veremos más adelante, dicha relación no siempre se tiene: Hay problemas donde con mínimos locales que no son puntos críticos, en el sentido de (4.10).

4.1.1. Condiciones geométricas de primer orden

Antes de comenzar nuestro análisis, es necesario recordar lo que es un cono en $\mathbb{R}^n$.

Definición 4.5. *Sea $C \subset \mathbb{R}^n$. Decimos que C es un **cono** si*

$$\forall y \in C, \forall \lambda \geq 0, \lambda y \in C. \tag{4.11}$$

*Si C es además cerrado, decimos que es un **cono cerrado**, y si C es además convexo, decimos que es un **cono convexo**.*

Universidad de O'Higgins

[Incluir Dibujo]

En optimización sin restricciones, los puntos críticos son aquellos puntos $x \in \mathbb{R}^n$ donde ninguna dirección $\nu \in \mathbb{R}^n$ es dirección de descenso para la función objetivo f. Esto sólo ocurre cuando $\nabla f(x) = 0$ precisamente pues f debe ser no-decreciente en cualquier dirección.

Sin embargo, si queremos replicar este mismo criterio en el contexto de optimización con restricciones, es claro que hay ciertas direcciones que no nos interesan: aquellas que, desde el punto x en cuestión, "se salen" inmediatamente del conjunto factible X. Esto quiere decir que la condición necesaria de primer orden debería ser que f sea no-decreciente a través de "ciertas" direcciones. Esto motiva la siguiente definición.

Definición 4.6 (Vector tangente y cono tangente)**.** *Sea X un conjunto no-vacío de $\mathbb{R}^n$ y sea $x \in X$. Decimos que un vector $\nu \in \mathbb{R}^n$ es un* ***vector tangente a X en x*** *si existen dos sucesiones $(\nu_k) \subset \mathbb{R}^n \setminus \{0\}$ y $(t_k) \subset (0, +\infty)$ tales que*

1. $\nu_k \to \nu$;
2. $t_k \to 0$; *y*
3. $x + t_k \nu_k \in X$ *para todo* $k \in \mathbb{N}$.

El conjunto de todos los vectores tangentes a X en x se denomina ***cono tangente de X en x*** *y se denota por $T_X(x)$ o bien por $T(x; X)$.*

Si tomamos un vector tangente ν a X en x distinto de cero, la Definición 4.6 nos dice que $d = \nu/\|\nu\|$ es un vector que podemos formar como un límite del tipo $d = \lim d_k$ donde

$$d_k = \frac{x_k - x}{\|x_k - x\|}$$

para (x_k) alguna sucesión en X convergente a x. Esto es precisamente la idea de que d como dirección se puede aproximar por direcciones que no se salen de X cerca del punto x.

En la misma línea, el nombre "cono tangente" que le damos al conjunto $T_X(x)$ no es antojadizo, si no que da cuenta de la estructura geométrica de este conjunto. La Figura 4.2 muestra varios ejemplos de conos tangentes en diferentes puntos de un mismo conjunto.

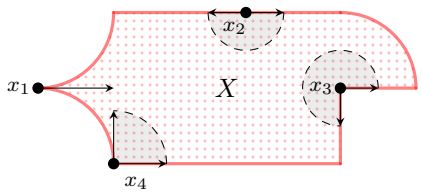


Figura 4.2: Ejemplo de conos tangentes (zona gris, que se proyecta a infinito entre las flechas negras). a) $T_X(x_1)$ es un rayo; b) $T_X(x_2)$ es un semiespacio, o sea, un cono de apertura π; c) $T_X(x_3)$ es un cono de apertura $3\pi/2$. d) $T_X(x_4)$ es un cono de apertura $\pi/2$.

Proposición 4.7. *Sea $X \subset \mathbb{R}^n$ y $x \in X$. Tenemos que*

1. *$T_X(x)$ es un cono cerrado no-vacío ($0 \in T_X(x)$).*
2. *$T_X(x) = T_{\overline{X}}(x)$.*
3. *Si $x \in \text{int}(X)$, entonces $T_X(x) = \mathbb{R}^n$.*
4. *El cono tangente es un concepto local, en el sentido que*

$$T(x; X) = T(x; X \cap B(x, \delta)), \quad \forall \delta > 0.$$

Universidad de O'Higgins

Demostración. Mostraremos cada afirmación por separado.

1. Claramente $0 \in T_X(x)$ (basta tomar $\nu_k = 0$ en la definición). Veamos ahora que $T_X(x)$ es un cono. En efecto, sea $\nu \in T_X(x)$ y $\lambda \geq 0$. Queremos demostrar que $\lambda\nu \in T_X(x)$. Claramente, si $\lambda = 0$ entonces $\lambda\nu = 0$ y el resultado se tiene. Supondremos entonces que $\lambda > 0$.

   Por definición, ν es vector tangente y por lo tanto existen $\nu_k \to \nu$ y $t_k \to 0$ (con $t_k > 0$), tal que $x + t_k\nu_k \in X$. Definiendo $s_k = t_k/\lambda$ y $u_k = \lambda\nu_k$, tenemos que $s_k \to 0$, que $u_k \to \lambda\nu_k$ y que
   $$x + s_k u_k = x + \frac{t_k}{\lambda}(\lambda\nu_k) = x + t_k\nu_k \in X, \quad \forall k \in \mathbb{N}.$$
   Por lo tanto, $\lambda\nu \in T_X(x)$. Esto demuestra que $T_X(x)$ es un cono.

   Veamos ahora que $T_X(x)$ es cerrado. Sea (ν_k) una sucesión en $T_X(x)$ convergente a $\nu \in \mathbb{R}^n$. Queremos demostrar que $\nu \in T_X(x)$. Por definición de $T_X(x)$, sabemos que para cada $k \in \mathbb{N}$, existe una sucesión $(\nu_{k,j})_j \subset \mathbb{R}^n$ y una sucesión $(t_{k,j})_j \subset (0, +\infty)$ tal que $\nu_{k,j} \to \nu_k$, $t_{k,j} \to 0$ y $x + t_{k,j}\nu_{k,j} \in X$ para todo $j \in \mathbb{N}$. Ahora, para cada $k \in \mathbb{N}$, podemos elegir $j_k \in \mathbb{N}$ lo suficientemente grande tal que
   $$\|\nu_k - \nu_{k,j_k}\| \leq \frac{1}{k} \quad \text{y} \quad t_{k,j_k} \leq \frac{1}{k}.$$
   Luego, si consideramos las sucesiones $(\nu_{k,j_k})_k$ y $(t_{k,j_k})_k$ seleccionadas de esta manera, tendremos que

   - $x + t_{k,j_k}\nu_{k,j_k} \in X$, para todo $k \in \mathbb{N}$;
   - $0 < t_{k,j_k} \leq \frac{1}{k} \xrightarrow{k\to\infty} 0$, por lo que $t_{k,j_k} \to 0$; y
   - $\nu_{k,j_k} \to \nu$. En efecto,
     $$\begin{aligned}\|\nu_{k,j_k} - \nu\| &= \|\nu_{k,j_k} - \nu_k + \nu_k - \nu\| \\ &\leq \|\nu_{k,j_k} - \nu_k\| + \|\nu_k - \nu\| \\ &\leq \frac{1}{k} + \|\nu_k - \nu\| \xrightarrow{k\to\infty} 0.\end{aligned}$$

   Concluimos entonces que $\nu \in T_X(x)$, como queríamos probar. Esto demuestra que $T_X(x)$ es también cerrado.

2. Claramente de la definición, $T_X(x) \subset T_{\overline{X}}(x)$ puesto que $X \subset \overline{X}$. Veamos ahora la inclusión reversa. Sea $\nu \in T_{\overline{X}}(x)$. Por definición, existe una sucesión $(\nu_k)_k$ convergiendo a ν y una sucesión $(t_k)_k \subset (0, +\infty)$ convergiendo a 0 tales que $x + t_k\nu_k \in \overline{X}$ para todo $k \in \mathbb{N}$. De la definición de adherencia, para cada $k \in \mathbb{N}$, podemos encontrar $u_k \in B(0, 1/k)$ tal que
   $$x + t_k(\nu_k + u_k) \in X.$$
   Luego, notando que $\nu_k + u_k \to \nu$ pues $u_k \to 0$, concluimos que que $\nu \in T_X(x)$, lo que muestra la inclusión reversa.

3. Supongamos que $x \in \text{int}(X)$ y sea $\nu \in \mathbb{R}^n$. Sabemos que existe $\lambda > 0$ lo suficientemente pequeño tal que el intervalo $[x, x + \lambda\nu] \subset X$. Luego, tomando $t_k = \lambda/k$ y $\nu_k = \nu$, concluimos que $\nu \in T_X(x)$. Como ν es arbitrario, tenemos que $T_X(x) = \mathbb{R}^n$.

4. Sea $\delta > 0$ fijo, y denotemos $Y = X \cap B(x, \delta)$. Como $Y \subset X$, tenemos directamente que $T_Y(x) \subset T_X(x)$. Por ende, demostraremos solo la inclusión reversa. Sea $\nu \in T_X(x)$. Por definición, existe una sucesión $(\nu_k)_k$ convergiendo a ν y una sucesión $(t_k)_k \subset (0, +\infty)$ convergiendo a 0 tales que $x + t_k\nu_k \in X$ para todo $k \in \mathbb{N}$. Notemos que
   $$\|t_k\nu_k\| = t_k\|\nu_k\| \leq t_k(\|\nu_k - \nu\| + \|\nu\|) \xrightarrow{k\to\infty} 0.$$

Universidad de O'Higgins

Esto nos dice que, definiendo $x_k = x + t_k\nu_k$, tenemos que la sucesión (x_k) converge a x. Esto quiere decir que existe $k_0 \in \mathbb{N}$ lo suficientemente grande tal que para todo $k \geq k_0$, $x_k \in B(x, \delta)$. Por lo tanto, $(x_k)_{k\geq k_0} \subset Y$, lo que implica que $\nu \in T_Y(x)$ al considerar las sucesiones $(\nu_k)_{k\geq k_0}$ y $(t_k)_{k\geq k_0}$. Esto concluye la demostración.

□

Así como tenemos la noción de vectores tangentes al conjunto X, que son aquellos vectores que apuntan "hacia dentro de X" (o límite de vectores que apuntan hacia adentro de X), también podemos introducir la noción de **vectores normales**, que apuntan "en dirección opuesta al conjunto X".

Definición 4.8 (Vector normal y Cono normal)**.** *Sea X un conjunto no-vacío y $x \in X$. Decimos que un vector $\xi \in \mathbb{R}^n$ es un **vector normal de X en x** si*

$$\langle \xi, \nu \rangle \leq 0, \quad \forall \nu \in T_X(x). \tag{4.12}$$

*El conjunto de todos los vectores normales a X en x se denomina **cono normal de X en x** y se denota por $N_X(x)$ o bien por $N(x; X)$.*

La Figura 4.3 ilustra como se construyen los conos normales a partir de los conos tangentes para diferentes puntos de un conjunto X.

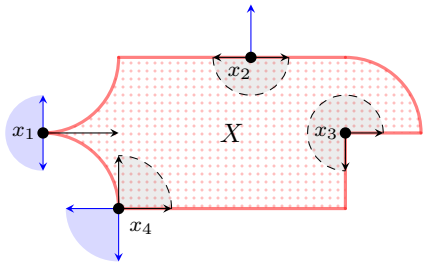


Figura 4.3: Ejemplo de conos normales (zona azul). a) $N_X(x_1)$ es el semiespacio inducido por la única dirección tangente; b) $N_X(x_2)$ es el único rayo ortogonal opuesto al semiespacio dado por $T_X(x_2)$; c) $N_X(x_3)$ es sólo el $\{0\}$. d) $N_X(x_4)$ es un cono de apertura $\pi/2$, el opuesto a $T_X(x_4)$.

Proposición 4.9. *Sea X un conjunto no-vacío y $x \in X$. El cono normal $N_X(x)$ es un cono convexo cerrado no-vacío ($0 \in N_X(x)$).*

Demostración. Claramente $0 \in N_X(x)$. Más aún, para todo $\xi \in N_X(x)$ y todo $\lambda > 0$ se tiene que

$$\langle \lambda\xi, \nu \rangle = \lambda\langle \xi, \nu \rangle \leq 0, \qquad \forall \nu \in T_X(x).$$

Por lo tanto, $\lambda\xi \in N_X(x)$, lo que muestra que $N_X(x)$ es un cono.

Para ver que es convexo, consideremos $\xi_1, \xi_2 \in N_X(x)$ y $t \in [0, 1]$. Tenemos que para todo $\nu \in T_X(x)$,

$$\langle t\xi_1 + (1-t)\xi_2, \nu \rangle = t\langle \xi_1, \nu \rangle + (1-t)\langle \xi_2, \nu \rangle \leq 0,$$

y por lo tanto $t\xi_1 + (1-t)\xi_2 \in N_X(x)$. Concluimos entonces que $N_X(x)$ es convexo.

Finalmente, para ver que es cerrado, tomemos una sucesión $(\xi_k) \subset N_X(x)$ con $\xi_k \to \xi \in \mathbb{R}^n$. Tenemos que para todo $\nu \in T_X(x)$

$$\langle \xi, \nu \rangle = \lim_k \langle \xi_k, \nu \rangle \leq 0,$$

y por lo tanto $\xi \in N_X(x)$. Concluimos entonces que $N_X(x)$ es también cerrado, lo que termina la demostración. □

Universidad de O'Higgins

Con estas dos nociones (conos tangente y normal) estamos listos para establecer las condiciones necesarias optimalidad de primer orden en el contexto de optimización sin restricciones.

Proposición 4.10. *Sea $x^* \in X$ un mínimo local del problema (4.1). Entonces se tiene que*

$$\forall \nu \in T_X(x^*), \quad \langle -\nabla f(x^*), \nu \rangle \leq 0, \tag{4.13}$$

o equivalentemente, $-\nabla f(x^*) \in N_X(x^*)$.

Demostración. Sea $\nu \in T_X(x)$. Sin perder generalidad, supongamos que $\nu \neq 0$. Por definición, existen dos sucesiones $(\nu_k) \subset \mathbb{R}^n$ y $(t_k) \subset (0, +\infty)$ tal que $\nu_k \to \nu$, $t_k \to 0$ y $x^* + t_k \nu_k \in X$ para todo $k \in \mathbb{N}$. Definamos $x_k = x^* + t_k \nu_k$ y tomemos $d_k = (x_k - x^*)/\|x_k - x^*\|$. Como $\nu \neq 0$, entonces (d_k) está bien definida, al menos a partir de algún $k_0 \in \mathbb{N}$ lo suficientemente grande. Más aún,

$$d_k = \frac{t_k \nu_k}{\|t_k \nu_k\|} = \frac{\nu_k}{\|\nu_k\|} \to \frac{\nu}{\|\nu\|}.$$

Luego, como x^* es mínimo local de (4.1), tenemos que $f(x_k) - f(x^*) \geq 0$ para $k \in \mathbb{N}$ lo suficientemente grande. Así, podemos escribir

$$\begin{aligned}
\langle -\nabla f(x^*), \nu/\|\nu\| \rangle &= \lim_k \langle -\nabla f(x^*), d_k \rangle \\
&= \lim_k \frac{f(x_k) - f(x^*)}{\|x_k - x^*\|} - \langle \nabla f(x^*), d_k \rangle + \underbrace{\frac{f(x^*) - f(x_k)}{\|x_k - x^*\|}}_{\leq 0} \\
&\leq \lim_k \frac{f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle}{\|x_k - x^*\|} = 0.
\end{aligned}$$

Concluimos entonces que $-\nabla f(x^*) \in N_X(x^*)$, lo que concluye la demostración. □

Lo que nos dice la proposición anterior es que cuando $x^* \in X$ es un mínimo local, entonces todas las direcciones $\nu \in \mathbb{R}^n$ que son direcciones de descenso de f en x^* se salen del conjunto factible X. El argumento es simple: si ν es dirección de descenso, entonces $\langle -\nabla f(x^*), \nu \rangle > 0$. Por lo tanto, $\nu \notin T_X(x^*)$, y eso significa que la dirección ν "se sale del conjunto". Dicho de otro modo, $-\nabla f(x^*)$ separa las direcciones factibles (es decir los vectores tangentes) de las direcciones de descenso. Esta idea se ilustra en la Figura 4.4.

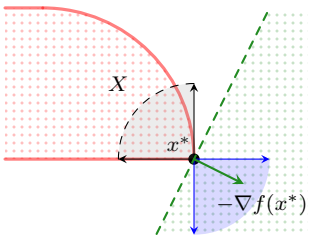


Figura 4.4: Ilustración de condición necesaria de primer orden geométrica: $-\nabla f(x^*)$ pertenece al cono normal $N_X(x^*)$ (en azul). Las posibles direcciones de descenso desde el punto x^* están el el hiperplano inducido por $-\nabla f(x^*)$ (en verde) separado del conjunto factible X (en rojo).

4.1.2. Teorema de Karush-Kuhn-Tucker

Volvamos ahora al caso donde X está dado por un conjunto de restricciones, de la forma (4.2). Para cada punto $x \in X$, lo que queremos ahora es relacionar de alguna manera el cono tangente $T_X(x)$ con las condiciones (4.9) con las que obtuvimos la definición de punto crítico.

Universidad de O'Higgins

Definición 4.11. *Sea X un conjunto no-vacío, de la forma (4.2) y sea $x \in X$. Definimos el* ***cono tangente linearizado de X en x*** *como*

$$\mathscr{F}_X(x) := \left\{ \nu \in \mathbb{R}^n \ : \ \begin{array}{ll} \langle \nabla h_i(x), \nu \rangle = 0, & \forall i \in I, \\ \langle \nabla g_j(x), \nu \rangle \leq 0, & \forall j \in \mathcal{A}(x), \end{array} \right\} \tag{4.14}$$

donde $\mathcal{A}(x)$ es el conjunto de restricciones activas de x (los $j \in J$ tal que $g_j(x) = 0$).

El cono $\mathscr{F}_X(x)$ se llama cono tangente linearizado pues es el cono tangente que se obtiene de reemplazar las restricciones h_i y g_j por sus aproximaciones de Taylor de primer orden en torno a x. Es decir, $\mathscr{F}_X(x) = T_Z(x)$, donde Z es el polítopo dado por

$$Z = \left\{ z \in \mathbb{R}^n \ : \ \begin{array}{ll} h_i(x) + \langle \nabla h_i(x), z - x \rangle = 0 & \forall i \in I \\ g_j(x) + \langle \nabla g_j(x), z - x \rangle \leq 0 & \forall j \in J \end{array} \right\}.$$

A priori, el polítopo Z es bastante distinto del conjunto original X. No obstante, el Teorema de Taylor nos hace pensar que Z y X debieran parecerse cerca del punto x, y por lo tanto uno podría esperar que $T_X(x)$ se pareciera a $\mathscr{F}_X(x)$.

Lema 4.12 (de Farkas)**.** *Sea X un conjunto no-vacío de la forma (4.2) y sea $x \in X$. Si $T_X(x) = \mathscr{F}_X(x)$, entonces*

$$N_X(x) = \left\{ \sum_{i \in I} \lambda_i \nabla h_i(x) + \sum_{j \in \mathcal{A}(x)} \mu_j \nabla g_j(x) \ \middle| \ \begin{array}{ll} \lambda_i \in \mathbb{R} & \forall i \in I \\ \mu_j \geq 0 & \forall j \in \mathcal{A}(x) \end{array} \right\}. \tag{4.15}$$

Demostración. Denotemos por K el conjunto de la derecha de (4.15). Primero, veamos que $K \subset N_X(x)$. En efecto, sea $\xi \in K$. Entonces, existen $(\lambda_i \ : \ i \in I)$ y $(\mu_j \ : \ j \in \mathcal{A}(x))$ tal que $\mu_j \geq 0$ para todo $j \in \mathcal{A}(x)$ y

$$\xi = \sum_{i \in I} \lambda_i \nabla h_i(x) + \sum_{j \in \mathcal{A}(x)} \mu_j \nabla g_j(x).$$

Sea $\nu \in T_X(x)$. Como $T_X(x) = \mathscr{F}_X(x)$, tenemos que

$$\langle \xi, \nu \rangle = \sum_{i \in I} \lambda_i \langle \nabla h_i(x), \nu \rangle + \sum_{j \in \mathcal{A}(x)} \mu_j \langle \nabla g_j(x), \nu \rangle \leq 0.$$

Como $\nu \in T_X(x)$ es arbitrario, concluimos que $\xi \in N_X(x)$. Así, tenemos que $K \subset N_X(x)$.

Ahora, para demostrar que $N_X(x) \subset K$, supongamos por contradicción que existe $\xi \in N_X(x) \setminus K$. Es fácil ver que K es un conjunto convexo y cerrado, y por lo tanto, podemos aplicar el teorema de separación 2.22 entre K y $\{\xi\}$, concluyendo que existe $x^* \in \mathbb{R}^n$ tal que

$$\sup_{\zeta \in K} \langle x^*, \zeta \rangle < \langle x^*, \xi \rangle.$$

Mostraremos ahora que $x^* \in \mathscr{F}_X(x)$.

- Tomemos $i \in I$. Tenemos que para todo $\lambda \in \mathbb{R}$, $\lambda \nabla h_i(x) \in K$ y por lo tanto

  $$\lambda \langle \nabla h_i(x), x^* \rangle \leq \langle \xi, x^* \rangle, \quad \forall \lambda \in \mathbb{R}.$$

  Lo anterior implica que $\langle \nabla h_i(x), x^* \rangle = 0$.
- Tomemos $j \in \mathcal{A}(x)$. Tenemos que para todo $\mu \geq 0$, $\mu \nabla g_j(x) \in K$. Por lo tanto,

  $$\mu \langle \nabla g_j(x), x^* \rangle \leq \langle \xi, x^* \rangle, \quad \forall \mu \geq 0.$$

  Lo anterior implica que $\langle \nabla g_j(x), x^* \rangle \leq 0$.

Universidad de O'Higgins

Mezclando ambos resultados, tenemos que

$$\begin{aligned}\langle \nabla h_i(x), x^* \rangle &= 0, \quad \forall i \in I \\ \langle \nabla g_j(x), x^* \rangle &\leq 0, \quad \forall j \in \mathcal{A}(x)\end{aligned}$$

y por lo tanto, $x^* \in \mathscr{F}_X(x)$. Pero como $\mathscr{F}_X(x) = T_X(x)$ y $\xi \in N_X(x)$, concluimos que

$$0 \geq \langle \xi, x^* \rangle > \sup_{\zeta \in K} \langle \zeta, x^* \rangle = 0,$$

lo cual es una contradicción. Esto prueba que $N_X(x) \subset K$, lo que concluye la demostración. □

Para escribir el Teorema de Karush-Kuhn-Tucker, es útil definir el *Lagrangiano* del problema (4.8), que es la función

$$\begin{aligned}\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^{|I|} \times \mathbb{R}^{|J|} &\to \mathbb{R} \\ (x, \lambda, \mu) &\mapsto f(x) + \sum_{i \in I} \lambda_i h_i(x) + \sum_{j \in J} \mu_j g_j(x).\end{aligned} \tag{4.16}$$

Cuando calculamos el gradiente de $\mathcal{L}$ sólo con respecto a la primera coordenada (la variable $x \in \mathbb{R}^n$), obtenemos que

$$\nabla_x \mathcal{L}(x, \lambda, \mu) = \nabla f(x) + \sum_{i \in I} \lambda_i \nabla h_i(x) + \sum_{j \in J} \mu_j \nabla g_j(x). \tag{4.17}$$

Con esta definición, la primera ecuación de (4.10) es equivalente a la condición $\nabla_x \mathcal{L}(x, \lambda, \mu) = 0$.

Teorema 4.13 (Karush-Kuhn-Tucker)**.** *Sea $\bar{x} \in \mathbb{R}^n$ un mínimo local del Problema (4.8). Si $T_X(\bar{x}) = \mathscr{F}(\bar{x})$, entonces $\bar{x}$ es punto crítico en el sentido de la Definición 4.4, es decir, $\bar{x} \in X$ y*

$$\exists (\lambda, \mu) \in \mathbb{R}^{|I|} \times \mathbb{R}^{|J|} \text{ tal que } \begin{cases} \nabla_x \mathcal{L}(\bar{x}, \mu, \lambda) = 0 \\ \mu_j g_j(\bar{x}) = 0, \quad \forall j \in J, \\ \mu \geq 0. \end{cases} \tag{4.18}$$

Demostración. Como $\bar{x}$ es un mínimo local, entonces tenemos que $-\nabla f(\bar{x}) \in N_X(\bar{x})$. Usando el Lema 4.12, tenemos que existen $(\lambda_i \ : \ i \in I)$ y $(\mu_j \ : \ j \in \mathcal{A}(\bar{x}))$, con $\mu_j \geq 0$ para todo $j \in \mathcal{A}(\bar{x})$, tal que

$$-\nabla f(\bar{x}) = \sum_{i \in I} \lambda_i \nabla h_i(\bar{x}) + \sum_{j \in \mathcal{A}(\bar{x})} \mu_j \nabla g_j(\bar{x}).$$

Completando el vector $(\mu_j \ : \ j \in \mathcal{A}(\bar{x}))$ a $(\mu_j \ : \ j \in J)$ imponiendo $\mu_j = 0$ si $j \notin \mathcal{A}(x_0)$, tenemos que los vectores λ y μ cumplen que

$$\begin{aligned}&\nabla f(\bar{x}) + \sum_{i \in I} \lambda_i \nabla h_i(\bar{x}) + \sum_{j \in J} \mu_j \nabla g_j(\bar{x}) = 0 \\ &\mu_j g_j(\bar{x}) = 0, \quad \forall j \in J, \\ &\mu \geq 0.\end{aligned}$$

Es decir, $\bar{x}$ es punto crítico en el sentido de la Definición 4.4. □

4.1.3. Condiciones de Calificación de Restricciones

En la práctica, el teorema de Karush-Kuhn-Tucker se utiliza de dos maneras:

1. Cuando es posible, resolver de manera analítica el sistema no-lineal de ecuaciones (4.10), y elegir de entre las soluciones aquel punto que minimice el valor de la función objetivo f.

Universidad de O'Higgins

2. Buscar algorítmicamente, a partir de un punto inicial x_0, una solución aproximada x^* del sistema no-lineal de ecuaciones (4.10) que garantice que $f(x^*) \leq f(x_0)$ y entregar el punto encontrado como candidato a solución.

Ambos enfoques están asumiendo que la conclusión del teorema 4.13 se tiene, es decir, que estamos trabajando con problemas donde los mínimos locales son puntos críticos. Sin embargo, esto es válido siempre y cuando los conos tangente y tangente linearizado coincidan en todos los mínimos locales. El siguiente ejemplo, muestra que dicho supuesto puede fallar.

Ejemplo 4.14 Consideremos el problema

$$\begin{cases} \text{mín} & f(x_1, x_2) = x_2 \\ s.a. & g(x_1, x_2) = x_1^2 - x_2^3 \leq 0. \end{cases}$$

La restricción $g(x_1, x_2) \leq 0$ es equivalente a $x_1^2 \leq x_2^3$, lo que implica en particular que si (x_1, x_2) es un punto factible, entonces $x_2 \geq 0$. Luego, el punto $(0, 0)$ es un mínimo global del problema. Sin embargo,

$$\nabla f(0,0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{y} \quad \nabla g(0,0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

En particular, $(0, 0)$ no es un punto crítico en el sentido de la Definición 4.4, y por lo tanto el Teorema 4.13 falla. Como conclusión, necesariamente $T_X(0,0) \neq \mathscr{F}_X(0,0)$. □

El problema del Ejemplo 4.14 se solucionaría si, antes de intentar resolver el problema de optimización (4.1), pudiéramos asegurar que

$$\forall x \in X, \quad T_X(x) = \mathscr{F}_X(x). \tag{4.19}$$

En esta sección, desarrollaremos algunas condiciones suficientes sobre el conjunto de restricciones X de tal manera que podamos asegurar (4.19). Estas condiciones suficientes se conocen como **Calificación de Restricciones**, pues el objetivo es garantizar la condición (4.19) que no depende de la función objetivo. Cuando estamos frente a un problema calificado (en el sentido que verifica (4.19)), podemos utilizar el Teorema 4.13 tranquilamente, confiando tanto en resultados analíticos como en algoritmos (al menos con la misma confianza que teníamos en problemas sin restricciones).

Lema 4.15. *Sea X un conjunto no-vacío y $x \in X$. Entonces*

$$T_X(x) \subset \mathscr{F}_X(x).$$

Demostración. Ejercicio. □

Como se mencionó previamente, el cono linearizado $\mathscr{F}_X(x)$ se obtiene como el cono tangente en x del polítopo Z resultante de reemplazar cada restricción que define X por su aproximación de primer orden en torno a x. Esto nos dice que si X es en si mismo un polítopo (dado por restricciones lineales), entonces (4.19) se debe verificar. La siguiente proposición formaliza este razonamiento.

Proposición 4.16. *Sea X un conjunto no-vacío de la forma (4.2). Si para todo $i \in I$ y para todo $j \in J$, las funciones h_i y g_j son funciones lineales afines, entonces*

$$\forall x \in X, \quad T_X(x) = \mathscr{F}_X(x).$$

Demostración. Sea $x \in X$. Si para todo $i \in I$ y para todo $j \in J$, h_i y g_j son funciones lineales afines, entonces se tiene que estas funciones deben coincidir con sus aproximaciones de primer orden

Universidad de O'Higgins

en torno a x. Es decir, para todo $z \in \mathbb{R}^n$ se tiene que

$$h_i(z) = h_i(x) + \langle \nabla h_i(x), z - x \rangle, \quad \forall i \in I,$$
$$g_j(z) = g_j(x) + \langle \nabla g_j(x), z - x \rangle, \quad \forall j \in J.$$

Sea ahora $\nu \in \mathscr{F}_X(x)$ con $\nu \neq 0$. Vamos a demostrar que existe $\bar{t} > 0$ tal que $[x, x + \bar{t}\nu] \subset X$. Para esto, veremos que $x + t\nu$ satisface las restricciones que definen X:

- Sea $i \in I$. Como $\nu \in \mathscr{F}_X(x)$, tenemos que $\langle \nabla h_i(x), \nu \rangle = 0$. Luego, que para todo $t > 0$ podemos escribir:
$$h_i(x + t\nu) = h_i(x) + t\langle \nabla h_i(x), \nu \rangle = 0.$$

- Sea $j \in \mathcal{A}(x)$. Como $\nu \in \mathscr{F}_X(x)$, tenemos que $\langle \nabla g_j(x), \nu \rangle \leq 0$. Luego, para todo $t > 0$ podemos escribir:
$$g_j(x + t\nu) = g_j(x) + t\langle \nabla g_j(x), \nu \rangle = t\langle \nabla g_j(x), \nu \rangle \leq 0.$$

- Sea $j \in J \setminus \mathcal{A}(x)$. Entonces tenemos que $g_j(x) < 0$. Por continuidad de g_j, existe $\delta_j > 0$ lo suficientemente pequeño tal que $g_j(z) < 0$ para todo $z \in B(x, \delta)$. Luego, tomando $\bar{t}_j = \frac{\delta}{2\|\nu\|}$, tenemos que
$$g_j(x + t\nu) < 0, \quad \forall t \in [0, \bar{t}_j].$$

Ahora, tomando

$$\bar{t} = \begin{cases} 1 & \text{si } \mathcal{A}(x) = J \\ \min\{\bar{t}_j \ : \ j \in J \setminus \mathcal{A}(x)\} & \text{si no,} \end{cases}$$

tenemos que $x + t\nu \in X$ para todo $t \in [0, \bar{t}]$. Luego, $\nu \in T_X(x)$: en la Definición 4.6, basta tomar $\nu_k = \nu$ y $t_k = \frac{\bar{t}}{k}$. □

La condición de que X esté definido únicamente por restricciones lineales (o sea, que todas las funciones h_i y g_j sean lineales afines), se conoce como la condición de **Calificación de Restricciones Lineales (LCQ)**. Si bien, (LCQ) garantiza (4.19), es demasiado restrictiva en el contexto de optimización no lineal: es de esperar que nos encontremos con restricciones no-lineales.

Para estudiar condiciones de calificación de restricciones mejor adaptadas al contexto no-lineal, necesitamos recordar un teorema de cálculo diferencial bastante célebre, llamado el Teorema de la Función Implícita.

Teorema 4.17 (Función Implícita)**.** *Sea $f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^p$ una función de clase $\mathcal{C}^1$ y $(x_0, z_0) \in \mathbb{R}^p \times \mathbb{R}^q$ tal que*

- *$f(x_0, z_0) = 0$; y*
- *El Jacobiano parcial $D_x f(x_0, z_0) = \left[\frac{\partial}{\partial x_1} f(x_0, z_0) \quad \cdots \quad \frac{\partial}{\partial x_p} f(x_0, z_0)\right]$ es invertible.*

Entonces, existen dos abiertos $U_x \subset \mathbb{R}^p$ y $U_z \subset \mathbb{R}^q$ y una función $\phi : U_z \to U_x$ de clase $\mathcal{C}^1$ tal que

1. *$x_0 \in U_x$, $z_0 \in U_z$ y $\phi(z_0) = x_0$.*
2. *Para todo $(x, z) \in U_x \times U_z$, se tiene que $f(x, z) = 0 \Leftrightarrow x = \phi(z)$.*

Basados en el Teorema de la Función Implícita, presentaremos dos tipos de calificación de restricciones. La primera, que es la condición de independencia lineal presentada a continuación en la Definición 4.18, es probablemente la más conocida y simple de enunciar.

Universidad de O'Higgins

Definición 4.18 (LICQ). *Sea X un conjunto no-vacío de la forma (4.2) y sea $x \in X$. Decimos que un punto x satisface la* ***calificación de independencia lineal (LICQ)*** *si el conjunto*

$$\{\nabla h_i(x) \ : \ i \in I\} \cup \{\nabla g_j(x) \ : \ j \in \mathcal{A}(x)\}$$

es linealmente independiente.

Teorema 4.19. *Sea X un conjunto no-vacío de la forma (4.2) y sea $x^* \in X$. Si x^* satisface (LICQ), entonces*

$$T_X(x^*) = \mathscr{F}_X(x^*).$$

Demostración. Sin perder generalidad, asumamos que $\mathcal{A}(x) = J$, y sea $m = |I| + |J|$. Sea $c : \mathbb{R}^n \to \mathbb{R}^m$ la función dada por $c(x) = (h_1(x), \ldots, h_{|I|}(x), g_1(x), \ldots, g_{|J|}(x))$ y tomemos la matriz

$$A(x^*) = Dc(x^*) = \begin{bmatrix} \nabla h_1(x^*) & \cdots & \nabla h_{|I|}(x^*) & \nabla g_1(x) & \cdots & \nabla g_{|J|}(x^*) \end{bmatrix}^T.$$

Como todas las columnas de A son linealmente independientes, necesariamente $m \leq n$ y A tiene rango igual a m. Vamos a suponer que $m < n$: el caso $m = n$ es similar (se omite la matriz Z que se presenta más adelante) y se deja como ejercicio. Sea Z una matriz de $n \times (n-m)$ cuyas columnas forman una base del espacio nulo de A. Esta construcción nos dice que el rango de Z es $n - m$ y que $A(x^*)Z = 0$. En particular, las columnas de Z y las filas de $A(x^*)$ forman una base de $\mathbb{R}^n$.

Ahora, sea $\nu \in \mathscr{F}_X(x)$. Definamos la función $R : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ dada por

$$R(z,t) = \begin{bmatrix} c(z) - tA(x^*)\nu \\ Z^T(z - x^* - t\nu) \end{bmatrix}.$$

Claramente R es de clase $\mathcal{C}^1$ y $R(x^*, 0) = 0$. Más aún,

$$D_x R(x^*, 0) = \begin{bmatrix} A(x^*) \\ Z^T \end{bmatrix},$$

que es de rango completo y por lo tanto invertible. Por el Teorema de la Función Implícita (Teorema 4.17, existen dos abiertos $(-\delta, \delta)$ y $U_z \subset \mathbb{R}^n$ y una función $\phi : (-\delta, \delta) \to U_z$ de clase $\mathcal{C}^1$ tal que $\phi(0) = x^*$ y tal que

$$R(\phi(t), t) = 0, \quad \forall t \in (-\delta, \delta).$$

Sea $t_k = \delta/2k$ y sea $x_k = \phi(t_k)$. Notemos que $x_k \in X$ para todo $k \in \mathbb{N}$. En efecto, recordando que $\nu \in \mathscr{F}_X(x^*)$, tenemos que

$$R(\phi(t), t) = 0 \implies c(\phi(t)) = tA(x^*)\nu \implies \begin{array}{ll} h_i(x_k) = t_k \langle \nabla h_i(x^*), \nu \rangle = 0, & \forall i \in I, \\ g_j(x_k) = t_k \langle \nabla g_j(x^*), \nu \rangle \leq 0, & \forall j \in J. \end{array}$$

Por continuidad de ϕ, tenemos que $x_k \to x^*$. Luego, tenemos que

$$\begin{aligned} 0 = R(t_k, x_k) &= \begin{bmatrix} c(x_k) - t_k A(x^*)\nu \\ Z^T(x_k - x^* - t_k\nu) \end{bmatrix} \\ &\quad \begin{bmatrix} A(x^*)(x_k - x^*) - t_k A(x^*)\nu \\ Z^T(x_k - x^* - t_k\nu) \end{bmatrix} + \underbrace{\begin{bmatrix} c(x_k) - A(x^*)(x_k - x^*) \\ 0 \end{bmatrix}}_{o_k} \\ &= \begin{bmatrix} A(x^*) \\ Z^T \end{bmatrix} (x_k - x^* - t_k\nu) + o_k \end{aligned}$$

Universidad de O'Higgins

Luego, definiendo $\nu_k = t_k^{-1}(x_k - x^*)$, tenemos que $x^* + t_k\nu_k \in X$. Además, tenemos que

$$\underbrace{\begin{bmatrix} A(x^*) \\ Z^T \end{bmatrix}^{-1}}_{=M} \frac{o_k}{t_k} = \nu_k - \nu.$$

Finalmente, recordando que $x_k = \phi(t_k)$ y $x^* = \phi(0)$, y denotando por $L = \|M\|$, podemos escribir

$$\begin{aligned} \|\nu_k - \nu\| &\leq L \frac{\|c(x_k) - A(x^*)(x_k - x^*)\|}{t_k} \\ &= L \underbrace{\frac{\|c(x_k) - c(x^*) - A(x^*)(x_k - x^*)\|}{\|x_k - x^*\|}}_{\to 0} \underbrace{\frac{\|\phi(t_k) - \phi(0)\|}{t_k}}_{\to \|\phi'(0)\|} \\ &\xrightarrow{k\to\infty} 0. \end{aligned}$$

Concluimos entonces que $\nu \in T_X(x^*)$ y por lo tanto, que $\mathscr{F}_X(x^*) \subset T_X(x^*)$. En vista del Lema 4.15, la demostración está completa. □

La condición (LICQ) en general solo se verifica en problemas pequeños (con pocas restricciones). En la práctica, es muy difícil de verificar. Una condición más relajada es considerar la independencia lineal de los gradientes solo para las restricciones de igualdad. En general, estas restricciones no pueden ser demasiadas (de lo contrario el conjunto factible X estaría dado por un sistema de ecuaciones sobredimensionado), y por lo tanto hay buenas opciones de verificar independencia lineal. Las condiciones de Mangasarian-Fromowitz, que se presentan a continuación en la Definición 4.20, mantienen el requerimiento de independencia lineal de las restricciones de igualdad y reemplazan el requerimiento sobre las restricciones de desigualdad por la existencia de una "dirección interior" en el cono linearizado.

Definición 4.20 (MFCQ)**.** *Sea X un conjunto no-vacío de la forma (4.2) y sea $x \in X$. Decimos que un punto x satisface la* ***calificación de Mangasarian-Fromowitz (MFCQ)*** *si*

(i) $\{\nabla h_i(x) \ : \ i \in I\}$ *es linealmente independiente; y*

(ii) Existe $\overline{\nu} \in \mathbb{R}^n$ *tal que*

$$\begin{aligned} \langle \nabla h_i(x), \overline{\nu} \rangle &= 0, \quad \forall i \in I, \\ \langle \nabla g_j(x), \overline{\nu} \rangle &< 0, \quad \forall j \in \mathcal{A}(x). \end{aligned} \tag{4.20}$$

Teorema 4.21. *Sea X un conjunto no-vacío de la forma (4.2) y sea $x^* \in X$. Si x^* satisface (MFCQ), entonces*

$$T_X(x^*) = \mathscr{F}_X(x^*).$$

Demostración. Sea $\nu \in \mathscr{F}_X(x^*)$ y denotemos $p = |I|$. Para que exista $\bar{\nu}$ satisfaciendo la condición (4.20), es necesario que, o bien $p < n$, o bien, $p = n$ y $\mathcal{A}(x) = \emptyset$. En el segundo caso, (MFCQ) se reduce a (LICQ), y por lo tanto el resultado se sigue del Teorema 4.19. Asumiremos entonces que $p < n$. Definamos la matriz

$$A = \begin{bmatrix} \nabla h_1(x^*) & \cdots & \nabla h_p(x^*) \end{bmatrix} \in \mathcal{M}_{n\times p}(\mathbb{R}),$$

y consideremos la función

$$\begin{aligned} \ell : \mathbb{R}^p \times \mathbb{R} &\to \mathbb{R}^p \\ (u, t) &\mapsto h(x^* + t\nu + Au), \end{aligned}$$

Universidad de O'Higgins

donde $h(x) := (h_1(x), \ldots, h_p(x)) \in \mathbb{R}^p$. Tenemos que la función ℓ es de clase $\mathcal{C}^1$, $\ell(0,0) = 0$ y

$$D_u\ell(0,0) = \begin{bmatrix}\nabla h_1(x^*) & \cdots & \nabla h_p(x^*)\end{bmatrix}^T A = A^T A \in \mathcal{M}_{p\times p}(\mathbb{R}).$$

Como A las columnas son linealmente independientes, se tiene que A^TA es definida positiva y por lo tanto, invertible. En efecto, claramente A^TA es semidefinida positiva. Si tomamos $y \in \mathbb{R}^p$ tal que $y^TA^TAy = 0$, tendríamos que

$$0 = y^TA^TAy = \|Ay\|^2,$$

lo que implicaría que $Ay = 0$. Luego, notando que

$$Ay = \sum_{i=1}^{p} y_i A_{\bullet i} = \sum_{i=1}^{p} y_i \nabla h_i(x^*),$$

concluimos que la independencia lineal de las columnas de A implica que $y = 0$. Esto muestra que A^TA es definida positiva como fue enunciado.

Luego, podemos aplicar el Teorema de la Función Implícita (Teorema 4.17), que nos asegura que existe $\delta > 0$, $V_u \subset \mathbb{R}^p$ abierto y $\phi : (-\delta, \delta) \to V$ tal que: $0 \in V_u$, ϕ es de clase $\mathcal{C}^1$ y

$$\forall (t,u) \in (-\delta,\delta) \times V_u, \quad \ell(u,t) = (0,0) \Leftrightarrow u = \phi(t).$$

Definamos ahora la curva $x : (-\delta, \delta) \to \mathbb{R}^n$ dada por

$$x(t) = x^* + t\nu + A\phi(t).$$

Tenemos que $x(0) = x^*$ y que $h_i(x(t)) = 0$ para todo $t \in (-\delta, \delta)$ y todo $i \in I$. Luego, podemos escribir

$$\begin{aligned} 0 = \frac{d}{dt}(h \circ x)(0) &= \begin{bmatrix}\nabla h_1(x(0))^T \cdot x'(0) \\ \vdots \\ \nabla h_p(x(0))^T \cdot x'(0)\end{bmatrix} \\ &= \begin{bmatrix}\nabla h_1(x^*)^T \cdot (\nu + A\phi'(0)) \\ \vdots \\ \nabla h_p(x^*)^T \cdot (\nu + A\phi'(0))\end{bmatrix} \\ &= \begin{bmatrix}\nabla h_1(x^*)^T \cdot A\phi'(0) \\ \vdots \\ \nabla h_p(x^*)^T \cdot A\phi'(0)\end{bmatrix} = A^TA\phi'(0). \end{aligned}$$

Como A^TA es invertible, tenemos que $\phi'(0) = 0$ y por lo tanto, concluimos que $x'(0) = \nu + A\phi'(0) = \nu$.

Supongamos primero que

$$\forall j \in \mathcal{A}(x^*), \quad \langle \nabla g_j(x^*), \nu\rangle < 0. \tag{4.21}$$

Vamos a demostrar que en tal caso existe $\varepsilon \in (0, \delta)$ tal que para todo $t \in (0, \varepsilon)$ y todo $j \in J$, se tiene que $g_j(x(t)) < 0$.

- Tomemos primero $j \in \mathcal{A}(x^*)$. Por regla de la cadena, tenemos que

$$0 > \langle \nabla g_j(x^*), \nu\rangle = \langle \nabla g_j(x(0)), x'(0)\rangle = \lim_{t\to 0} \frac{g_j(x(t)) - g_j(x(0))}{t}.$$

  Luego, recordando que $g_j(x^*) = 0$, la desigualdad anterior nos asegura que existe $\varepsilon_j \in (0, \delta)$ tal que $g_j(x(t)) = g_j(x(t)) - g_j(x(0)) < 0$ para todo $t \in (0, \varepsilon_j)$.

Universidad de O'Higgins

- Tomemos ahora $j \in J \setminus \mathcal{A}(x^*)$. Como la curva $x : (-\delta, \delta) \to \mathbb{R}^n$ es continua y $g_j(x(0)) = g_j(x^*) < 0$, existe $\varepsilon_j \in (0, \delta)$ lo suficientemente pequeño tal que $g_j(x(t)) < 0$ para todo $t \in (0, \varepsilon_j)$.

Finalmente, definiendo $\varepsilon = \min\{\varepsilon_j \ : \ j \in J\}$, tenemos que

$$g_j(x(t)) < 0, \quad \forall j \in J, \quad \forall t \in (0, \varepsilon).$$

Finalmente, con esta construcción tenemos que $x(t) \in X$ para todo $t \in (0, \varepsilon)$ y por lo tanto, podemos definir las sucesiones (ν_k) y (t_k) de la Definición 4.6 como

$$t_k = \frac{\varepsilon}{2k} \quad \text{y } \nu_k = \frac{x(t_k) - x(0)}{t_k}.$$

Así, tenemos que $x^* + t_k \nu_k = x(t_k) \in X$ para todo $k \in \mathbb{N}$, que $t_k \to 0$ y finalmente que $\nu_k \to x'(0) = \nu$. Por lo tanto, $\nu \in T_X(x^*)$.

Ahora, para el caso general (sin suponer (4.21)), la condición (MFCQ) nos dice que existe $\bar{\nu} \in \mathscr{F}_X(x^*)$ tal que $\langle \nabla g_j(x^*), \bar{\nu} \rangle < 0$ para todo $j \in \mathcal{A}(x^*)$. En tal caso, para $k \in \mathbb{N}$ podemos definir

$$\nu_k = \nu + \frac{1}{k}\bar{\nu}$$

que cumple que $\nu_k \in \mathscr{F}_X(x^*)$ y además verifica (4.21). Repitiendo todo el desarrollo para ν_k en vez de ν, concluimos que $\nu_k \in T_X(x^*)$. Luego, como $T_X(x^*)$ es cerrado (ver Proposición 4.7), concluimos que

$$\nu_k \to \nu \in T_X(x^*).$$

Como ahora $\nu \in \mathscr{F}_X(x^*)$ es arbitrario, concluimos que $\mathscr{F}_X(x^*) \subset T_X(x^*)$, lo que concluye la demostración. □

4.2 Algoritmo de Punto Interior

[PENDIENTE]

4.3 Algoritmos de Lagrangiano Aumentado

[PENDIENTE]

4.4 Ejercicios Capítulo 4

P1.

Universidad de O'Higgins

Capítulo 5

Algoritmos para optimización no-diferenciable

[PENDIENTE]

5.1 Método de Máximo Descenso con subgradientes

[PENDIENTE]

5.2 Método de Descenso Coordinado

[PENDIENTE]

5.3 Metaheuristica de Enjambre de partículas

[PENDIENTE]

Universidad de O'Higgins

Apéndice A

Contenidos previos de Cálculo multivariado

Universidad de O'Higgins

Bibliografía

[1] Álvarez, F. *Análisis Convexo y Dualidad - Apuntes del curso*. Departamento de Ingeniería Matemática, Universidad de Chile, 2012.

[2] Bertsekas, D. P. *Convex optimization theory*, Third ed. Athena Scientific Optimization and Computation Series. Athena Scientific, Belmont, MA, 2009.

[3] Bertsekas, D. P. *Nonlinear Programming*, Third ed. Athena Scientific Optimization and Computation Series. Athena Scientific, Belmont, MA, 2016.

[4] Eaton, J. W., Bateman, D., Hauberg, S., and Wehbring, R. *GNU Octave version 5.2.0 manual: a high-level interactive language for numerical computations*, 2020.

[5] The Mathworks, Inc. *MATLAB R2021a*. Natick, Massachusetts, 2021.

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

[7] Phelps, R. R. *Convex functions, monotone operators, and differentiability*. Springer-Verlag, Berlin New York, 1993.

Universidad de O'Higgins