# FUNDAMENTOS DE OPTIMIZACIÓN

Notas de curso

Marcelo Fiori

2025

# Índice general

## Capítulo 1

# Introducción

En este curso nos vamos a ocupar del problema de minimizar funciones. Vamos a cubrir aspectos básicos de la teoría y de lo métodos que están detrás de muchísimas aplicaciones modernas en todas las ramas de la ingeniería.

Formalmente, si tenemos una función $f : \mathbb{R}^n \to \mathbb{R}$, y queremos encontrar el valor mínimo que toma $f$ en un subconjunto $\mathscr{X} \subset \mathbb{R}^n$, escribimos ese problema como

$$\min_{x \in \mathscr{X}} f(x) \tag{1.1}$$

En realidad, en muchas aplicaciones, en vez de hallar el valor mínimo que toma $f$, lo que importa es el lugar donde se alcanza el mínimo. Es decir, el punto de $\mathscr{X}$ donde la función toma su valor mínimo. A este problema lo denotamos así:

$$\arg\min_{x \in \mathscr{X}} f(x) \tag{1.2}$$

aunque muchas veces haremos un abuso de notación y nos vamos a referir simplemente al problema de minimizar una función.

El mínimo de una función puede existir o no, se puede alcanzar en un único punto o en varios, y tanto el conjunto $\mathscr{X}$ como la función $f$ pueden tener propiedades que ayuden a la tarea, o no.

**Ejercicio 1.1**
*Para cada item de abajo, pensar en funciones de una variable y conjuntos $\mathscr{X} \subset \mathbb{R}$ tales que:*

- *$f$ no tenga mínimo en $\mathscr{X}$*
- *$f$ sea acotada pero no tenga mínimo*

- *$f$ tenga un mínimo que se alcance en un único punto*
- *$f$ alcance el mínimo en varios puntos*

Durante todo el curso vamos a hablar de minimizar una función, pero observar que si el problema que queremos resolver en realidad implica maximizar una función $f$, basta considerar la función opuesta $-f$, y convertimos el problema entonces en uno de minimización. Es decir, no perdemos generalidad al concentrarnos en problemas de minimización, y nos permite facilitar la notación utilizada.

Ya durante los primeros cursos de cálculo buscamos minimizar funciones, y también aparece este concepto en los cursos de álgebra lineal. Por ejemplo, la proyección ortogonal de un punto $p$ a un subespacio es aquel punto del subespacio que minimiza la distancia con el punto $p$. En los casos trabajados, estos problemas tienen soluciones más o menos simples, como por ejemplo buscar ceros de la derivada. La realidad en general es más compleja que esto, y en este curso veremos métodos para resolver problemas más generales, y en particular los utilizados en aplicaciones modernas.

**Ejemplo 1.2**
*El problema de Mínimos Cuadrados, que ya habrán visto en algún curso de álgebra lineal o de métodos numéricos, es uno de los ejemplos más conocidos de problemas de optimización, y nos servirá a lo largo del curso para mostrar varias ideas.*

*Consideremos, como motivación, el problema de regresión lineal. Esto es, tenemos N datos que de la forma $(a_i, b_i)$, y queremos encontrar la mejor función lineal para expresar los $b_i$ en términos de los $a_i$. Es decir, buscamos una recta $f(a) = am + n$, de manera tal que para los datos que tenemos, $f(a_i) \approx b_i$.*

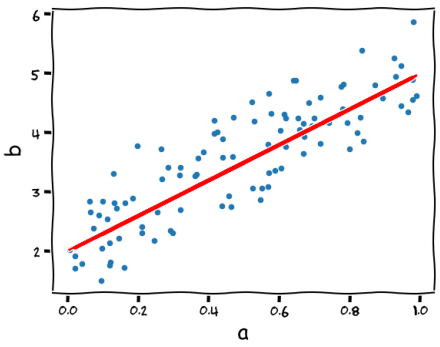



Figura 1.1: Regresión lineal.

*Si por "mejor" entendemos que minimice el error cuadrático medio, es decir*

$$\min_{m,n} \sum_{i=1}^{N} \|b_i - (ma_i + n)\|_2^2$$

*entonces estamos ante un problema de mínimos cuadrados.*

*Este problema se puede escribir en general (y en más dimensiones) de forma matricial:*

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|_2^2,$$

*donde agrupamos todas las variables de la función lineal en el vector x.*

*Tanto de forma analítica (derivando) como geométrica (como proyección ortogonal), se puede encontrar la solución de forma cerrada, y por eso es un modelo muy utilizado: por su simplicidad, y su baja complejidad algorítimica para encontrar la solución.*

*Sin embargo, incluso ligeras modificaciones del problema hacen que la solución ya no se pueda calcular de forma cerrada, y se necesiten métodos como los que veremos en el curso para resolverlo.*

Comencemos revisando algunos conceptos de cursos anteriores que van a ser fundamentales en lo que sigue.

## 1.1. Repaso

**Definición 1.3.** Sea $f:\mathbb{R}^n\to\mathbb{R}$, decimos que $x^\star$ es un mínimo local o mínimo relativo si existe $\delta>0$ tal que $f(x)\geqslant f(x^\star)$ para todo $x\in B(x^\star,\delta)$.

**Definición 1.4.** Sea $f:\mathbb{R}^n\to\mathbb{R}$, decimos que $x^\star$ es un mínimo global o mínimo absoluto si $f(x)\geqslant f(x^\star)$ para todo $x\in\mathbb{R}^n$.

Muchas veces vamos a buscar mínimos en un subconjunto $\mathscr{X}$. En ese caso, las definiciones anteriores se generalizan trivialmente, considerando que las desigualdades se cumplen para $x\in\mathscr{X}$, o para $x\in\mathscr{X}\cap B(x^\star,\delta)$.

Como fue mencionado arriba, el objetivo será siempre encontrar mínimos (locales o globales) de funciones. Observar entonces que el dominio de la función (que en la definición es $\mathbb{R}^n$ o $\mathscr{X}$) puede ser un conjunto más general. Por ejemplo, el mínimo local se puede definir en cualquier conjunto donde tengamos una distancia, para poder hablar de entorno de $x^\star$. Ahora, el codominio no puede ser cualquier conjunto: necesitamos un orden. Por lo tanto vamos a trabajar siempre con funciones a valores reales.

Otra noción fundamental que vamos a utilizar es la de diferenciabilidad:

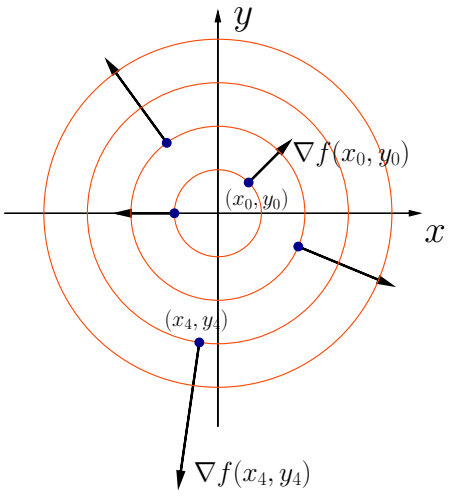
**Definición 1.5.** Sean $f:\mathbb{R}^n\to\mathbb{R}$, y $a\in\mathbb{R}^n$. Decimos que la función $f$ es diferenciable en $a$ sii existe una transformación lineal $df_a:\mathbb{R}^n\to\mathbb{R}$ tal que

$$f(a+h)=f(a)+df_a(h)+r(h)$$

con $r:\mathbb{R}^n\to\mathbb{R}$ una función que cumple $\lim_{h\to 0}\frac{r(h)}{||h||}=0$.

Dada $f : \mathbb{R}^n \to \mathbb{R}$ diferenciable en el punto $a$, definimos el vector gradiente de $f$ en $a$ como $\nabla f(a) = \left(\frac{\partial f}{\partial x_1}(a), \frac{\partial f}{\partial x_2}(a), \ldots, \frac{\partial f}{\partial x_n}(a)\right)$.

Recordemos que si $f$ es diferenciable, podemos expresar la derivada direccional como el producto interno del gradiente con la dirección: $\frac{\partial f}{\partial v}(a) = \langle \nabla f(a), v \rangle$. Esto tiene una consecuencia muy importante. Si $v$ es un vector unitario, como $\frac{\partial f}{\partial v}(a) = \langle \nabla f(a), v \rangle$, entonces la derivada direccional se maximiza cuando los vectores $v$ y $\nabla f(a)$ son colineales. Es decir, la dirección del gradiente es la dirección de máximo crecimiento de $f$.
En la figura podemos observar algunas curvas de nivel la función $f(x,y) = x^2 + y^2$, y los vectores gradientes en algunos puntos.




Esta función tiene su mínimo (relativo y absoluto) en $(0,0)$, y el gradiente vale $\nabla f(0,0) = (0,0)$. Este hecho, que ya lo conocemos de cursos anteriores, será muy importante en este curso.

**Proposición 1.6** (Condición necesaria de optimalidad)**.** Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función diferenciable en un entorno de $x^\star$, y supongamos que $x^\star$ es un mínimo relativo de $f$ e interior al dominio de $f$, entonces $\nabla f(x^\star) = 0$.

Por supuesto que no es una condición suficiente:

**Ejercicio 1.7**
*Buscar una función $f$ y un punto $a$ tal que $\nabla f(a) = 0$ pero que $a$ no sea un mínimo ni máximo relativo de $f$.*

Volviendo al ejemplo anterior, observar en la figura que el vector opuesto al gradiente apunta siempre hacia el mínimo (el origen en este caso). Esto no es general, pero sí observar que si la dirección del gradiente es la de máximo crecimiento, entonces la dirección opuesta es la de máximo **decrecimiento**. Recordar que este es un concepto *local*, pero va a ser muy útil en lo que sigue.

Más aún, supongamos que estamos en algún punto del dominio, y nos queremos mover en una dirección "corecta", es decir, una dirección donde la función decrezca respecto a donde estamos (recordar que estamos buscando minimizar el valor funcional). Entonces lo que necesitamos es que la derivada direccional en la dirección que nos pretendemos mover sea negativa. Y si recordamos que la derivada direccional se obtiene como el producto interno del

gradiente con la dirección, entonces las direcciones que nos sirven son aquellas que tienen un producto interno negativo con el gradiente. Esto es: que formen un ángulo de más de 90° con el gradiente.

**Digresión**

Antes de pasar a los primeros métodos de optimización, veamos algo de notación, y cómo derivar algunas funciones matriciales.

Recordemos que la norma euclídea en $\mathbb{R}^n$ se define como $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$, y esta norma proviene del producto interno usual. En efecto, $\|x\|_2^2 = \langle x, x\rangle$. En las matrices tenemos una norma similar, en el sentido que suma los cuadrados de todas las entradas. Se denomina norma de Frobenius, se define como $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$, y también proviene de un producto interno. Tenemos que $\|A\|_F^2 = \langle A, A\rangle$ para el producto interno definido como $\langle A, B\rangle = tr(AB^T)$.

Sea $A$ una matriz $n \times n$, consideremos la función $f : \mathbb{R}^n \to \mathbb{R}$ definida como $f(x) = \frac{1}{2}\|Ax\|_2^2$, e intentemos calcular su gradiente.

Observar que se puede escribir $f(x) = \frac{1}{2}\langle Ax, Ax\rangle$.

Veamos qué resulta de evaluar la función en $x+h$. La idea será que $h$ va a ser un incremento, y veremos si podemos identificar los elementos de la definición de diferenciabilidad de la expresión.

Tenemos entonces:

$$\begin{aligned} f(x+h) &= \tfrac{1}{2}\langle A(x+h), A(x+h)\rangle = \tfrac{1}{2}\left(\langle Ax, Ax\rangle + \langle Ax, Ah\rangle + \langle Ah, Ax\rangle + \langle Ah, Ah\rangle\right) \\ &= f(x) + \langle Ax, Ah\rangle + \tfrac{1}{2}\langle Ah, Ah\rangle \\ &= f(x) + \langle A^\top Ax, h\rangle + \tfrac{1}{2}\langle Ah, Ah\rangle. \end{aligned}$$

Tenemos entonces que $f(x+h)$ lo podemos escribir como $f(x)$ más un término lineal en $h$ ($\langle A^\top Ax, h\rangle$), y un término de orden mayor (cuadrático en $h$). Por lo tanto, el término lineal tiene que ser el diferencial aplicado a $h$, y en particular el gradiente de $f$ es $\nabla f(x) = A^\top Ax$.

Este procedimiento puede ser útil cuando tenemos funciones vectoriales o matriciales que podemos describir en términos de productos internos u otros operadores. Por las dudas: el procedimiento consiste en escribir $f(x+h)$, y buscar en la expresión resultante un término lineal en $h$, y que el resto sea de orden superior.

Una práctica usual para corroborar que hicimos bien las cuentas, es calcular numéricamente las derivadas parciales, y comparar con el resultado algebraico que obtuvimos. Por ejemplo, en un punto $x$ cualquiera, y con $h = (1 \times 10^{-7}, 0, 0, \ldots, 0)$, podemos estimar la derivada parcial respecto a la primera coordenada como $\dfrac{f(x+h) - f(x)}{1 \times 10^{-7}}$.

## 1.2. Métodos iterativos

En cursos anteriores hemos resuelto problemas de optimización sencillos. Por ejemplo, si queremos hallar el mínimo de la función $f(x) = x^2 - 2x + 5$, entonces simplemente derivamos, $f'(x) = 2x - 2$, e igualamos a cero. El único punto donde se anula la derivada es $x = 1$, y es fácil comprobar que es mínimo local y global.

En estos casos podíamos hallar la solución de forma cerrada, pero en general no vamos a tener esa suerte. Una alternativa muy usada es utilizar métodos iterativos para hallar o aproximar soluciones. Esto es: supongamos que queremos hallar un mínimo de una función $f$ sobre un conjunto $\mathscr{X}$, como arriba. En un método iterativo construimos una sucesión de puntos en $\mathbb{R}^n$, que llamaremos $x_k$, tal que $\lim_{k\to\infty} x_k = x^\star$, donde $x^\star$ puede ser un mínimo local o global, dependiendo de varios elementos, como por ejemplo el método, las propiedades de la función, o del conjunto.

Supongamos que somos capaces de idear un método así, que sea convergente a un mínimo local por ejemplo. ¿Cómo caracterizamos qué tan rápido converge a este mínimo?

Para esto necesitamos una manera de medir qué tan lejos estamos en cada momento, y eso lo haremos con una función de error, que debe ser una función $e : \mathbb{R}^n \to \mathbb{R}$, continua, y que cumpla $e(x) \geqslant 0$ para todo $x$, y que $e(x^\star) = 0$. Por ejemplo, podríamos considerar $e(x) = \|x - x^\star\|$, o $e(x) = |f(x) - f(x^\star)|$.

**Definición 1.8.** Decimos que un método converge linealmente si $\limsup_{k\to\infty} \dfrac{e(x_{k+1})}{e(x_k)} \leqslant \beta < 1$.

Podemos pensar en el límite tradicional en vez del límite superior, que va a ser suficiente en la gran mayoría de los casos en este curso. El término "lineal" se refiere a que de una iteración a la siguiente, el error se reduce linealmente, con un factor $\beta$. Pero observar que esto quiere decir que $e(x_k)$ decrece aproximadamente como $\beta^k$, ya que $e(x_{k+1}) \approx e(x_k)\beta$, y podemos repetir este argumento de forma recursiva. Formalmente, podemos escribir $e(x_k) = O(\beta^k)$.

**Definición 1.9.** Cuando $\limsup_{k\to\infty} \dfrac{e(x_{k+1})}{e(x_k)} = 0$ decimos que la convergencia es superlineal.

Un caso particular de la convergencia superlineal es la convergencia cuadrática, que se define de la siguiente forma:

**Definición 1.10.** Decimos que un método tiene convergencia cuadrática si se cumple que $\limsup_{k\to\infty} \dfrac{e(x_{k+1})}{e^2(x_k)} \leqslant cte < +\infty$.

Por supuesto que la convergencia cuadrática es más rápida que la lineal. En general los métodos cuadráticos suelen ser muy rápidos. Los métodos lineales pueden llegar a ser lentos,

sobre todo si el valor de $\beta$ es muy cercano a uno, pero son suficientemente rápidos para muchas aplicaciones.

Veamos algunos primeros ejemplos de métodos de optimización en una variable. Esto nos va a permitir introducir algunos conceptos, pero además tiene interés en sí mismo, ya que algunos métodos más avanzados tienen como sub-problema una optimización en $\mathbb{R}$, por ejemplo para determinar cuánto avanzar en una dirección. Comencemos con las ideas más sencillas.

**Idea 1.** Supongamos que queremos hallar el mínimo de una función en un intervalo $[a,b]$. Una primera idea, bastante brutal, es dividir el intervalo con varios puntos intermedios, evaluar la función en todos los puntos, y quedarnos con el punto con menor valor funcional. Es decir, construimos una grilla en $[a,b]$, y buscamos el mínimo en ese conjunto discreto. Esto se denomina *grid search*, y, dependiendo de la aplicación, puede llegar a ser útil. Es una idea que escala terriblemente mal para dimensiones mayores.

**Idea 2.** Una segunda idea consiste en ir dividiendo el intervalo $[a,b]$, de manera similar al método de bipartición que se puede utilizar para encontrar raíces. Como ahora buscamos un mínimo, hay que modificar un poco el método. Por ejemplo, en el método de bipartición en cada paso tenemos dos puntos (un intervalo), mientras que en este método vamos a tener tres puntos en cada iteración.

Veamos la idea intuitiva con un ejemplo, y luego formalizamos algunos parámetros. Supongamos que tenemos una función unimodal, que tiene un solo mínimo relativo en $[a,b]$, y que tenemos un punto intermedio $x_2$, con evaluaciones como en la figura (luego veremos cómo ubicar $x_2$). El punto de partida son estos tres puntos. Ahora evaluamos la función en un nuevo punto $x_3$. Puede suceder que el valor funcional en $x_3$ sea mayor o menor que $f(x_2)$. En el primer caso, el mínimo estará en el intervalo $(a,x_3)$, y en el otro caso estará en el $(x_2,b)$ (ver figura). Nos quedamos con la terna de puntos donde se encuentra el mínimo, y volvemos a estar en las mismas condiciones del inicio (tres puntos, y el punto del medio con un valor funcional menor que los otros dos). Repetimos entonces el proceso, hasta llegar a un intervalo suficientemente chico.

Bien, la pregunta entonces es cómo elegir los parámetros. En cada paso, elegiremos uno u otro intervalo dependiendo de la evaluación. Sería bueno que estos intervalos estén balanceados (que sean de igual tamaño). Por otro lado, también sería bueno que en cada iteración el intervalo se reduzca por un mismo factor (igual que en el caso de bisección).

Llamemos $A, B$ y $C$ a las longitudes de los intervalos que muestra la figura. Recordemos que, en función del valor de $f(x_3)$, los posibles intervalos resultantes de la iteración son $(a,x_3)$, que tiene tamaño $A+C$, y $(x_2,b)$, que tiene tamaño $B$. Imponiendo que tengan el mismo largo, obtenemos que $A+C=B$.

Impongamos ahora que el factor de reducción sea constante. En la terna original $(a,x_2,b)$, consideremos el cociente entre el largo total $(A+B)$ y el largo del intervalo "chico" $(A)$, es decir,

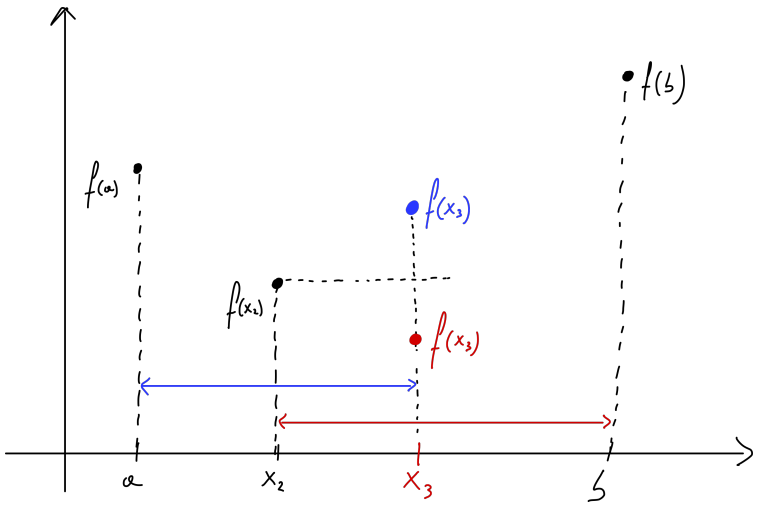



Figura 1.2: Golden section method: mecanismo e iteración.

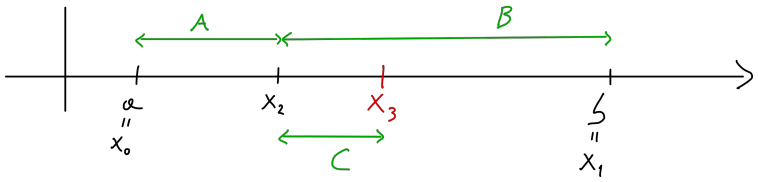



Figura 1.3: Golden section method: parámetros.

$\frac{A+B}{A}$. Ahora hagamos lo mismo en los intervalos de la siguiente iteración (alcanzaría hacerlo para uno de ellos, porque ya impusimos que se comportan igual). En el caso de $(a, x_2, x_3)$ el cociente análogo resulta $\frac{A+C}{C}$, y en el caso de $(x_2, x_3, b)$, resulta $\frac{B}{C}$. Tenemos que imponer entonces que:

$$\frac{A+B}{A} = \frac{A+C}{C} = \frac{B}{C}.$$

En realidad, como decíamos, con una de las igualdades es suficiente. Tomemos la primera, y usemos que $A+C=B$ para eliminar una de las variables:

$$\begin{aligned} \frac{A+B}{A} &= \frac{A+C}{C} \\ \frac{A+B}{A} &= \frac{B}{B-A} \Longrightarrow AB = (A+B)(A-B) \Longrightarrow A^2 + AB - B^2 = 0 \\ &\Longrightarrow A = \frac{-B \pm \sqrt{5B^2}}{2} = \left(\frac{-1 \pm \sqrt{5}}{2}\right) B. \end{aligned}$$

De los dos valores de $\left(\frac{-1\pm\sqrt{5}}{2}\right)$, el positivo es $\left(\frac{-1+\sqrt{5}}{2}\right) \approx 0{,}618$, que será entonces la razón que usaremos para elegir en cada iteración, el próximo punto a evaluar.

Este valor ($\left(\frac{-1+\sqrt{5}}{2}\right)$) es el inverso del denominado número áureo, y por eso este método se denomina "método de la sección áurea" (o *Golden Section Method* en inglés) [11, 12].

Si bien no es una generalización de este algoritmo, existe un método con una filosofía similar, basándose solamente en evaluaciones de la función, que funciona en dimensiones mayores. Se denomina Nelder-Mead [7], y es muy utilizado. Por ejemplo, es uno de los métodos implementados en scipy.optimize.minimize, de Python.

**Idea 3.** Observar que en las dos ideas anteriores solamente utilizamos información de la función, pero no de su derivada (de hecho la función a optimizar perfectamente podría ser no derivable). Sin embargo, ya vimos que el gradiente, o la derivada en este caso de funciones de una variable, tiene mucha información valiosa. Por un lado, nos indica crecimiento o decrecimiento (o la dirección de máximo crecimiento en el caso de varias variables), y por otro lado, en general buscamos puntos críticos, para encontrar candidatos a mínimos.

Una forma de utilizar la derivada es la siguiente. Pensemos que estamos en un momento dado en un punto $x_k$, con derivada $f'(x_k)$. Si esta derivada es positiva, quiere decir que localmente la función está creciendo, y por lo tanto para buscar un mínimo debería moverme "hacia atrás" un poco, hacia un nuevo punto $x_{k+1}$. ¿Cuánto? Ya veremos, por ahora llamemos $\alpha$ a ese poco que nos movemos. Si por otro lado la derivada es negativa, entonces la función es localmente decreciente, y por lo tanto es conveniente seguir avanzando "hacia la derecha" para disminuir el valor funcional. Estas dos situaciones las podemos resumir como

$$x_{k+1} = x_k - \alpha f'(x_k).$$

Naturalmente este método repite esta iteración, como en la figura. Observar que este método "se estanca" cuando la derivada es cero (o muy chica), que es justamente lo que buscamos: un punto crítico, que puede llegar a ser un mínimo local o global.

La versión en varias variables de este método (y sus variantes) es uno de los algoritmos más usados en muchas aplicaciones modernas, y será muy discutido a lo largo de este curso.

A menos de determinar el valor del paso $\alpha$, el algoritmo entonces es: partimos de un punto inicial $x_0$, y repetimos $x_{k+1} = x_k - \alpha f'(x_k)$ hasta llegar a un punto con derivada suficientemente chica.

**Ejercicio 1.11**
*Pensar gráficamente una función para la cual el resultado del algoritmo dependa del punto inicial $x_0$, y otra función para la cual el punto al que converge es independiente del punto inicial.*

**Idea 4.** Pensemos ahora que la función es más suave, por ejemplo tres veces derivable, y de nuevo estamos parados en un punto $x_k$. Si la función es suave, podemos hacer uso de nuestra

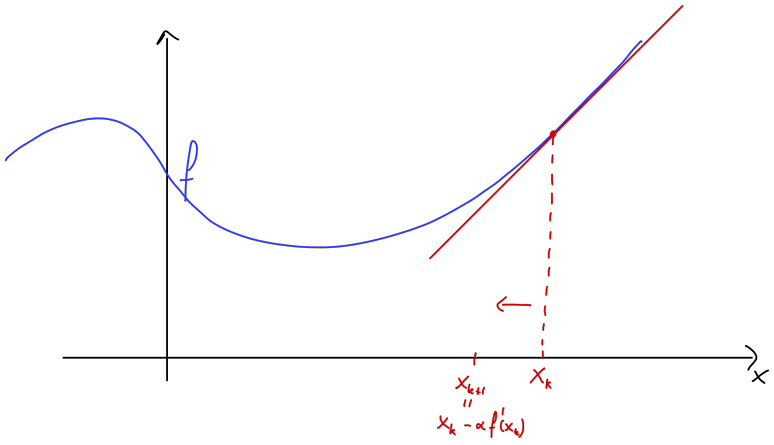



Figura 1.4: En un punto $x_k$, la derivada positiva nos indica que debemos movernos "hacia atrás".

herramienta favorita para entender cómo es localmente la función: Taylor. Tenemos que $f$ es localmente parecida a su polinomio de Taylor de segundo orden,

$$P_2(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2} f''(x_k)(x - x_k)^2.$$

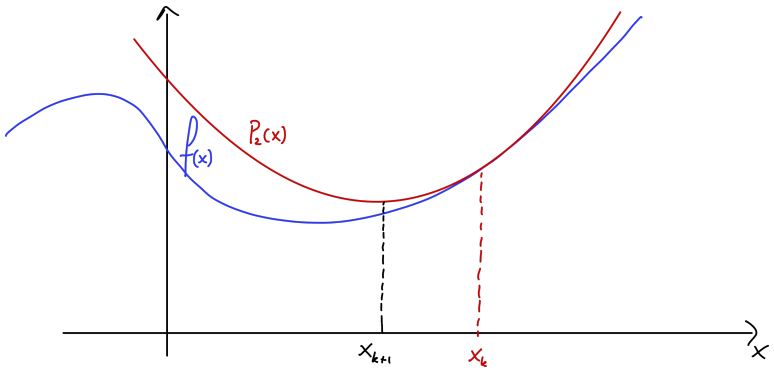



Figura 1.5: Aproximación de Taylor de orden 2 en $x_k$.

Olvidemos por ahora que esto es una aproximación, y supongamos que $f$ es exactamente este polinomio cuadrático, para fijar ideas con derivada segunda positiva como en la figura. Entonces sabemos encontrar su mínimo explícitamente, derivando e igualando a cero.

$$f'(x_k) + f''(x_k)(x - x_k) = 0,$$

por lo que el polinomio $P_2(x)$ se minimiza en $x = x_k - \frac{f'(x_k)}{f''(x_k)}$. Como en realidad trabajamos sobre una aproximación, este $x$ no será el mínimo de $f$, pero si la aproximación es buena,

entonces $x$ será una mejor aproximación al mínimo del punto $x_k$ donde estábamos. Naturalmente el método iterativo es entonces

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)},$$

que es el denominado *Método de Newton*.

Si usted, lector/a, ha pasado por un curso de métodos numéricos o similar, esto puede recordarle al método de Newton-Raphson ($\sim$1670) para hallar raíces. Efectivamente, este método de Newton al que llegamos aquí se puede ver como la aplicación del método de Newton-Raphson pero a la derivada de $f$, para hallar una raíz (es decir, un punto crítico de $f$).

El método de Newton tiene un comportamiento que hace que el estudio de cuándo converge a un mínimo local o a otro, sea extremadamente difícil[1]. Sin embargo, si estamos en la denominada "cuenca de atracción" de un mínimo, entonces el método de Newton converge cuadráticamente. Es decir, muy rápido.

Para ver esto, llamemos $g$ a la derivada de $f$, $g(x) = f'(x)$, y consideremos la iteración

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}.$$

Es decir, el método de Newton escrito de otra forma (en función de $g$), o el método de Newton-Raphson aplicado a $g$.

Sea entonces $g$ dos veces derivable, y $x^\star$ un punto tal que $g(x^\star) = 0$. Entonces, si $x_0$ está "suficientemente cerca" de $x^\star$, entonces la iteración de arriba converge (al menos) cuadráticamente a $x^\star$.

Hagamos el desarrollo de Taylor de $g$ en $x_k$, utilizando la fórmula del resto de Lagrange:

$$g(x) = g(x_k) + g'(x_k)(x - x_k) + \frac{1}{2}g''(c)(x - x_k)^2,$$

donde $c$ es un punto entre $x$ y $x_k$.

Este desarrollo es válido para todo $x$, en particular para $x = x^\star$. Evaluamos entonces en $x = x^\star$:

$$g(x^\star) = g(x_k) + g'(x_k)(x^\star - x_k) + \frac{1}{2}g''(c)(x^\star - x_k)^2,$$

Ahora, como $g(x^\star) = 0$, podemos escribir, dividiendo todo entre $g'(x_k)$:

$$0 = \frac{g(x_k)}{g'(x_k)} + (x^\star - x_k) + \frac{g''(c)}{2g'(x_k)}(x^\star - x_k)^2.$$

Reordenando los términos, llegamos a que

[1] De hecho funciones muy simples dan lugar a comportamientos fractales, como en LINK AL EJEMPLO DE Z3.

$$\underbrace{x_k - \frac{g(x_k)}{g'(x_k)}}_{x_{k+1}} - x^\star = \frac{g''(c)}{2g'(x_k)}(x^\star - x_k)^2.$$

Es decir:

$$|x_{k+1} - x^\star| = \left|\frac{g''(c)}{2g'(x_k)}\right| |x^\star - x_k|^2 .$$

Para la función de error $e(x) = |x - x^\star|$, tenemos entonces que

$$\frac{e(x_{k+1})}{e(x_k)^2} = \left|\frac{g''(c)}{2g'(x_k)}\right| \leqslant \frac{C_2}{2C_1}.$$

donde usamos que como $g \in C^2$, localmente podemos acotar superiormente la derivada segunda por $C_2$, e inferiormente la derivada primera por $C_1$. Observar que este decrecimiento en la función de error es exactamente la definición de convergencia cuadrática.

# Capítulo 2

# Convexidad

La convexidad es muy importante en optimización por varios motivos. Por un lado, en los problemas convexos no puede haber varios mínimos aislados, y existen métodos generales para encontrar un mínimo global. Por otro lado, los problemas convexos sirven como problemas auxiliares para resolver problemas más complejos, o como relajaciones de problemas no convexos. Por estos motivos la optimización convexa es muy estudiada, y hay varios libros de referencia como [4, 8, 9, 3], además de que cada libro sobre optimización tiene un apartado sobre convexidad.

**Definición 2.1.** Decimos que una función $f : \mathbb{R}^n \to \mathbb{R}$ es convexa, si se cumple:

$$f(tx+(1-t)y) \leqslant tf(x)+(1-t)f(y), \quad \forall\, x,y \in \mathbb{R}^n,\ \forall\, t \in [0,1].$$

Además, decimos que $f$ es **estrictamente** convexa si la desigualdad anterior es estricta para todo $x \neq y$, y todo $t \in (0,1)$.

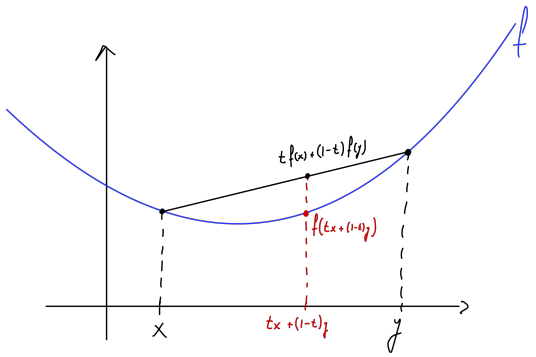



Figura 2.1: Definición de convexidad.

Veamos geométricamente qué es la convexidad en una función de una variable. Al variar $t$ en $[0,1]$, la expresión de la izquierda $f(tx+(1-t)y)$ va recorriendo la gráfica de la función, entre los puntos $(x,f(x))$ e $(y,f(y))$. Por otro lado, en la expresión de la derecha lo que hacemos es combinar linealmente los valores de $f(x)$ y $f(y)$, y lo resultantes es justamente la recta que une los puntos (la cuerda). La definición de arriba lo que dice entonces, es que una función es convexa cuando el gráfico queda siempre por debajo de cualquier cuerda.

En el caso en que además la función es diferenciable, tenemos el siguiente resultado.

**Proposición 2.2.** Si $f:\mathbb{R}^n \to \mathbb{R}$ es una función convexa y diferenciable, entonces

$$f(z) \geqslant f(x) + \nabla f(x)^T (z-x),\ \forall\, x,z \in \mathbb{R}^n.$$

Antes de ver la prueba, interpretemos nuevamente esta desigualdad de forma gráfica, en el caso univariado. En este caso la desigualdad es $f(z) \geqslant f(x) + f'(x)(z-x),\ \forall\, x,z \in \mathbb{R}$. Observar que, fijado $x$ y pensando la función en términos de $z$, a la izquierda tenemos la función en sí, y a la derecha tenemos la recta tangente a $f$ por $(x,f(x))$. Es decir, si $f$ es convexa, el gráfico de la función está siempre soportado por la reta tangente en todo punto. Esta interpretación se generaliza a varias variables, y resulta que el gráfico de $f$ está soportado por el hiperplano tangente en todo punto.

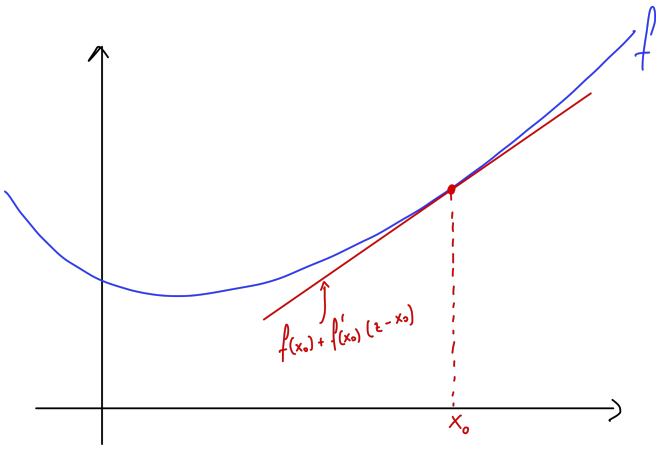



Figura 2.2: Convexidad y diferenciabilidad: la gráfica de $f$ está soportada por la recta tangente en todo punto.

*Demostración.* Como $f$ es convexa, tenemos entonces que se cumple

$$f(tz+(1-t)x) \leqslant tf(z)+(1-t)f(x), \quad \forall\, x,z \in \mathbb{R}^n,\ \forall\, t \in [0,1].$$

Observar que esto es equivalente a:

$$f(x+t(z-x)) \leqslant f(x)+t\,(f(z)-f(x)), \quad \forall\, x,z \in \mathbb{R}^n,\ \forall\, t \in [0,1].$$

Y ahora reordenando términos, tenemos que

$$\frac{f(x+t(z-x))-f(x)}{t} \leqslant f(z)-f(x), \quad \forall\, x,z \in \mathbb{R}^n,\ \forall\, t \in (0,1].$$

Como $f$ es diferenciable, existen todas sus derivadas direccionales. En particular existe la derivada direccional en la dirección $v := z-x$. Por lo tanto, tomando límite en la desigualdad anterior, se obtiene:

$$\frac{\partial f}{\partial v}(x) = \lim_{t\to 0^+} \frac{f(x+tv)-f(x)}{t} \leqslant f(z)-f(x), \quad \forall\, x,z \in \mathbb{R}^n.$$

Finalmente, como la función es diferenciable, sus derivadas direccionales se pueden expresar como un producto interno del gradiente con la dirección:

$$\frac{\partial f}{\partial v}(x) = \nabla f(x)^T (z-x), \quad v = z-x.$$

Reemplazando en la desigualdad anterior, se obtiene el resultado buscado:

$$\nabla f(x)^T (z-x) \leqslant f(z)-f(x), \quad \forall\, x,z \in \mathbb{R}^n.$$

□

Esta idea nos lleva a una definición de un concepto más fuerte que el de convexidad, y que de hecho se denomina convexidad fuerte:

**Definición 2.3.** Decimos que $f$ es fuertemente convexa si existe $m \in \mathbb{R}$ tal que

$$f(z) \geqslant f(x) + \nabla f(x)^T (z-x) + \frac{m}{2}\|z-x\|^2,\ \forall\, x,z \in \mathbb{R}^n.$$

Es decir, que no solo el gráfico de la función queda por encima de la recta tangente, sino que hay una función cuadrática que queda por debajo de la gráfica.

En el caso que $f$ sea de clase $C^2$, esto se puede escribir como $\nabla^2 f(x) \succcurlyeq mI$ para todo $x$. Retomaremos esta noción de convexidad fuerte en el siguiente capítulo.

Hablando de derivadas de segundo orden, veamos que también tienen información sobre la convexidad. En efecto, enunciaremos el siguiente resultado sin probarlo:

**Proposición 2.4.** Sea $f : \mathbb{R}^n \to \mathbb{R}$ de clase $C^2$, y tal que su matriz Hessiana $\nabla^2 f(x)$, formada por las derivadas parciales segundas, está definida en todo punto. Se cumple:

1. $\nabla^2 f(x) \succeq 0,\ \forall\, x \in \mathbb{R}^n \Rightarrow f$ convexa en $\mathbb{R}^n$.
2. $\nabla^2 f(x) \succ 0,\ \forall\, x \in \mathbb{R}^n \Rightarrow f$ estrictamente convexa en $\mathbb{R}^n$.

Donde la notación $\nabla^2 f(x) \succeq 0$ significa que la matriz es semi-definida positiva, y naturalmente $\nabla^2 f(x) \succ 0$ indica que es definida positiva.

**Ejercicio 2.5**
*Buscar una función que sea convexa pero no estrictamente convexa. ¿Cuál es el valor del mínimo? ¿en dónde se alcanza ese mínimo?*

Observar que si $f$ es una función univariada, recuperamos algo que ya sabíamos. La primera condición del resultado anterior, es: $f''(x) \geqslant 0$, $\forall\, x \in \mathbb{R}$, es decir que la función tiene concavidad no negativa en todo punto. El otro concepto que tenemos asociado a la convexidad es la concavidad:

**Definición 2.6.** Decimos que una función $f$ es cóncava si $-f$ es convexa.

Así como la convexidad son buenas noticias cuando tenemos un problema de minimización, la concavidad es buena cuando queremos maximizar una función.

**Comentarios sobre la convexidad.**

- Una función $f$ es convexa si y sólo si toda restricción en rectas es convexa. Es decir, supongamos que tenemos una función $f : \mathbb{R}^n \to \mathbb{R}$. Las restricciones a rectas son simplemente evaluar la función en una recta incluida en el dominio (o segmento de recta). Por ejemplo se puede escribir estas restricciones como $g(t) = f(x+tv)$, con $x, v \in \mathbb{R}^n$ y $t \in \mathbb{R}$. El resultado es una función univariada, para la cual naturalmente tenemos muchas herramientas para estudiar la convexidad[1].

- Una función $f$ es convexa si y sólo si el conjunto denominado epigrafo, y definido como

$$epi(f) = \{(x,t) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leqslant t\}$$

  es un conjunto convexo (definiremos conjunto convexo en breve).

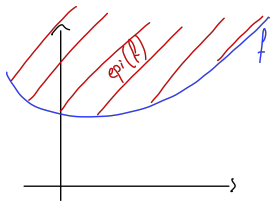



Figura 2.3: Epígrafo de una función convexa..

[1] S. Boyd, que es un referente en optimización y convexidad, cuenta con mucho humor cómo usar este hecho. Recomendable: https://youtu.be/ijD2KSXWDyo?si=Q7mx660-2tfF63QB

- Si $f$ es convexa, entonces los conjuntos de subnivel

$$C_\alpha = \{x \in dom f : f(x) \leqslant \alpha\}$$

son conjuntos convexos. El recíproco no es cierto (puede buscar un contraejemplo).

**Ejemplo 2.7**
*Algunas funciones convexas y cóncavas*

1. $f(x) = e^{ax}$ *es convexa,* $\forall\, a \in \mathbb{R}$.
2. $f(x) = x^a$ *es convexa si* $a \geqslant 1$, *o si* $a \leqslant 0$.
3. $f(x) = |x|^p$ *es convexa si* $p \geqslant 1$.
4. $f(x) = -\log(x)$ *es convexa en* $x > 0$ *(no está definida para* $x \leqslant 0$*).*
5. $f(x) = x\log(x)$ *es convexa en* $x > 0$.
6. *Cualquier norma en* $\mathbb{R}^n$ *es una función convexa.*
7. $f(x) = \max\{x_1, \ldots, x_n\}$ *es convexa.*
8. $f(X) = \log(\det(X))$ *es cóncava en el conjunto de matrices simétricas y definidas positivas:*
$$S^n_{++} = \{X \in \mathbb{R}^{n\times n} \,/\, X = X^T,\, X \succ 0\}.$$
9. *Una función lineal es cóncava* **y** *convexa.*

**Operaciones que preservan la convexidad** Dadas dos o más funciones convexas $f_1, \ldots, f_n$, veremos algunas operaciones que tienen como resultado otra función convexa. Esto es muy útil para estudiar la convexidad de una función objetivo, si logramos descomponerla en operaciones con funciones más simples.

1. Combinación lineal de funciones convexas, con coeficientes no negativos. Es decir: si $f_1, \ldots, f_n$ son funciones convexas, entonces
$$f(x) = \alpha_1 f_1(x) + \ldots + \alpha_n f_n(x), \quad \text{con} \alpha_i \geqslant 0,\ \forall\, i = 1, \ldots, n,$$
también es una función convexa.
2. Composición de una función convexa con una función lineal:
$$f \text{ convexa} \Rightarrow f(Ax+b) \text{ convexa}.$$

3. Máximo y supremo (punto a punto) de funciones convexas:

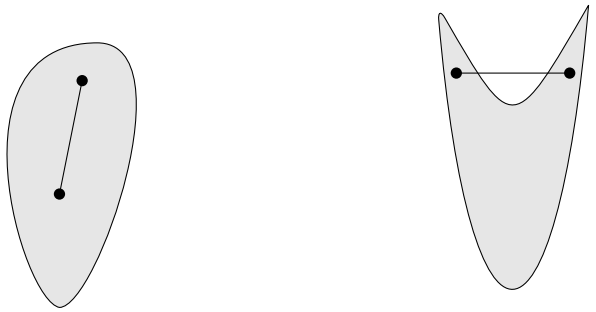
$$f_1, \ldots, f_n \text{ convexas} \Rightarrow f(x) = \text{máx}\{f_1(x), \ldots, f_n(x)\} \text{ convexa.}$$

Pasemos ahora a estudiar brevemente los **conjuntos convexos**.

**Definición 2.8.** Decimos que un conjunto $C \subset \mathbb{R}^n$ es convexo, si se cumple:

$$\forall\, x, y \in C,\ tx + (1-t)y \in C,\ \forall\, t \in [0,1].$$

Es decir: para cualquier par de puntos $x$ e $y$ de $C$, el segmento de recta que une dichos puntos está contenido en $C$.



convexo no convexo

Figura 2.4: Ejemplo de conjunto convexo y no convexo.

**Ejemplo 2.9**
*Ejemplos de conjuntos convexos.*

- *Conjunto afines: soluciones de un sistema lineal $Ax = b$. Esto incluye rectas, planos e hiperplanos.*
- *Bola de radio $R > 0$, centrada en el origen, con la norma usual (euclideana):*

$$B(\vec{0}, R) = \{x \in \mathbb{R}^n \ / \ \|x\|_2 \leqslant R\} \subset \mathbb{R}^n.$$

- *Bola de radio $R > 0$, con cualquier centro, y con cualquier norma.*

  *Es decir: toda combinación lineal de dos puntos de C, con pesos no negativos, es un punto de C.*
- *Cono semi-definido positivo. Es el conjunto formado por las matrices simétricas y semi-definidas positivas:*

$$S_+^n = \{X \in M_{n\times n}(\mathbb{R}) \ / \ X = X^T,\ X \succeq 0\}.$$

Estamos en condiciones de formular entonces **problemas de optimización convexa**, y lo haremos de dos formas.

La primera, que es la más cómoda y concisa, es la siguiente: decimos que un problema de optimización es convexo, si es de la forma:

$$\min_{x \in C} f(x), \quad \text{con } f \text{ función convexa, y } C \text{ conjunto convexo.} \tag{2.1}$$

La segunda forma, que es equivalente, es la formulación usada por ejemplo por S. Boyd, y que es la siguiente.

$$\begin{cases} \min & f(x), \quad \text{con} \quad f, f_1, \ldots, f_m \quad \text{convexas.} \\ f_i(x) \leqslant 0, & \forall\, i = 1, \ldots, m \\ a_i^T x = b_i, & \forall\, i = 1, \ldots, p \end{cases} \tag{2.2}$$

En esta segunda formulación, podemos identificar el conjunto $C$, como aquel conjunto que cumple todas las restricciones de la formulación (2.2).

Una de las propiedades más importantes de los problemas convexos, que adelantamos al inicio del capítulo, es la siguiente.

**Proposición 2.10.** En un problema convexo, todo mínimo local es también mínimo global.

Es decir que si de alguna forma logramos encontrar un mínimo local (como veremos en el capítulo siguiente), entonces habremos resuelto el problema.

Hay algunos problemas convexos con funciones objetivo y/o restricciones particulares. Por ejemplo los problemas lineales, denominados problemas de programación lineal (LP por sus siglas en inglés), los *Quadratic Constraint Quadratic Programming* (QCQP), los *Second-order cone programming* (SOCP), y los de *Semi-Definite Programming* (SDP). No entraremos en este curso en instancias particulares de estas clases, pero sí es importante tener en cuenta lo siguiente. Estas clases están anidadas. Por ejemplo, un problema de programación lineal es un caso particular de un QCQP, pero "más sencillo". Al ser casos particulares, cada tipo de problema tiene métodos específicos para resolverlos, y tienden a ser más eficientes que métodos generales. Por ejemplo, en los siguientes capítulos vamos a ver métodos que nos podríamos aplicar a cualquier problema convexo con relativo éxito, pero si lo que tenemos que resolver es un LP, será mucho más eficiente utilizar un método especializado en resolver problemas de ese tipo.

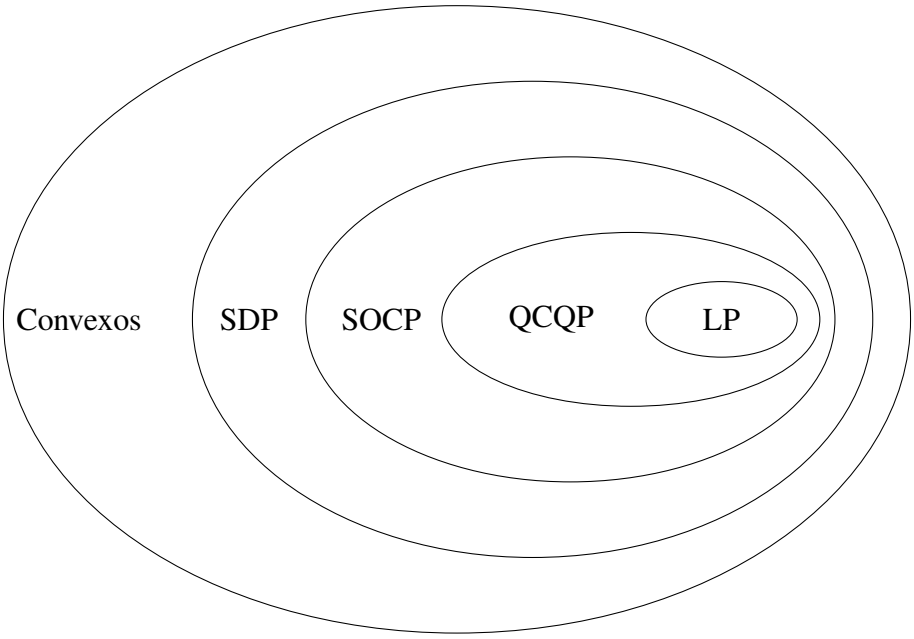



Figura 2.5: Jerarquía de algunas familias de problemas de optimización.

Capítulo 3

# Optimización sin restricciones

En este capítulo vamos a considerar el problema de minimizar una función sin restricciones. Vamos a trabajar con $\mathbb{R}^n$ como dominio, pero muchos resultados son generalizables a espacios más generales. El problema en esta parte entonces será

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

donde $f:\mathbb{R}^n\to\mathbb{R}$ es una función cualquiera, no necesariamente convexa. Vamos a ver propiedades del problema, de los mínimos, y métodos para resolverlo, o al menos encontrar mínimos locales.

Una buena referencia para el contenido de este capítulo es el libro de Dimitri Bertsekas [2].

## 3.1. Condiciones de optimalidad

La siguiente condición, formulada originalmente por Fermat en $\sim$1637, ya debería ser conocida por el público de este curso.

**Proposición 3.1** (Condición necesaria de optimalidad de primer orden)**.** Sea $f:\mathbb{R}^n\to\mathbb{R}$ diferenciable, y sea $x^*\in\mathbb{R}^n$ un punto donde $f$ alcanza un mínimo local. Entonces, necesariamente:

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

Si además $f\in C^2$, entonces $\nabla^2 f(x^\star)$ es semi-definida positiva.

*Demostración.* Tomemos una dirección $v\in\mathbb{R}^n$. Entonces la derivada direccional respecto a $v$ en el punto $x^\star$ es

$$\frac{\partial f}{\partial v}(x^\star)=\lim_{h\to 0^+}\frac{f(x^*+hv)-f(x^\star)}{h}=\langle\nabla f(x^\star),v\rangle,$$

donde usamos que $f$ es diferenciable para escribir la derivada direccional como producto interno del gradiente con la dirección. Ahora, como $f$ tiene un mínimo local en $x^\star$, tenemos que localmente $f(x^\star + hv) \geqslant f(x^\star)$, por lo que concluimos que $\langle \nabla f(x^\star), v\rangle \geqslant 0$.

Como $v$ era una dirección genérica, esto es cierto para toda dirección. En particular si tomamos $-v$, también obtenemos $\langle \nabla f(x^\star), -v\rangle \geqslant 0$, por lo que necesariamente debe ser $\langle \nabla f(x^\star), v\rangle = 0$.

Nuevamente, como la dirección $v$ es cualquiera, tenemos que $\langle \nabla f(x^\star), v\rangle = 0, \forall v \in \mathbb{R}^n$, por lo que el gradiente debe ser nulo: $\nabla f(x^\star) = 0$. □

Esta propiedad nos indica que si queremos buscar mínimos de una función diferenciable, parece buena idea buscar puntos donde se anula el gradiente. Estos puntos se denominan puntos críticos, o puntos estacionarios.

**Ejercicio 3.2**
*Mostrar con un ejemplo que la condición $\nabla f(x^\star) = 0$ no es suficiente para que $f$ tenga un mínimo local en $x^\star$.*

Si la función $f$ es convexa, la condición necesaria $\nabla f(x^\star) = 0$, es además suficiente para garantizar que se alcanza un mínimo local, que como vimos en el capítulo anterior, es además un mínimo global.

**Proposición 3.3.** Sea $f : \mathbb{R}^n \to \mathbb{R}$ convexa y diferenciable. Entonces, se cumple:

$$\nabla f(x^\star) = 0 \Leftrightarrow f \text{ alcanza mínimo } \textbf{global} \text{ en } x^\star.$$

*Demostración.* ($\Leftarrow$) Si $f$ tiene un mínimo global en $x^\star$, en particular tiene un mínimo local en $x^\star$. Por lo tanto, por la condición necesaria de primer orden, se debe cumplir: $\nabla f(x^*) = 0$.

($\Rightarrow$) Como $f$ es diferenciable y convexa, sabemos que se cumple:

$$f(x) \geqslant f(x^*) + \nabla f(x^*)(x - x^*), \quad \forall\, x \in \mathbb{R}^n.$$

Ahora, si $\nabla f(x^*) = 0$ tenemos directamente que $f(x) \geqslant f(x^*), \quad \forall\, x \in \mathbb{R}^n$.

□

La siguiente proposición da una condición suficiente para garantizar que un punto estacionario es mínimo local.

**Proposición 3.4** (Condición suficiente de optimalidad)**.** Sea $f : \mathbb{R}^n \to \mathbb{R}$ con derivada segunda continua, y sea $x^\star \in \mathbb{R}^n$ un punto estacionario de $f$. Es decir: $\nabla f(x^\star) = 0$. Si la matriz Hessiana $\nabla^2 f(x^\star)$ es definida positiva (estricta), entonces $f$ alcanza mínimo local en $x^\star$. A su vez, este mínimo local es estricto.

## 3.2. Algoritmos de optimización sin restricciones

Recapitulemos: el objetivo es encontrar el mínimo global de una función. Si la función es convexa, sabemos que corremos con ventaja. Para una función general, muchas veces solamente podremos aspirar a encontrar puntos críticos (que serán mínimos locales eventualmente, en vista de la condición de optimalidad que vimos recién).

También vimos en el primer capítulo que los problemas para los que se pueden hallar soluciones de forma cerrada son contados con las manos. Esgrimimos algunas ideas para minimizar funciones de una variable, y en este capítulo vamos a profundizar en la que denominamos "idea 3": usar la información de primer orden para construir un método iterativo. Asumiremos entonces en todo este capítulo que la función $f$ es diferenciable, aunque no necesariamente convexa.

Recordemos que un método iterativo construye una sucesión $x_k$ en el dominio $\mathbb{R}^n$, y queremos que $\lim_{k\to\infty} x_k = x^\star$, con $x^\star$ punto crítico.

Como queremos minimizar la función, tiene sentido pretender que el método "mejore" paso a paso. Es decir, que el valor funcional sea menor que en la iteración anterior. Denominemos esto como métodos de descenso:

**Definición 3.5.** Decimos que un método es de descenso, si se cumple:

$$f(x^{k+1}) < f(x^k), \quad \forall\, k \geqslant 0.$$

Esta definición es general, no necesariamente para métodos de primer orden. De hecho, por ejemplo el método de *golden-section* que vimos en la "idea 2", es un método de descenso. Pero en este capítulo queremos explotar la información de las derivadas primeras, ya que, como vimos, nos dice hacia dónde deberíamos movernos. Nos concentraremos entonces en métodos de la forma $x_{k+1} = x_k + \alpha_k d_k$, donde $\alpha_k \in \mathbb{R}_+$ es un paso, y $d_k \in \mathbb{R}^n$ es la dirección en la que nos movemos.

**Definición 3.6.** Decimos que el método es "de gradiente", cuando las direcciones utilizadas cumplen:

$$d_k \,/ \langle \nabla f(x_k), d_k \rangle < 0, \quad \forall\, k \geqslant 0.$$

Claramente esto se condice con lo mencionado en el primer capítulo: la condición $\langle \nabla f(x_k), d_k \rangle < 0$, equivale a que la derivada direccional de $f$ en la dirección $d_k$ sea negativa. Geométricamente, esto es que la dirección de movimiento $d_k$ forme un ángulo mayor a $\pi/2$ con el gradiente $\nabla f(x_k)$.

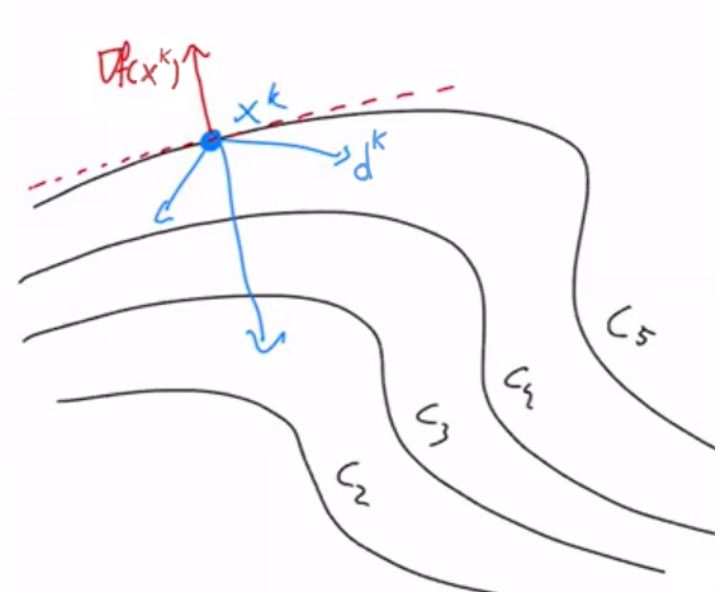



Figura 3.1: Direcciones $d_k$ que cumplen la condición de gradiente: $\langle \nabla f(x^k), d_k \rangle < 0$.

A una dirección $d_k$ que cumpla esta condición se la denomina dirección de descenso. Ahora, que la **dirección** sea de descenso no implica que el **método** sea de descenso, ya que si el valor funcional decrece en la iteración dependerá del paso $\alpha_k$ que tomemos. La condición de la dirección de "ser de descenso" es una condición local, por lo que si tomamos un paso muy grande, no tenemos garantías de lo que puede suceder. Eso sí, lo que tenemos es que si la dirección $d_k$ es de descenso, entonces se cumple que

$$\exists \delta > 0 \,/\, f(x_k + \alpha d_k) < f(x_k), \quad \forall \alpha \in (0, \delta).$$

Esto último no es otra cosa que decir que la función $f$ decrece localmente en la dirección $d_k$, que es consecuencia de que la derivada direccional en esa dirección es negativa. Entonces, si bien no todo paso $\alpha_k$ nos dará un decrecimiento en la función, sabemos que hay un intervalo de pasos que sí lo hará.

Tenemos entonces que el método descrito generará una sucesión según el siguiente pseudo-código:

**Algorithm 1** Métodos de descenso por gradiente.

**Require:** Inicialización $x_0 \in \mathbb{R}^n$

1: $k = 0$
2: **while** no se cumple condición de parada **do**
3: Elegir una dirección $d_k$, tal que: $\langle \nabla f(x_k), d_k \rangle < 0$.
4: Elegir un paso $\alpha_k > 0$
5: Actualizar la estimación: $x_{k+1} = x_k + \alpha_k d_k$.
6: $k = k + 1$
7: **end while**
8: **return** $x_k$

Esta descripción da lugar a varias preguntas. Por ejemplo ¿cómo elegimos la dirección $d_k$?, ¿cómo elegimos el paso $\alpha_k$?, y ¿qué condición de parada utilizamos?. Discutiremos estos elementos en lo que sigue.

### 3.2.1. Elección de la dirección de gradiente

Vamos a escribir genéricamente la dirección $d_k$ como $d_k = -D_k \nabla f(x_k)$, con $D_k \in S^n_{++}$ (matriz simétrica definida positiva). Las iteraciones resultan entonces

$$x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k),$$

y la elección de la dirección se reduce a elegir una matriz simétrica definida positiva $D_k$.

¿Por qué esto? Observemos que la condición $D_k \in S^n_{++}$ asegura que la dirección $d_k = -D_k \nabla f(x_k)$ cumple la condición para que el método sea de gradiente. En efecto, utilizando la expresión del producto interno usual en $\mathbb{R}^n$, tenemos que:

$$\langle \nabla f(x_k), d_k \rangle = \langle \nabla f(x_k), -D_k \nabla f(x_k) \rangle = -\left(\nabla f(x_k)\right)^T D_k \nabla f(x_k) < 0, \quad \forall\, \nabla f(x_k) \neq 0;$$

donde la última desigualdad se debe a que $D_k \in S^n_{++}$.

Veamos algunas elecciones particulares para la dirección.

**Caso I: Máximo descenso** Si tomamos la identidad $D_k = I$, obtenemos un método de gradiente que utiliza directamente la dirección opuesta al gradiente en cada paso, $d_k = -\nabla f(x_k)$. La iteración resulta entonces

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

Esta es la versión en varias variables del método que vimos en la "idea 3" del Capítulo 1, y se conoce como método de máximo descenso (o *steepest descent* en inglés). Recordar que

la derivada parcial se maximiza cuando tomamos una dirección colineal con el gradiente, por lo tanto la dirección de máximo descenso es la dirección opuesta al gradiente, que es la que estamos eligiendo en este caso.

Si bien $-\nabla f(x_k)$ es la dirección de máximo descenso local, y por lo tanto uno pensaría que es la mejor decisión que puede tomar, puede no ser la elección más conveniente globalmente. En particular, para algunas funciones, el método de máximo descenso puede tener convergencia muy lenta debido a un fenómeno de "zig-zageo".

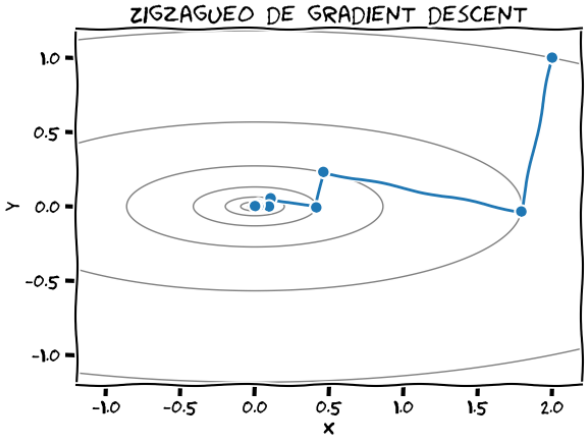



Figura 3.2: Función con curvas de nivel elípticas. El método de máximo descenso podría moverse en "zig-zag" y converger lentamente hacia el mínimo.

En la figura del ejemplo se toma la función $f(x,y) = x^2 + 10y^2$, y se corre el método de descenso por gradiente con dirección $-\nabla f$, y paso $\alpha$ óptimo. Es decir, el $\alpha$ que minimiza la función en la dirección elegida. En este caso, que la función es cuadrática, este valor de $\alpha$ se puede encontrar analíticamente.

Observar que, precisamente como el $\alpha$ es el que minimiza la función en esa dirección, el siguiente iterado estará donde la dirección es tangente a la curva de nivel. Como la siguiente dirección será colineal con el gradiente, entonces siempre los ángulos son rectos.

**Caso II: Newton** En el caso en que la función $f$ sea dos veces diferenciable, y la matriz Hessiana es invertible en todo punto, podemos definir el método que utiliza matriz de dirección $D_k = \left(\nabla^2 f(x_k)\right)^{-1}$. Es decir, el método resulta :

$$x_{k+1} = x_k - \alpha_k \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k). \tag{3.2}$$

Este método utiliza información de segundo orden (derivadas segundas), y es el equivalente al método de Newton univariado que vimos en la "idea 4" al comienzo de estas notas. Al igual

que en ese caso, la iteración surge de aproximar la función por su polinomio de Taylor de segundo orden, en este caso en $\mathbb{R}^n$:

$$f(x) \approx f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k).$$

**Ejercicio 3.7**
*Comprobar que el x que minimiza la aproximación anterior es exactamente el que determina la iteración (3.2), con* $\alpha_k = 1$.

Al igual que vimos antes, este método tiene convergencia cuadrática, por lo que suele converger en mucho menos iteraciones que el método de máximo descenso. Esta ganancia en la velocidad de convergencia viene de la mano de una desventaja: es necesario calcular las derivadas segundas para hallar la matriz Hessiana, y luego invertir dicha matriz (o calcular la solución de un sistema lineal de ecuaciones). Esto lo debemos hacer además en cada iteración, por lo que el método de Newton puede resultar muy costoso computacionalmente. Comparativamente, el método de máximo descenso requiere calcular un vector gradiente en cada iteración.

Como último comentario sobre este método, recordemos que para el caso univariado la condición para la convergencia cuadrática era que el iterado inicial esté "suficientemente cerca" del mínimo, y eso nos permitía acotar las derivadas. En este caso, observar que para estar en las condiciones en que venimos trabajando, la Hessiana debe ser definida positiva (su inversa es la que juega el papel de $D_k$). Bueno, tenemos entonces un comportamiento similar: en un mínimo local, la matriz Hessiana es definida positiva, y tenemos un abierto alrededor de él donde esta condición se sigue cumpliendo. Entonces, si comenzamos "suficientemente cerca" del mínimo, garantizamos que la matriz sea definida positiva, el método está bien definido, y tenemos convergencia cuadrática.

**Caso III: Escalado diagonal ("diagonal scaling")** Decíamos que el método de Newton tiene la desventaja del cálculo de todas las derivadas segundas (que son del orden de $n^2$), e invertir una matriz $n \times n$ (o resolver un sistema lineal). Una opción intermedia es aproximar la Hessiana por su diagonal. Utilizamos como matriz $D_k$:

$$D_k = \begin{pmatrix} d_1^k & 0 & 0 & \dots & 0 \\ 0 & d_2^k & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \dots & d_n^k \end{pmatrix}, \quad \text{con} \quad d_i^k = (f_{x_i,x_i}(x_k))^{-1} = \frac{1}{f_{x_i,x_i}(x_k)}.$$

Cuando la matriz Hessiana es diagonal, esto coincide exactamente con el método de Newton, y en otro caso es simplemente una aproximación. Tiene la ventaja de ser menos costoso computacionalmente, ya que la cantidad de derivadas que hay que calcular es de orden

lineal en $n$, y no hay que invertir una matriz, aprovechando la estructura diagonal. Al mismo tiempo, incorpora "algo" de información de segundo orden, por lo que puede ofrecer una ventaja frente al método de máximo descenso, reduciendo el fenómeno de zigzagueo.

### 3.2.2. Elección del paso

Pasamos al siguiente paso en el pseudo-código. Una vez que tenemos elegida la dirección $d_k$, debemos elegir el paso $\alpha_k$.

**Caso I: Búsqueda lineal ("line search")** Una opción ambiciosa (pero muchas veces realizable) es elegir el paso $\alpha_k$ que minimiza el valor de $f$ en la dirección de gradiente seleccionada $d^k$. Es decir:

$$\alpha_k \ / \ f(x_k + \alpha_k d_k) = \min_{\alpha \geqslant 0} f(x_k + \alpha d_k). \tag{3.3}$$

Observar que estamos ante un problema de optimización univariada, y podemos utilizar algunos de los métodos que discutimos en el Capítulo 1.

Para algunas funciones $f$ (por ejemplo cuadráticas en $\mathbb{R}^n$), es posible calcular este paso "óptimo" de forma cerrada.

**Ejercicio 3.8**
*Supongamos que estamos aplicando un método de gradiente a una función $f : \mathbb{R}^2 \to \mathbb{R}$, en la iteración $k$ estamos parados en un punto $x_k$, y ya elegimos la dirección $d_k$. Interprete geométricamente, en términos de las curvas de nivel, hacia qué punto $x_{k+1}$ nos moveremos si elegimos el paso $\alpha_k$ que minimiza el valor funcional, como en la ecuación (3.3).*

**Caso II: Búsqueda lineal acotada ("limited line search")** En lugar de hallar el mínimo de $g(\alpha) = f(x_k + \alpha d_k)$ en toda la semirrecta $\alpha > 0$, se busca el mínimo solamente en un segmento acotado: $\alpha \in [0, s]$, para cierto $s > 0$. Es decir:

$$\alpha_k \ / \ f(x_k + \alpha_k d_k) = \min_{\alpha \in [0,s]} f(x_k + \alpha d_k).$$

Esto permite utilizar otros sub-métodos para encontrar un buen valor de $\alpha$, como el método de *golden-section* que vimos antes.

**Caso III: Regla de Armijo** Una alternativa consiste en probar con un paso inicial $s > 0$, e ir reduciendo dicho paso hasta que el valor $f(x^k + sd^k)$ sea "suficientemente" menor al de la estimación anterior $f(x^k)$. La denominada regla de Armijo propone una forma de llevar a cabo esto. Si bien este es uno de los métodos más usados, la descripción es más compleja que las alternativa que discutimos, por lo que dejamos la lectura e implementación de esta regla como un ejercicio extra.

**Caso IV: Paso constante** La elección más sencilla: seleccionar el mismo paso para cada iteración: $\alpha_k = s > 0,\ \forall\ k \geqslant 0$. Por supuesto, si el paso seleccionado es "muy grande", no se puede asegurar convergencia, y si el paso es "muy chico", la convergencia puede ser muy lenta. Qué significa "muy grande" o "muy chico", dependerá de cada función. En particular, se puede ver que es posible seleccionar un paso fijo "satisfactorio" en función de algunos parámetros de la derivada de la función.

**Caso V: Paso decreciente** Consiste en utilizar una sucesión de pasos $\alpha_k$, que tienda a cero, pero no demasiado rápido. Formalmente, debe ser una sucesión $\{\alpha_k\}_{k\in\mathbb{N}}$, tal que:

$$\alpha_k > 0,\ \forall\ k, \quad \lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty.$$

Por ejemplo, se podría utilizar: $\alpha_k = \frac{1}{k}$, pero no $\alpha_k = \frac{1}{k^2}$.

Al igual que en la estrategia de paso constante, el uso de paso decreciente no garantiza que el algoritmo sea de descenso en cada iteración, aún cuando la dirección sea de gradiente.

### 3.2.3. Condición de parada

Recordemos que lo que estamos buscando son puntos que verifican la condición de optimalidad, es decir, puntos críticos, donde el gradiente se anula.

Por lo tanto, una condición de parada típica para los algoritmos de optimización, consiste en exigir que la norma del gradiente sea "suficientemente" pequeña:

$$\|\nabla f(x^k)\| \leqslant \varepsilon, \quad \text{para cierto } \varepsilon > 0,$$

o una versión normalizada (o relativa) con respecto a la norma inicial:

$$\|\nabla f(x^k)\| \leqslant \varepsilon \|\nabla f(x^0)\|.$$

En todo caso, siempre se debe agregar la condición de no superar una cierta cantidad de iteraciones máximas.

## 3.3. Análisis de convergencia

Hasta ahora hemos descrito un método (con sus variantes por la elección de la dirección, paso, etc), pero aún no sabemos nada sobre su funcionamiento.

Veamos brevemente el enunciado de un resultado de convergencia. Luego veremos una demostración parcial, asumiendo algunos elementos, y finalmente haremos un análisis más

detallado del orden de convergencia para un caso particular. En lo que sigue vamos a asumir que $f$ es diferenciable y con gradiente Lipschitz.

El siguiente resultado da garantías para el uso del método de descenso por gradiente.

**Proposición 3.9.** Si las direcciones $d_k$ son de descenso, y el paso $\alpha_k$ es elegido según la regla de Armijo, entonces todo límite de $x_k$ es un punto crítico.

El resultado es válido también para cuando el paso $\alpha_k$ es un valor de paso fijo (para ciertos valores de paso), y para cuando se utiliza un paso decreciente según los criterios descritos arriba (tendiente a cero pero de serie divergente).

Veamos una demostración parcial para el caso de paso fijo, y dirección $d_k = \nabla f(x_k)$. Es decir, tenemos el método

$$x_{k+1} = x_k - \alpha \nabla f(x_k),$$

con $\alpha$ fijo.

Consideremos el mapa $\psi(x) : \mathbb{R}^n \to \mathbb{R}^n$ definido como $\psi(x) = x - \alpha \nabla f(x)$. Observemos que la iteración anterior la podemos escribir en términos de este mapa como $x_{k+1} = \psi(x_k)$, y además que si $x^\star$ es un punto crítico, como $\nabla f(x^\star) = 0$, tenemos que $\psi(x^\star) = x^\star$. Es decir, $x^\star$ es un punto fijo del mapa $\psi$. Entonces, si asumimos que $\psi$ es una contracción[1], podemos usar un teorema de punto fijo para concluir la convergencia. Hagamos los cálculos explícitos:

$$\|x_{k+1} - x^\star\| = \|x_k - \alpha \nabla f(x_k) - x^\star\| = \|\psi(x_k) - \psi(x^\star)\| \leqslant \beta \|x_k - x^\star\|.$$

Repetimos este argumento de forma recurrente, $k$ veces, y llegamos a que

$$\|x_{k+1} - x^\star\| \leqslant \beta^k \|x_0 - x^\star\|.$$

Como $\beta < 1$, el término de la derecha tiende a cero (geométricamente). Si usamos una función de error $e(x) = \|x - x^\star\|$, tenemos que $\dfrac{e(x_{k+1})}{e(x_k)} \leqslant \beta$, por lo que tenemos convergencia lineal.

Recién asumimos que el mapa $\psi$ era una contracción, y quizás esto es asumir demasiado. En realidad se puede ver que con esta hipótesis en realidad estamos pidiendo que la función sea (localmente) fuertemente convexa.

**Digresión**

Para quienes tengan nociones de ecuaciones diferenciales y métodos numéricos, es interesante observar que el método de descenso por gradiente, con la dirección de máximo descenso y paso fijo, es una discretización tradicional (*forward Euler*) de la ecuación diferencial $\dot{x} = -\nabla f(x)$. Esto permite estudiar las propiedades de convergencia de ambos de forma relacionada. Por otro lado, otras discretizaciones (como *backward Euler*) dan lugar a otros métodos muy interesantes.

[1] Recordar que un mapa $T : \mathbb{R}^n \to \mathbb{R}^n$ es una contracción si existe $\beta \in [0,1)$ tal que $\|T(x) - T(y)\| \leqslant \beta \|x - y\|$ para todo $x, y \in \mathbb{R}^n$.

## 3.4. Análisis del caso cuadrático

Para analizar en detalle cómo depende la convergencia de algunos elementos de la función, vamos a considerar una función cuadrática $f : \mathbb{R}^n \to \mathbb{R}$, de la forma

$$f(x) = \frac{1}{2} x^T Q x, \quad \text{con } Q \text{ simétrica y definida positiva.}$$

Observar que esta función se minimiza en $x^\star = 0$.

Es fácil ver que el gradiente de $f$ es:

$$\nabla f(x) = \frac{1}{2}\left(Q^T x + Qx\right) = Qx.$$

Como $Q$ es invertible, el único punto estacionario es el origen.

También se puede ver que la matriz Hessiana de $f$ es $\nabla^2 f(x) = Q$, y como esta matriz es definida positiva, la función es fuertemente convexa.

Resumiendo, tenemos una función fuertemente convexa, que alcanza su mínimo (local y global) en su único punto crítico $x^\star = 0$.

La iteración del método de descenso por gradiente con $d_k = -\nabla f(x_k)$ entonces es:

$$x_{k+1} = x_k - \alpha \nabla f(x_k) = x_k - \alpha Q x_k = (I - \alpha Q)\, x_k.$$

Estudiemos la distancia de los iterados al óptimo, es decir $\|x_{k+1} - x^\star\|$. Como el mínimo se alcanza en el origen, esto resulta simplemente $\|x_{k+1}\|$. Estudiemos su cuadrado para realizar un cálculo:

$$\|x_{k+1}\|^2 = \langle x_{k+1}, x_{k+1} \rangle = x_k^T \left(I - \alpha Q\right)^2 x_k$$

Podemos acotar esta forma cuadrática por su valor propio más grande[2]:

$$\|x_{k+1}\|^2 = x_k^T \left(I - \alpha Q\right)^2 x_k \leqslant \lambda_{\text{máx}} \left(\left(I - \alpha Q\right)^2\right) \|x_k\|^2.$$

Observar que tenemos la distancia de $x_{k+1}$ al origen, acotada por la distancia del iterado anterior, $x_k$. Esto es exactamente lo que necesitamos para obtener algo como $\dfrac{e(x_{k+1})}{e(x_k)}$. Afinemos entonces las expresiones.

Si los valores propios de $Q$ son $\lambda_i$, se puede ver fácilmente que los valores propios de la matriz $(I - \alpha Q)^2$, son de la forma: $(1 - \alpha\lambda_i)^2$. En particular, si $m$ y $M$ son el menor y mayor valor propio de $Q$, respectivamente, se obtiene:

$$\lambda_{\text{máx}} \left(\left(I - \alpha Q\right)^2\right) = \max_i \left\{(1 - \alpha\lambda_i)^2\right\} = \text{máx}\left\{(1 - \alpha m)^2, (1 - \alpha M)^2\right\}.$$

[2]Recordar que para una matriz cuadrada $A$ tenemos: $z^T A z \leqslant \lambda_{\text{máx}}(A)\|z\|_2^2$, para todo $z \in \mathbb{R}^n$

Tenemos entonces que

$$\frac{\|x_{k+1}\|}{\|x_k\|} \leqslant \max\{|1-\alpha m|, |1-\alpha M|\}. \tag{3.4}$$

Haciendo cuentas, y graficando las funciones $|1-\alpha m|$ y $|1-\alpha M|$ en función de $\alpha$ (y el máximo de las dos), se puede ver que el paso $\alpha > 0$ que minimiza esta cota superior es aquel donde estas funciones coinciden (ver Figura 3.3). El paso óptimo resulta $\alpha_\star = \frac{2}{M+m}$. También es posible ver que para que la cota sea menor que uno, el valor de $\alpha$ debe estar en $\alpha \in (0, \frac{2}{M})$.

**Ejercicio 3.10** *1. Corroborar que el valor de $\alpha$ que minimiza el máximo de la ecuación (3.4) es $\alpha_\star = \frac{2}{M+m}$.*

*2. Corroborar que la cota es menor que uno para $\alpha \in (0, \frac{2}{M})$.*

Para este valor de $\alpha_\star$ que minimiza la cota, tenemos que

$$\frac{\|x_{k+1}\|}{\|x_k\|} \leqslant \frac{M-m}{M+m} < 1,$$

lo que prueba la convergencia lineal, con parámetro $\frac{M-m}{M+m}$.

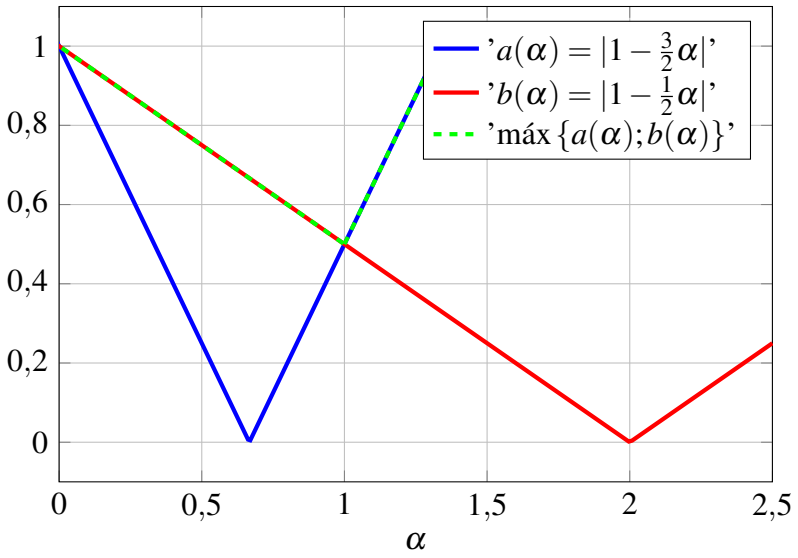



Figura 3.3: Ejemplo con $m = \frac{1}{2}$, $M = \frac{3}{2}$. La cota superior se minimiza en: $\alpha_\star = \frac{2}{M+m} = 1$.

El cociente de los valores propios $\frac{M}{m}$ es el denominado número de condición de la matriz $Q$, al que vamos a denominar $\kappa$:

$$\kappa := \frac{M}{m} \geqslant 1.$$

A partir del resultado anterior, se puede ver que un número de condición pequeño: $\frac{M}{m} \simeq 1$, se traduce en una convergencia más rápida, pues la cota es cercana a cero: $\frac{M-m}{M+m} \simeq \frac{0}{M+m} = 0$. Por otro lado, si el número de condición es grande: $\frac{M}{m} \gg 1$, es de esperarse que la convergencia sea más lenta, dado que la cota es cercana a uno: $\frac{M-m}{M+m} \simeq \frac{M}{M} = 1$.

El número de condición se relaciona con la elipticidad de las curvas de nivel de la función cuadrática. En particular: para un número de condición cercano a uno, las curvas de nivel serán más parecidas a una circunferencia; mientras que para número de condición muy grande, las curvas de nivel serán más elípticas, y más "achatadas" cuanto más grande sea el número de condición. Esto último enlentece la convergencia del método de máximo descenso, cosa que ya mencionamos con el fenómeno de zigzagueo (Figura 3.4).

**Digresión (sobre la convexidad y número de condición)**

Consideremos una función $f : \mathbb{R}^n \to \mathbb{R}$ convexa y diferenciable. Habíamos visto que el gráfico queda soportado por la tangente en cualquier punto:

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

Resulta que si la función es *fuertemente* convexa, entonces la función no solo queda por encima de la recta tangente, sino que además está por encima de una cuadrática:

$$f(z) \geqslant f(x) + \langle \nabla f(x), z - x \rangle + \frac{m}{2}\|z - x\|^2, \quad \forall x, z \in \mathbb{R}^n.$$

**Ejercicio 3.11**
*Considere una función convexa (pero que no será fuertemente convexa), y encuentre un punto del dominio donde sea imposible ubicar una cuadrática soportando el gráfico como en la desigualdad anterior.*

Otro concepto importante, que bajo ciertas condiciones garantiza la convergencia de los métodos de descenso por gradiente a un punto crítico, es que el gradiente de $f$ sea Lipschitz[3]. Es decir, que

$$\nabla f(x) - \nabla f(y)\| \| \leqslant M\|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

Utilizando el desarrollo de Taylor, se puede ver que si el gradiente es Lipschitz de constante $M$, entonces

$$f(z) \leqslant f(x) + \langle \nabla f(x), z - x \rangle + \frac{M}{2}\|z - x\|^2, \quad \forall x, z \in \mathbb{R}^n.$$

Es decir, que si tenemos una función fuertemente convexa con gradiente Lipschitz, entonces en todo punto $f$ está atrapada entre dos cuadráticas.

[3]Recordemos que una función $g : \mathbb{R}^n \to \mathbb{R}^n$ es Lipschitz de constante $L$ si $\|g(x) - g(y)\| \leqslant L\|x - y\|$, para todo $x, y \in \mathbb{R}^n$.

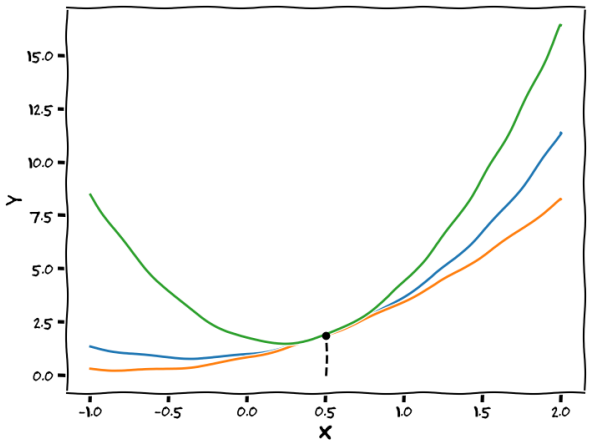



Figura 3.4: Función convexa $f(x) = e^x + x^2$, y las cuadráticas por abajo y por arriba, en el punto $x = 0{,}5$.

Se puede ver también que estas constantes $m$ y $M$ son los valores propios más chico y más grande de $\nabla^2 f$ respectivamente.

## 3.5. Tasa de convergencia y métodos con momentum

Hicimos el análisis detallado para el caso cuadrático, pero en general vale que si $f$ es fuertemente convexa y de gradiente Lipschitz, entonces vale que si tomamos como paso fijo $\alpha = \frac{2}{m+M}$, se tiene

$$\|x_k - x^\star\| \leqslant \left(1 - \frac{2}{\kappa + 1}\right)^k \|x_0 - x^\star\|, \tag{3.5}$$

donde reemplazamos $\frac{M-m}{M+m}$ por la expresión equivalente en términos del número de condición $\left(1 - \frac{2}{\kappa+1}\right)$.

Esta tasa de convergencia es la alcanzada con un paso fijo, por lo que es sensato preguntarse si se puede hacer algo mejor. Lo cierto es que sí es posible mejorar la performance del método con elecciones como la regla de Armijo, pero esto **no varía esencialmente la tasa de convergencia**.

El método de Newton sí ofrece una mejor convergencia (cuadrática de hecho, que es bastante más que mejorar la constante), pero utiliza información de segundo orden.

Entonces, la pregunta natural es, utilizando solamente información de primer orden (i.e., el gradiente), ¿se puede mejorar la tasa de convergencia?

Volvamos a dos elementos que ya comentamos: por un lado, el método de descenso por

gradiente, en particular con dirección opuesta al gradiente, sufre de un fenómeno de zigzagueo. Por otro lado, este método se puede interpretar como la discretización del flujo gradiente. Es decir, si en la siguiente ecuación diferencial

$$\dot{x} = -\nabla f(x)$$

aproximamos la derivada por el cociente incremental para un paso $h$ $\left(\dot{x} \approx \frac{x_{k+1}-x_k}{h}\right)$, resulta

$$\frac{x_{k+1}-x_k}{h} = -\nabla f(x_k) \quad \Longrightarrow \quad x_{k+1} = x_k - h\nabla f(x_k).$$

Ahora bien, si tenemos un fenómeno dinámico que presenta un zigzagueo indeseado, nos podemos preguntar, ¿qué haría un físico? Pues agregar fricción.

Tomemos la ecuación diferencial entonces, y agreguemos un término de rozamiento. Consideremos entonces

$$\ddot{x} = -b\dot{x} - \nabla f(x).$$

Si discretizamos esta ecuación diferencial como antes, obtenemos

$$x_{k+1} = x_k - \alpha\nabla f(x_k) + \beta\,(x_k - x_{k-1})\,. \tag{3.6}$$

Este método, con $\alpha$ y $\beta$ constantes, se denomina *Heavy-ball method*[4], y es debido a Boris Polyak, en la década de los 60*s* en la URSS.

Observar que la iteración (3.6) tiene una parte idéntica al descenso por gradiente tradicional, seguido de un término de "memoria" o "momentum". En particular, tiene una dependencia no sólo del término en la iteración $k$, sino también del anterior, $x_{k-1}$. Esta iteración también se puede escribir de la siguiente forma equivalente:

$$\begin{cases} x_{k+1} = x_k + \alpha p_k \\ p_k = -\nabla f(x_k) + \tilde{\beta} p_{k-1} \end{cases}$$

En esta forma, tenemos una iteración "tradicional" en la primera línea, es decir, nos movemos de $x_k$ a $x_{k+1}$ tomando una dirección $p_k$ y moviéndonos un paso $\alpha$. La diferencia se hace evidente con la segunda línea: la dirección elegida en cada paso no es exactamente la opuesta al gradiente, sino que tiene una "corrección", o una inercia, que se ve en el sumando de la dirección anterior $p_{k-1}$.

Veamos qué propiedades de convergencia tiene este nuevo método que acabamos de describir.

[4]Cuando este método se aplica a una función cuadrática, se obtiene el método de Chebyshev.

### 3.5.1. Análisis de convergencia del Heavy-ball method

En realidad vamos a dar una idea breve, sin detalles técnicos.

Consideremos el vector $w_k \in \mathbb{R}^{2n}$, que construimos como

$$w_k = \begin{bmatrix} x_{k+1} - x^\star \\ x_k - x^\star \end{bmatrix}$$

y hagamos un análisis local, cerca de $x^\star$. En efecto, haciendo un desarrollo de Taylor alrededor de $x^\star$, es fácil ver que

$$w_{k+1} \approx \begin{bmatrix} -\alpha\nabla^2 f(x^\star) + (1+\beta)I & \beta I \\ I & 0 \end{bmatrix} w_k \,,$$

donde la matriz está compuesta por cuatro bloques $n \times n$. Los dos bloques de abajo (la identidad y la matriz nula) son los que "mueven" el término $x_{k+1} - x^\star$ hacia abajo. Los dos bloques de arriba son los que codifican (Taylor mediante) la iteración (3.6).

Con un argumento simple de álgebra lineal tenemos que los valores propios de la matriz

$$\begin{bmatrix} -\alpha\nabla^2 f(x^\star) + (1+\beta)I & \beta I \\ I & 0 \end{bmatrix}$$

son los mismos que la matriz

$$\begin{bmatrix} -\alpha\Lambda + (1+\beta)I & \beta I \\ I & 0 \end{bmatrix}$$

donde $\Lambda$ es una matriz diagonal con los $n$ valores propios de $\nabla^2 f(x^\star)$.

Luego podemos elegir los valores de $\alpha$ y $\beta$ explícitamente para minimizar el valor propio más grande de esta matriz, que resultan:

$$\alpha = \frac{4}{M} \frac{1}{\left(1 + \frac{1}{\sqrt{\kappa}}\right)^2} \quad , \qquad \beta = \left(1 - \frac{2}{1+\sqrt{\kappa}}\right)^2 .$$

Los valores de los parámetros no son importantes (sobre todo porque si quisiéramos usarlos en una aplicación real, necesitaríamos conocer los valores propios de la Hessiana en el punto crítico). Lo importante es que con estos valores se obtiene una tasa de convergencia de $\left(1 - \frac{2}{1+\sqrt{\kappa}}\right)$. Comparar con la tasa de convergencia del descenso por gradiente usual, en (3.5). Tenemos:

$$\begin{aligned} \text{Descenso por gradiente "simple"} \qquad & \text{tasa } \left(1 - \frac{2}{1+\kappa}\right) = \left(\frac{\kappa - 1}{\kappa + 1}\right), \\ \text{Heavy-ball method} \qquad & \text{tasa } \left(1 - \frac{2}{1+\sqrt{\kappa}}\right) = \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right). \end{aligned}$$

Por ejemplo, para llegar a reducir el error inicial $\|x_0 - x^\star\|$ en un factor de $\varepsilon$, es decir, una iteración $k$ tal que $\|x_k - x^\star\| \leqslant \varepsilon \|x_o - x^\star\|$, necesitamos:

$$\begin{aligned} k &\geqslant \tfrac{\kappa}{2}\log(1/\varepsilon) \quad \text{iteraciones con descenso por gradiente,} \\ k &\geqslant \tfrac{\sqrt{\kappa}}{2}\log(1/\varepsilon) \quad \text{iteraciones con Heavy-ball.} \end{aligned}$$

Si tenemos un número de condición $\kappa = 100$, con el Heavy-ball method necesitamos 10 veces menos iteraciones.

Unos pocos años antes, en USA[5], se estaban desarrollando métodos similares (método del gradiente conjugado), que tenían expresiones más explícitas para los parámetros, pero análisis de convergencia que servían bien solamente para el caso de funciones cuadráticas.

Recapitulemos. Teníamos el método de descenso por gradiente, con una convergencia lineal, y nos preguntamos si es posible hacer algo mejor, usando solamente información de primer orden. El método de Polyak contesta afirmativamente esa pregunta, pero nos podríamos preguntar de nuevo si se puede hacer algo mejor.

En 1983 Yurii Nesterov, estudiante de Polyak, muestra dos resultados que fueron sorprendentes para la época. Por un lado, da un método similar el de Polyak, pero con constantes mucho más manejables, y por otro lado prueba que no se puede alcanzar mejores tasas de convergencia que la de su método, utilizando información de primer orden. Es decir, el método de Nesterov es óptimo en ese sentido.

Este método es:

$$\begin{cases} x_{k+1} = x_k + \alpha_k p_k \\ p_k = -\nabla f\left(x_k + \beta_k (x_k - x_{k-1})\right) + \beta_k p_{k-1} \end{cases}$$

Observar que es muy similar al método de Polyak, pero con un "extragradient step": el gradiente no está evaluado en $x_k$ sino en una versión corregida con inercia.

Para el parámetro $\alpha$ se puede usar $\alpha = \frac{1}{M}$, y para $\beta$ hay varias opciones. Una muy usada es $\beta = \frac{k-1}{k+2}$, y también hay una expresión en el paper de Beck y Teoulle [1].

Como antes, también podemos expresar este método de una forma alternativa:

$$\begin{cases} y_k = x_k + \beta_k (x_k - x_{k-1}) \\ x_{k+1} = x_k - \alpha_k \nabla f(y_k). \end{cases}$$

En términos de implementación, observar que es simplemente agregar "una línea" de código al método de descenso por gradiente tradicional, por lo que, si tenemos una implementación para nuestro problema del descenso por gradiente, muchas veces vale la pena intentar un método acelerado como el de Nesterov, porque implica un esfuerzo relativamente bajo.

[5] Recordemos que estamos hablando de las décadas de los 50*s* y 60*s*, plena Guerra Fría.

## 3.6. Método del Gradiente Conjugado

Mencionamos más arriba el método del gradiente conjugado. Este es un método debido a Hestenes y Stiefel, en 1952, que veremos brevemente en lo que sigue. Algunas buenas referencias para este tema son [12, 10].

Consideremos una matriz $A \in S^n_{++}$, es decir, $A \succ 0$ definida positiva, y simétrica. El Gradiente Conjugado se puede pensar como un método para resolver $Ax = b$, o para minimizar $f(x) = \frac{1}{2}x^T Ax - b^T x$. En particular, observemos que si $A \succ 0$, entonces $f$ es convexa y el mínimo se da en $\nabla f(x) = 0 = Ax - b$.

Este es un método que suele ser útil para matrices grandes y esparsas, ya que lo único que necesitamos es saber cómo calcular el producto de $A$ con un vector $v$ ($Av$), y no es necesario conocer $A$ explícitamente.

Para el caso de $Ax = b$, o equivalentemente para minimizar la función cuadrática $f$ de arriba, el método tiene garantizado encontrar la solución en $n$ iteraciones. Muchas veces no se necesita correr el método hasta el final, y con algunas iteraciones ya se tiene una buena aproximación de la solución.

Es también un método iterativo, así que tendremos iterados $x_k$ como antes. Llamaremos $r = Ax - b$, y en general $r_k = Ax_k - b$, y lo denominamos residuo. Observar que $r_k = \nabla f(x_k)$.

Hay un enfoque basado en los subespacios de Krylov, que son los espacios generados así: $K_k = span\{b, Ab, \ldots, A^{k-1}b\}$. Entonces se define el iterado $x_k$ como el que minimiza $f$ en el subespacio $K_k$:

$$x_k = \arg\min_{x \in K_k} f(x)$$

Se puede ver (de forma no trivial) que existe una recurrencia

$$x_{k+1} = x_k + \alpha_k r_k + \beta_k (x_k - x_{k-1})$$

para ciertos valores $\alpha_k, \beta_k$.

Pero veamos un enfoque más parecido a lo que veníamos haciendo en este capítulo. Vamos a movernos en direcciones $d_k$, es decir, con iteraciones de la forma $x_{k+1} = x_k + \alpha_k d_k$ como siempre. Las direcciones van a estar relacionadas con el opuesto del gradiente, pero no serán exactamente eso. La propiedad que van a cumplir las direcciones es que son **conjugadas**, es decir, ortogonales respecto a la matriz $A$. Esto es, que el producto interno asociado a $A$ entre dos direcciones, sea nulo:

$$\langle d_{k-1}, d_k \rangle_A = d_{k-1}^T A d_k = 0.$$

Veamos cómo calcular la dirección $d_k$, que será una combinación del residuo (o gradiente) y la dirección anterior:

$$d_k = -r_k + \beta_k d_{k-1}.$$

Imponiendo que las direcciones sean conjugadas ($d_{k-1}^T A d_k = 0$), se puede despejar el valor de $\beta_k$ que resulta

$$\beta_k = \frac{r_k^T A d_{k-1}}{d_{k-1}^T A d_{k-1}} = \frac{\langle r_k, d_{k-1} \rangle_A}{\langle d_{k-1}, d_{k-1} \rangle_A}.$$

Ahora que tenemos definida la dirección, tenemos que escoger el paso $\alpha_k$, igual que antes. En este caso, que la función es cuadrática, podemos encontrar el paso que minimiza la función en la dirección elegida, en forma cerrada:

$$\alpha_k = \frac{r_k^T d_k}{d_k^T A d_k} = \frac{\langle r_k, d_k \rangle}{\langle d_k, d_k \rangle_A}.$$

Es posible actualizar los parámetros de forma más astuta, llevando a la implementación del pseudo-código en el Algoritmo 2.

**Algorithm 2** Método del gradiente conjugado

**Require:** $A$ matriz simétrica definida positiva, $b \in \mathbb{R}^n$, $x_0$

$r_0 \leftarrow A x_0 - b$
$d_0 \leftarrow -r_0$
$k \leftarrow 0$
**while** $r_k \neq 0$ **do**
  $\alpha_k \leftarrow \dfrac{r_k^T r_k}{d_k^T A d_k}$
  $x_{k+1} \leftarrow x_k + \alpha_k d_k$
  $r_{k+1} \leftarrow r_k + \alpha_k A d_k$
  $\beta_{k+1} \leftarrow \dfrac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$
  $d_{k+1} \leftarrow -r_{k+1} + \beta_{k+1} d_k$
  $k \leftarrow k + 1$
**end while**

Se puede ver que la tasa de convergencia es $\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)$, pero el método puede ser más rápido dependiendo de la estructura de los valores propios de $A$. En particular, si los valores propios se encuentran en "clusters".

## 3.7. Coordinate Descent

En esta última sección de este capítulo, más que un método vamos a ver una estrategia para minimizar funciones, por medio de resolver minimizaciones más sencillas (de menor dimensionalidad). Siempre con métodos iterativos.

Por ejemplo, si tenemos una función de dos variables, digamos $f(x,y)$, y comenzamos desde $(x_1,y_1)$, podemos minimizar respecto a la variable $x$, dejando la $y$ fija en el valor $y=y_1$, obteniendo un valor $x_2$, y luego minimizar respecto de $y$ dejando fija $x=x_2$. Repetimos este procedimiento alternadamente. ¿Llegaremos a un punto crítico al menos?

Sea $f: X \to \mathbb{R}$ en principio convexa y diferenciable, donde el conjunto $X \subset \mathbb{R}^n$ se puede descomponer como $X = X_1 \times X_2 \times \cdots \times X_m$, donde $X_i \subset \mathbb{R}^{n_i}$. Podría ser eventualmente $n_i = 1$ para todo $i$, y en ese caso $m = n$. Para fijar ideas, podemos pensar que $X = \mathbb{R}^n$, y que lo descomponemos en cada una de sus componentes unidimensionales (es decir, el producto de $n$ veces $\mathbb{R}$).

A un punto $x \in X$, lo escribimos en función de sus coordenadas (o bloques de coordenadas) pertenecientes a cada $X_i$, es decir

$$x = \left(x_{(1)}, x_{(2)}, \ldots, x_{(m)}\right).$$

En el caso que decíamos, donde $X_1 = \mathbb{R}$, esta descomposición es simplemente las coordenadas de $x$.

En lo que resta de la sección vamos a escribir $x^k$ para el iterado $k$, para no saturar los subíndices.

La idea es, a partir del iterado $x^k = \left(x^k_{(1)}, x^k_{(2)}, \ldots, x^k_{(m)}\right)$, generar el iterado $x^{k+1}$ de forma secuencial, de a una coordenada (o en realidad, bloque de coordenadas) por vez.

$$x^{k+1}_{(i)} = \arg\min_{z \in X_{(i)}} f\left(x^{k+1}_{(1)}, x^{k+1}_{(2)}, \ldots, x^{k+1}_{(i-1)}, z, x^k_{(i+1)}, \ldots, x^k_{(m)}\right).$$

Es decir, vamos actualizando coordenada a coordenada (de nuevo, o bloques de coordenadas), y a medida que calculamos un valor para una coordenada, ya la utilizamos cuando hagamos la optimización de la siguiente[6]. Esto es la base del método de descenso por coordenadas (CDM).

Este método puede ser atractivo cuando estos problemas de minimización en una coordenada (o bloque de coordenadas) son sensiblemente más fáciles que el problema original. Ahora, ¿este procedimiento converge?

El siguiente resultado, que se puede consultar en [3][6.5.1], asegura que bajo ciertas condiciones, este mecanismo nos asegura convergencia a un mínimo:

**Teorema 3.12.** Sea $f: \mathbb{R}^n \to \mathbb{R}$ convexa y de clase $C^1$. Sea $X = X_1 \times X_2 \times \cdots \times X_m$ con los $X_i$ cerrados y convexos. Si la función $f\left(x_{(1)}, x_{(2)}, \ldots, x_{(i-1)}, z, x_{(i+1)}, \ldots, x_{(m)}\right)$, mirada como función de $z$, tiene un único mínimo en $X_i$, para todo $i$, entonces la sucesión generada por CDM cumple que todo punto límite minimiza $f$ sobre $X$.

[6] Para quienes lo vieron en un curso de Métodos Numéricos, es equivalente un método de Gauss-Seidel, en contraposición al método de Jacobi.

**Ejercicio 3.13**
*Buscar una función $f : \mathbb{R}^2 \to \mathbb{R}$, no diferenciable, para la cual este procedimiento quede estancado en un punto que no minimiza $f$. Sugerencia: solamente dibujar las curvas de nivel.*

Este resultado se puede generalizar a una función $f$ no necesariamente convexa (ver [3][6.5.2]), pidiendo a estas funciones miradas como función de $z$ que tengan un único mínimo y que sean monótonas entre $x_i^k$ y $x_i^{k+1}$. En este caso se asegura convergencia a un mínimo que satisface la condición de optimalidad que veremos luego. Como ya no pedimos que la función sea convexa, no se puede asegurar convergencia a un mínimo.

Los métodos de tipo descenso por coordenadas, o por bloques de coordenadas, son muy utilizados cuando la minimización respecto a cada bloque es muy poco costosa computacionalmente. Muchas veces se pueden conseguir mejores resultados aleatorizando el orden en que se recorren las coordenadas. Hay varios ejemplos de el uso de estos métodos en la literatura, para problemas particulares. Por ejemplo el problema de mínimos cuadrados con regularización $\ell_1$, denominado LASSO, se puede resolver de forma muy eficiente con estos métodos [5]. Relacionado con lo anterior, en [6] utilizan descenso por bloques de coordenadas para resolver un problema matricial, de inferencia de la matriz inversa de covarianza.

**Resumen del capítulo**

En este capítulo estudiamos el problema de optimización sin restricciones, es decir,

$$\min_{x\in\mathbb{R}^n} f(x).$$

Vimos condiciones de optimalidad para funciones diferenciables, y vimos que la convexidad es sumamente útil: por un lado permite demostrar la convergencia de algoritmos en muchos casos, y por otro lado ser mínimo local implica ser mínimo global.

Luego estudiamos el método de descenso por gradiente, que tiene iteraciones del tipo $x_{k+1} = x_k + \alpha_k d_k$. Tomando la dirección $d_k = \nabla f(x_k)$ el método se denomina *steepest descent*, y vimos que puede sufrir problemas de zigzagueo. Sin embargo, también vimos que el método tiene convergencia lineal a un punto crítico, con tasa $\frac{\kappa-1}{\kappa+1}$.

Para solucionar parcialmente el zigzagueo, vimos los métodos acelerados, o con momentum, que utilizan información de los iterados anteriores. El apellido a recordar aquí es Nesterov. Vimos que se lograba una tasa óptima (para métodos de primer orden), de $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$.

Luego vimos el método de gradiente conjugado, que tiene muy buenas propiedades para funciones cuadráticas (converge en $n$ iteraciones), pero que de todas maneras se puede utilizar en otros problemas. Se denomina "non-linear conjugate gradient", y si bien pierde un poco la belleza del caso cuadrático, puede funcionar muy bien.

Terminamos el capítulo con métodos por descenso por coordenadas, que nos permiten encontrar puntos críticos de funciones, bajo ciertas hipótesis, realizando minimizaciones alternadas por (bloques de) coordenadas.

Vale la pena incluir en este resumen final al método de Nelder-Mead, que mencionamos al pasar en el Capítulo 1, y que busca mínimos locales de funciones sin utilizar información de las derivadas.

# Capítulo 4

# Optimización con restricciones

En este capítulo vamos a agregar restricciones al problema, y en lugar de buscar mínimos en todo el espacio ambiente, lo haremos en un subconjunto $\mathscr{X} \subset \mathbb{R}^n$, que vamos a suponer convexo y cerrado. Como escribimos antes, el problema a resolver es

$$\min_{x \in \mathscr{X}} f(x) \quad ,$$

recordando además que muchas veces lo que nos interesa es el punto donde se alcanza el mínimo, más que el valor del mismo. Vamos a suponer también que $f$ es diferenciable.

A un punto que verifique las restricciones ($x \in \mathscr{X}$) lo llamaremos *factible* (o *feasible* en inglés).

## 4.1. Condiciones de optimalidad para minimización con restricciones

Comencemos por generalizar las condiciones de optimalidad que vimos en el capítulo anterior (que el gradiente se anule). En este caso, tenemos lo siguiente

**Proposición 4.1.** Sea $f : \mathscr{X} \to \mathbb{R}$ diferenciable, donde $\mathscr{X}$ es un subconjunto convexo de $\mathbb{R}^n$.

a) Si $f$ tiene un mínimo local en $x^\star$, entonces $\langle \nabla f(x^\star), x - x^\star \rangle \geqslant 0$ para todo $x \in \mathscr{X}$.

b) Si además $f$ es convexa, entonces la condición anterior es suficiente.

Para interpretar esto, analicemos primero el término $x - x^\star$. Observemos que, como $\mathscr{X}$ es convexo, si $x$ es un punto del conjunto, $x - x^\star$ es una dirección tal que, partiendo de $x^\star$ y

moviéndonos hacia $x$, permanecemos dentro del conjunto. Más aún, cualquier dirección en la que nos queramos mover desde $x^\star$, la podemos describir como $x - x^\star$ para algún $x \in \mathscr{X}$. Es decir, $x - x^\star$, con $x$ variando en todo $\mathscr{X}$, son las **direcciones factibles** desde $x^\star$.

La interpretación de la primera condición entonces es clara: el producto interno $\langle \nabla f(x^*), x - x^\star \rangle$ no es otra cosa que la derivada direccional de $f$ en la dirección $(x - x^\star)$, y por lo tanto la condición se puede leer así: en todas las direcciones que uno se pueda mover desde $x^\star$, la función $f$ *crece*, y por lo tanto $x^\star$ tiene que ser un mínimo local.

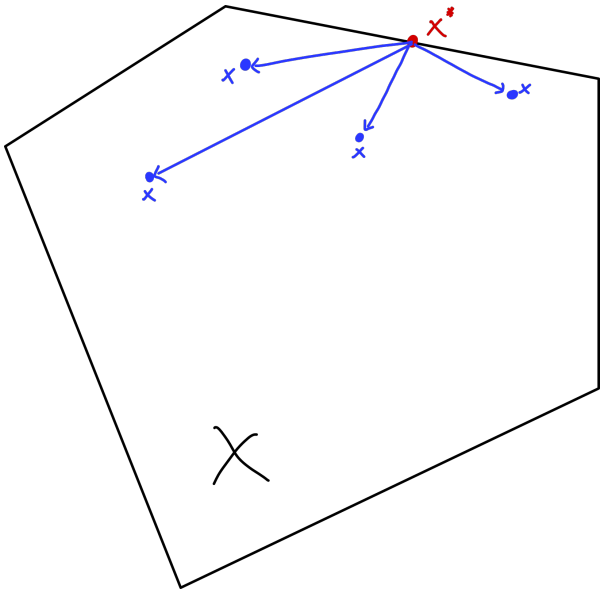
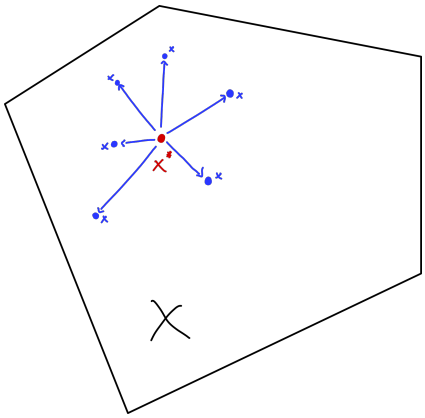


Figura 4.1: Punto $x^\star \in \mathscr{X}$, y algunas direcciones factibles $x - x^\star$, para varios puntos $x \in \mathscr{X}$.

Observemos también que si $x^\star$ es un punto interior a $\mathscr{X}$, entonces las direcciones factibles son todas. En este caso, el producto interno del gradiente con todas las direcciones debe ser no negativa, y por lo tanto el gradiente debe ser cero. Es decir, en este caso recuperamos la condición de optimalidad que teníamos en el caso sin restricciones.



Incorporando la geometría de las curvas de nivel de $f$, volvemos a tener una interpretación natural. Veamos en la Figura 4.2 las curvas de nivel de una función diferenciable, creciendo desde dentro hacia afuera, y el punto $x^\star$ en el conjunto $\mathscr{X}$, donde se alcanza el mínimo. Es decir, el punto de $\mathscr{X}$ que toca la curva de nivel más baja. Dibujamos en azul las direcciones

factibles $x - x^\star$ de la condición de optimalidad, y en rojo el gradiente de $f$ (perpendicular a la curva de nivel). La condición del producto interno es que el ángulo que forma el gradiente con todas las posibles direcciones factibles, es menor o igual a $\pi/2$.

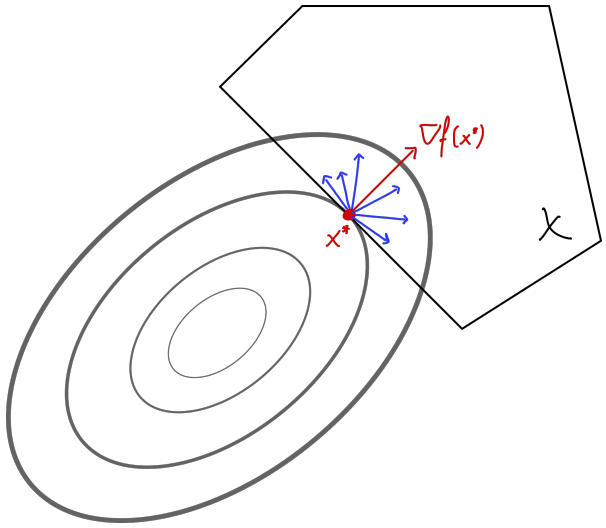



Figura 4.2: Curvas de nivel de $f$ diferenciable, mínimo de $f$ restricto a $\mathscr{X}$ en el punto $x^\star$. El gradiente $\nabla f(x^\star)$ forma ángulos agudos con todas las direcciones factibles (en azul).

**Ejercicio 4.2**
*Comprobar que la condición de convexidad es necesaria para que el resultado de la Proposición 4.1 sea válido. En particular: dibujar curvas de nivel de una función $f$ diferenciable, un conjunto $\mathscr{X}$ no convexo, y encontrar un punto $x \in \mathscr{X}$ tal que la dirección $x - x^\star$ tenga un ángulo de más de $\pi/2$ con el gradiente.*

Tomemos como primer ejemplo una aplicación que además será muy útil en el resto del capítulo.

**Definición 4.3.** Sea $\mathscr{X} \in \mathbb{R}^n$ un conjunto convexo, y $z \in \mathbb{R}^n$. Definimos la proyección de $z$ sobre $\mathscr{X}$ como el punto

$$\arg\min_{x \in \mathscr{X}} \|z - x\|^2.$$

Es decir, el punto del conjunto $\mathscr{X}$ que está más cerca de $z$. Al mapa que asigna a cada punto $z \in \mathbb{R}^n$, su proyección en el conjunto $\mathscr{X}$, lo denotamos $P_{\mathscr{X}} : \mathbb{R}^n \to \mathscr{X}$. Es decir

$$P_{\mathscr{X}}(z) = \arg\min_{x \in \mathscr{X}} \|z - x\|^2.$$

La proyección está definida como un problema de optimización con restricciones (minimizar la función distancia a $z$, en el conjunto $\mathscr{X}$), y la función objetivo es convexa, por lo que el mínimo verifica la condición de optimalidad. Veamos qué dicen en este caso esas condiciones.

El gradiente de $f(x) = \|z-x\|^2$ es $\nabla f(x) = -2(z-x)$. Evaluando en el óptimo $x\star = P_{\mathscr{X}}(z)$, el gradiente es entonces $\nabla f(P_{\mathscr{X}}(z)) = -2(z-P_{\mathscr{X}}(z))$, por lo que la condición de optimalidad queda:

$$\langle -2\,(z-P_{\mathscr{X}}(z))\,, x-P_{\mathscr{X}}(z)\rangle \geqslant 0 \quad \forall x \in \mathscr{X}\,,$$

o lo que es lo mismo,

$$\langle z-P_{\mathscr{X}}(z), x-P_{\mathscr{X}}(z)\rangle \leqslant 0 \quad \forall x \in \mathscr{X}\,. \tag{4.1}$$

El primer término del producto interno, $z-P_{\mathscr{X}}(z)$, es el vector que une la proyección de $z$ con el mismo $z$ (ver figura). Los términos $x-P_{\mathscr{X}}(z)$, con $x$ variando en el conjunto $\mathscr{X}$, son todos los vectores que se pueden formar desde $P_{\mathscr{X}}(z)$ hacia puntos del conjunto. La condición de optimalidad entonces dice que el ángulo entre estos dos vectores tiene que ser mayor o igual a $\pi/2$, que es lo que uno esperaría para la proyección.

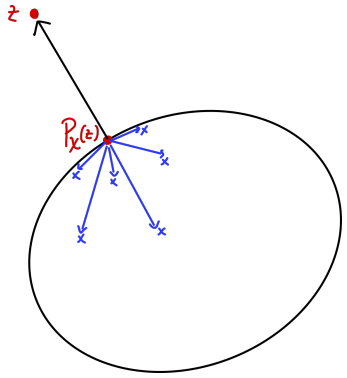



Figura 4.3: Condición de optimalidad para la proyección.

**Observación 4.4.** Si $\mathscr{X}$ es un subespacio vectorial (que en particular es un conjunto convexo), la condición de optimalidad resulta $z-P_{\mathscr{X}}(z) \in \mathscr{X}^{\perp}$, y la proyección es la proyección ortogonal que aprendimos en álgebra lineal.

En la siguiente observación vemos cómo la condición de optimalidad nos da información del comportamiento de la proyección.

**Observación 4.5.** Sea $\mathscr{X}$ un conjunto convexo, $x$ e $y$ dos puntos en $\mathbb{R}^n$, y sean $P_{\mathscr{X}}(x)$ y $P_{\mathscr{X}}(y)$ sus respectivas proyecciones. Entonces se cumple

$$\|P_{\mathscr{X}}(x) - P_{\mathscr{X}}(y)\| \leqslant \|x-y\|.$$

Esta propiedad se denomina *no expansividad*. La proyección es un operador no expansivo. Observar que esto es menos fuerte que ser una contracción, porque podría suceder que se de la igualdad (ejercicio: piense un ejemplo donde se de la igualdad).

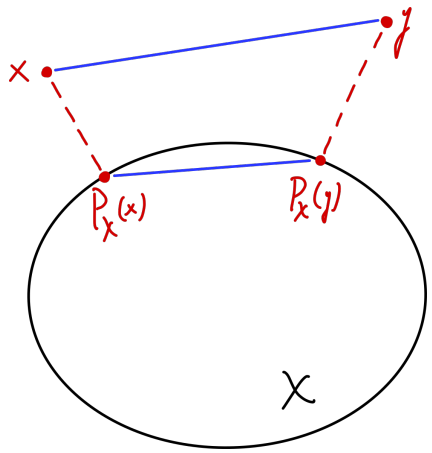



Figura 4.4: La proyección es un operador no expansivo.

**Ejercicio 4.6**
*Usando la condición de optimalidad para $(x, P_{\mathcal{X}}(x))$ y para $(y, P_{\mathcal{X}}(y))$, combinando las expresiones, y utilizando la desigualdad de Cauchy-Schwarz, probar la desigualdad anterior.*

Volviendo al problema original, es decir,

$$\min_{\mathcal{X}} f(x)$$

para una función $f$ diferenciable cualquiera, veamos que existe una manera alternativa de escribir la condición de optimalidad, que nos resultará muy útil para pensar los algoritmos.

Veamos que la condición de optimalidad es equivalente a la siguiente condición:

$$x^\star = P_{\mathcal{X}}\left(x^\star - \nabla f(x^\star)\right). \tag{4.2}$$

En efecto, tomemos la condición de optimalidad:

$$\langle \nabla f(x^\star), x - x^\star \rangle \geqslant 0 \quad \forall x \in \mathcal{X},$$

y en el primer término multiplicamos por menos uno, y sumamos y restamos $x^\star$:

$$\langle x^\star - \nabla f(x^\star) - x^\star, x - x^\star \rangle \leqslant 0 \quad \forall x \in \mathcal{X}.$$

Ahora basta reconocer a esta expresión como la condición de optimalidad de (4.2). Ejercicio: ver que esta última desigualdad es exactamente la condición de optimalidad para la proyección, con $z = x^\star - \nabla f(x^\star)$.

Esta condición se denomina condición de estacionareidad, y es fácil ver que también es cierto si agregamos un $\alpha > 0$:

$$x^\star = P_{\mathcal{X}}\left(x^\star - \alpha \nabla f(x^\star)\right). \tag{4.3}$$

Intepretemos esta condición. Dice que si un punto $x^\star$ es un mínimo local del problema con restricciones, entonces sucede lo siguiente: si estamos parados en $x^\star$, nos movemos cualquier paso $\alpha$ en la dirección opuesta al gradiente, y proyectamos al conjunto $\mathscr{X}$, entonces caemos en el mismo $x^\star$. Recordemos que movernos en la dirección opuesta al gradiente era, localmente, lo mejor que podíamos hacer. Eso eventualmente nos puede llevar fuera del conjunto, y entonces proyectamos de vuelta a $\mathscr{X}$. Si este procedimiento "se estanca" para cualquier paso $\alpha$, quiere decir que ya estábamos en el mejor punto (localmente).

Este mecanismo, moverse según el gradiente y proyectar, es de hecho uno de los algoritmos que veremos en breve.

**Observación 4.7.** Recordemos que en el capítulo anterior habíamos usado el mapa $\psi(x) = x - \nabla f(x)$, y particularmente vimos que si este mapa $\psi$ era una contracción, entonces el descenso por gradiente converge a un punto crítico. Ahora a este mapa le estamos agregando una proyección: $\psi(x) = P_{\mathscr{X}}(x - \nabla f(x))$, pero vimos que la proyección es no expansiva, por lo que estaremos componiendo un mapa no expansivo con una contracción, y el resultado es entonces una contracción.

## 4.2. Condiciones de optimalidad para funciones no diferenciables

Veamos en esta sección que se pueden extender las condiciones de optimalidad cuando la función objetivo no es diferenciable. Para esto necesitaremos una extensión del concepto de diferenciabilidad.

Trabajemos con funciones convexas. Recordemos que si $f$ es convexa y diferenciable, entonces la recta tangente por cualquier punto soporta el gráfico de la función. Para funciones no diferenciables, también puede haber rectas (o hiperplanos) que soporten el gráfico de la función en un punto, aunque este punto no sea de diferenciabilidad. Vamos a llamar *subgradiente* a estos elementos Más precisamente:

**Definición 4.8.** Dada una función $f : \mathbb{R}^n \to \mathbb{R}$ convexa y $x \in \mathbb{R}^n$, decimos que $d \in \mathbb{R}^n$ es un subgradiente de $f$ en $x$ sii $f(z) \geqslant f(x) + \langle z - x, d\rangle$ para todo $z \in \mathbb{R}^n$.

Como decíamos, esto es una generalización del concepto de plano tangente en un punto: $d$ es un subgradiente si el plano que define queda completamente por debajo de la función. En la figura 4.5 se puede ver una interpretación geométrica en el caso unidimensional.

Cuando $f$ es diferenciable en un punto, entonces el único subgradiente posible es el gradiente tradicional. En los puntos de no diferenciabilidad, puede haber más de un subgradiente,como se ve en la figura. Al conjunto de todos los subgradientes de $f$ en un punto $x$ se lo denomina subdiferencial, y se denota como $\partial f(x)$.

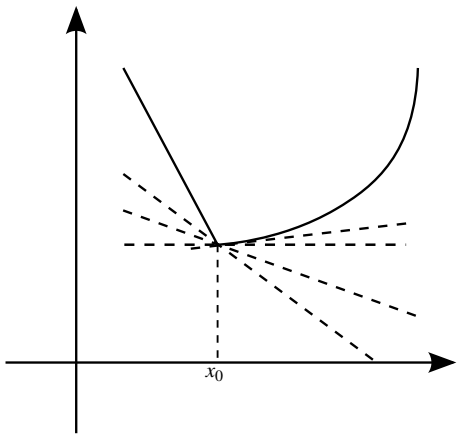



Figura 4.5: Función no diferenciable en $x_0$, con rectas que dejan el gráfico de la función completamente por arriba (subgradientes).

**Ejemplo 4.9**
*Para la función $f(x) = |x|$, tenemos por ejemplo $\partial f(1) = \{1\}$ y $\partial f(0) = [-1,1]$. Es decir, todas las rectas por el origen con pendiente entre $-1$ y $1$, dejan al gráfico de $f$ por arriba.*

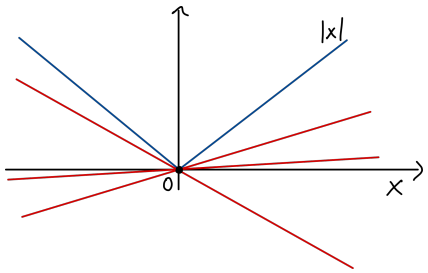



Figura 4.6: Subgradientes de $|x|$ en 0. En azul la gráfica de la función, y en rojo algunas rectas que soportan el gráfico en $x = 0$.

Ahora podemos dar una condición de optimalidad más general:

**Proposición 4.10.** Sea $f : \mathscr{X} \to \mathbb{R}$ convexa, donde $\mathscr{X}$ es un subconjunto convexo de $\mathbb{R}^n$. Entonces $x^\star$ minimiza $f$ sii existe un subgradiente $d \in \partial f(x^\star)$ tal que $\langle d, z - x^\star \rangle \geqslant 0$ para todo $z \in \mathscr{X}$.

Observar que si $f$ es diferenciable, obtenemos exactamente la condición anterior, y que si $x^\star$ está en el interior de $\mathscr{X}$, la condición queda $0 \in \partial f(x^\star)$, que generaliza la condición de anular el gradiente.

**Ejercicio 4.11**
*Interprete geométricamente la condición $0 \in \partial f(x^\star)$.*

Veamos una interpretación geométrica de las condiciones de optimalidad generales de la Proposición 4.10. Para esto, definamos antes el cono normal:

**Definición 4.12.** Dado un conjunto convexo $\mathscr{X}$ y un punto $x \in \mathscr{X}$, llamamos cono normal a $\mathscr{X}$ por $x$ al conjunto

$$N_{\mathscr{X}}(x) = \{g \in \mathbb{R}^n : \langle g, (z-x)\rangle \leqslant 0, \forall z \in \mathscr{X}\}.$$

Es decir, son las direcciones $g$ que forman un ángulo de al menos $\pi/2$ con las direcciones factibles $z-x$. Se puede pensar también como los vectores $g$ que definen un hiperplano que soporta al conjunto $\mathscr{X}$ en $x$.

Esto permite generalizar el concepto que el gradiente es normal a las curvas de nivel. Si $f$ es una función convexa (no necesariamente diferenciable), fijamos $x_0 \in \mathbb{R}^n$, y consideramos el conjunto de sub-nivel $C = \{z \in \mathbb{R}^n : f(x) \leqslant f(x_0)\}$. Entones es fácil ver (ejercicio) que la condición para que $g \in \partial f(x_0)$, es equivalente a que $g$ pertenezca al cono normal al conjunto de nivel por $x_0$: $g \in N_C(x_0)$. Si el borde de $C$ es diferenciable, entones el cono normal es simplemente la semirrecta perpendicular al borde, por lo que en este caso, recuperamos lo sabido del gradiente y las curvas de nivel.

Volvamos ahora a las condiciones de optimalidad. La condición dada en la Proposición 4.10 (que exista un subgradiente $d$ tal que $\langle d, z-x^*\rangle \geqslant 0$ para todo $z \in \mathscr{X}$), se traduce en que exista un subgradiente $d \in \partial f(x^*)$ tal que $-d \in N_{\mathscr{X}}(x^*)$. Veamos un par de ejemplos.

Consideremos primero una función $f$ diferenciable, pero un conjunto $\mathscr{X}$ convexo, con borde no diferenciable. Como $f$ es diferenciable (y por lo tanto el único subgradiente es el gradiente), la condición de optimalidad es $-\nabla f(x^*) \in N_{\mathscr{X}}(x^*)$. Este es el caso que se puede ver en la Figura 4.7.

Si ahora $f$ es no diferenciable, pero asumimos que el mínimo se alcanza en un punto $x^*$ donde el borde del conjunto $\mathscr{X}$ es diferenciable, entonces el cono normal a $\mathscr{X}$ será simplemente una semirrecta, a la cual deberá pertenecer el opuesto de algún subgradiente de $f$. Este es el caso de la Figura 4.8. Se puede pensar, como ejercicio, cómo es geométricamente la condición de optimalidad cuando la función es no diferenciable y el mínimo se alcanza en un punto de no diferenciabilidad del borde de $\mathscr{X}$.

## 4.3. Algoritmos para minimización con restricciones

Volviendo al problema de minimizar una función diferenciable $f$ en un conjunto convexo y cerrado $\mathscr{X}$, vamos a considerar métodos similares a los del capítulo anterior.

Específicamente, vamos a considerar métodos de la forma $x_{k+1} = x_k + \alpha_k d_k$, donde volvemos a movernos de $x_k$ al punto siguiente, avanzando un paso $\alpha_k$ según una dirección $d_k$. Lo

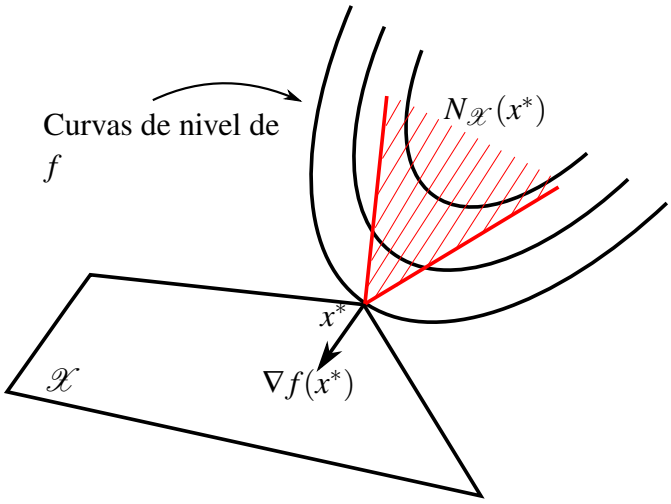



Figura 4.7: Curvas de nivel de una función diferenciable, y el mínimo se alcanza en un vértice de $\mathscr{X}$. En rojo, el cono normal a $\mathscr{X}$ por $x^*$, y se puede observar que el opuesto del gradiente pertenece al cono normal.

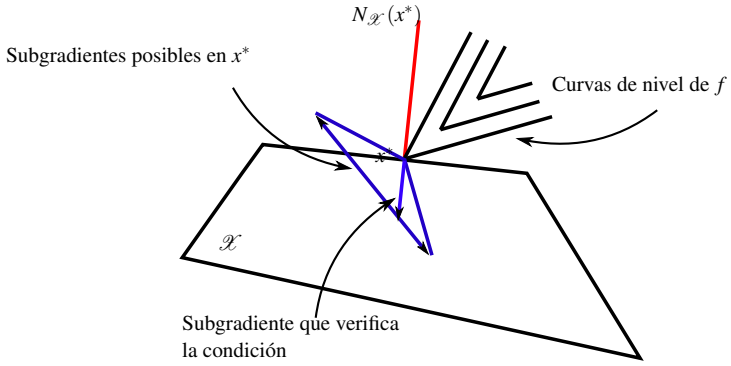



Figura 4.8: Curvas de nivel de una función no diferenciable. Como el mínimo se alcanza en un punto de diferenciabilidad del borde de $\mathscr{X}$, el cono normal es una semirrecta (en rojo). En azul, los posibles subgradientes de $f$ en $x^*$ (observar que forman un ángulo de $\pi/2$ con la curva de nivel por $x^*$), y el subgradiente cuyo opuesto pertenece al cono normal.

que vamos a pedir en esta oportunidad, además de que la dirección sea de descenso, es que la dirección sea factible. Es decir, que sea una dirección que nos permita movernos (al menos un poquito) dentro del conjunto $\mathscr{X}$. Recordemos que las direcciones factibles desde $x_k$ son las que podemos escribir como $d_k = x - x_k$ para algún $x \in \mathscr{X}$.

Para métodos de esta forma, hay un resultado de convergencia a un punto crítico, pidiendo algunas hipótesis razonables sobre el paso y las direcciones[1].

[1]Para las direcciones, lo que debemos pedir es que sean lo que se denomina *gradient-related*. Se puede consultar en [2]

Veamos algunas formas de elegir direcciones y pasos, que dan lugar a métodos con distintas denominaciones.

### 4.3.1. Método de Frank-Wolfe (o del gradiente condicional)

El primer método es el denominado de Frank-Wolfe. Recordemos que debemos elegir la dirección $d_k$, y que esto es equivalente a elegir un punto en el conjunto hacia el cual dirigirnos desde $x_k$.

Denominemos esto de la siguiente forma: $d_k = \bar{x_k} - x_k$. Es decir, estando parados en el punto $x_k \in \mathscr{X}$, elegimos un punto $\bar{x_k} \in \mathscr{X}$ que nos definirá la dirección factible $d_k = \bar{x_k} - x_k$.

Esta dirección debe ser de descenso, como ya discutimos varias veces en estas notas. Ser de descenso, recordemos, es que la derivada direccional sea negativa, lo que podemos escribir como

$$\langle \nabla f(x_k), \bar{x_k} - x_k \rangle < 0.$$

**Observación 4.13.** Si **no existe** ningún $\bar{x_k} \in \mathscr{X}$ de forma que ese producto interno sea negativo, eso quiere decir que para todo punto del conjunto, el producto interno es positivo. Es decir, que

$$\langle \nabla f(x_k), x - x_k \rangle \geqslant 0, \quad \forall x \in \mathscr{X},$$

y esto es exactamente la condición de optimalidad para minimización con restricciones. Es decir que si pasa eso, es porque ya estamos en un punto que verifica la condición de optimalidad, un punto estacionario, que es lo máximo a lo que puedo aspirar en un contexto general.

Hay entonces al menos uno (en general infinitos) posibles elecciones para el punto $\bar{x_k}$. El método de Frank-Wolfe consiste en elegir el punto $\bar{x_k}$ de manera que minimice justamente el producto interno $\langle \nabla f(x_k), x - x_k \rangle$. Es decir:

$$\bar{x_k} = \arg\min_{x \in \mathscr{X}} \langle \nabla f(x_k), x - x_k \rangle. \tag{4.4}$$

De alguna forma, es como buscar la "dirección de máximo descenso", donde se minimice la derivada direccional. Esto antes lo teníamos trivialmente con la dirección opuesta al gradiente, pero con las restricciones deja de ser trivial. Vamos a ver en breve que en realidad hay una sutileza, y esta dirección no es la que minimiza la derivada direccional.

**Observación 4.14.** Encontrar la dirección (4.4), que es algo que tenemos que hacer en cada iteración, es a su vez un problema de optimización, y de hecho tiene las mismas restricciones que el problema original ($x \in \mathscr{X}$). Pero observemos que en este problema la función a minimizar es lineal en $x$, por lo que en general es un problema más fácil.

El paso $\alpha_k$ lo podemos elegir de forma similar a descenso por gradiente sin restricciones, y en particular la regla de Armijo es una opción muy utilizada.

Volviendo al problema (4.4), observemos que en realidad al minimizar este producto interno, como no estamos imponiendo nada sobre la norma de la dirección, en realidad el resultado depende de dos factores: cuán alineada está la dirección con el opuesto del gradiente, y cuán "lejos" está el punto $\bar{x_k}$ del punto $x_k$. Ver Figura 4.9.

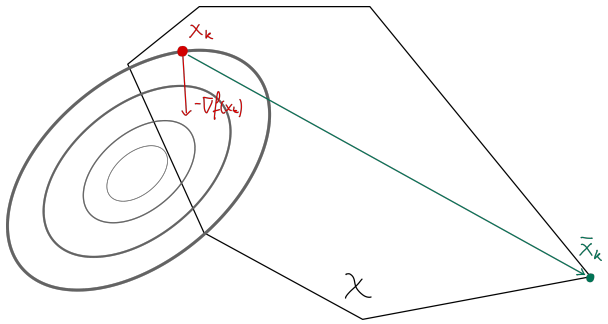



Figura 4.9: Elección de $\bar{x_k}$ en Frank-Wolfe. Observar que al minimizar el producto interno con el gradiente, influye la distancia de $\bar{x_k}$ a $x_k$.

Cuando el conjunto $\mathscr{X}$ es un poliedro (por ejemplo, dado por restricciones que sean igualdades y desigualdades lineales), el problema (4.4) es lineal, y la solución $\bar{x_k}$ es genéricamente un vértice del poliedro.

El método consiste entonces en elegir la dirección $d_k$ como indicamos, un paso $\alpha_k$ de forma similar al capítulo anterior, y repetir el procedimiento hasta llegar a un punto estacionario.

### 4.3.2. Método del gradiente proyectado

El otro método que veremos en este capítulo, denominado Método del gradiente proyectado, es también de la forma $x_{k+1} = x_k + \alpha_k d_k$, con una dirección $d_k = \bar{x_k} - x_k$, pero varía en la forma en la que se elije $\bar{x_k}$.

Específicamente, se establece que

$$\bar{x_k} = P_{\mathscr{X}}\left(x_k - s_k \nabla f(x_k)\right), \tag{4.5}$$

donde $s_k > 0$ es otro parámetro, y en este caso el paso $\alpha_k$ debe estar en $(0,1]$.

Veamos qué es lo que estamos haciendo. Elegimos la dirección, es decir, el punto hacia el cual vamos a "mirar", de la siguiente forma: desde $x_k$ hacemos un cierto paso $s_k$ de descenso por gradiente (tradicinal), y el resultado lo proyectamos al conjunto $\mathscr{X}$, pues eventualmente

nos podríamos haber salido del conjunto. Una vez elegida la dirección, decidimos cuánto nos movemos en esa dirección con $\alpha_k$.

Para fijar ideas, analicemos el caso en que $\alpha_k = 1$. En ese caso, tenemos que

$$x_{k+1} = x_k + 1(\bar{x_k} - x_k) = \bar{x_k} \quad \Longrightarrow \quad x_{k+1} = P_{\mathscr{X}}\left(x_k - s_k \nabla f(x_k)\right).$$

Esto es exactamente lo que discutimos cuando vimos la condición de estacionareidad (4.3): estando parados en $x_k$, nos movemos en la dirección opuesta al gradiente, y si eso nos saca del conjunto proyectamos.

¿Cuándo se estanca este método? Es decir, ¿cuándo ocurre que el iterado $x_{k+1}$ coincide con $x_k$? Bueno, si ocurre eso se cumple la condición de estacionareidad, que vimos que es equivalente a la condición de optimalidad.

Veamos geométricamente cómo se comporta este método, siguiendo con el caso $\alpha_k = 1$.

Comencemos viendo algunos iterados de una trayectoria usual. En la Figura 4.10 tenemos algunas curvas de nivel de una función diferenciable, y un conjunto $\mathscr{X}$. Comenzando en $x_0 \in \mathscr{X}$, nos movemos en la dirección opuesta del gradiente un cierto paso $s_1$, que resulta que nos saca del conjunto, por lo que proyectamos. Ahora $x_1$ está en el borde de $\mathscr{X}$, calculamos el gradiente, y nuevamente nos movemos en la dirección opuesta un cierto paso $s_2$, y también volvemos a proyectar. Esto se repite hasta eventualmente llegar al punto $x^\star$ marcado en la figura, donde se puede ver que la curva de nivel es tangente al borde del conjunto, como discutimos antes en este capítulo.

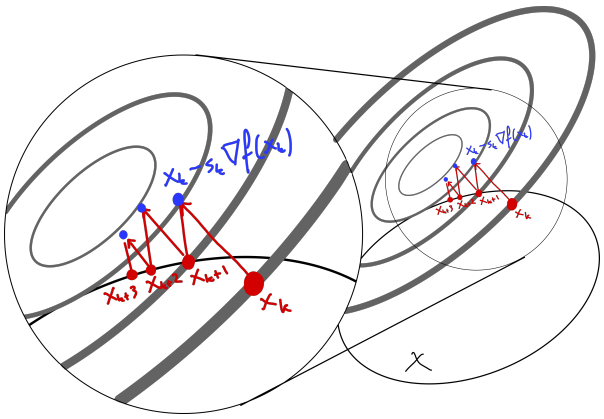


Figura 4.10: Algunos iterados del método de gradiente proyectado.

Siguiendo con el caso $\alpha_k = 1$, y estando parados en cierto punto $x_k$, veamos cómo es el lugar geométrico de los puntos $x_{k+1}$ en función de la elección del paso $s_k$. En la Figura 4.11 vemos esto: para valores chicos de $s_k$, el paso que damos en la dirección opuesta al gradiente todavía nos deja dentro del conjunto. Por lo tanto nos vamos moviendo, conforme $s_k$ aumenta,

en el segmento que une $x_k$ con el borde del conjunto, en la dirección opuesta al gradiente. Para valores de paso $s_k$ más grandes, se activa la proyección, y nos comenzamos a mover a lo largo del borde de $\mathscr{X}$.

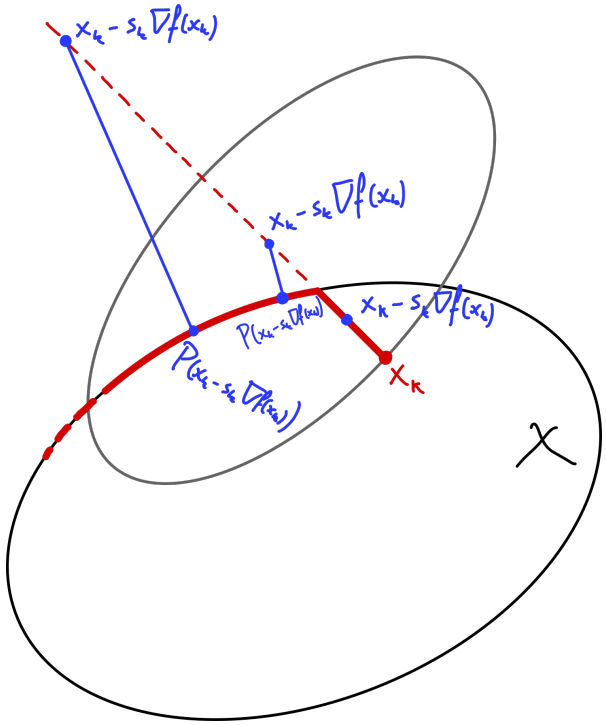



Figura 4.11: Dado un punto $x_k \in \mathscr{X}$, lugar geométrico de los posibles puntos $x_{k+1}$, para distintas elecciones de paso $s_k$. Para valores pequeños de $s_k$, el próximo punto se encontrará en el segmento que une $x_k$ con el borde del conjunto, en la dirección opuesta al gradiente. Para valores más grandes de $s_k$, el punto $x_k - s_k \nabla f(x_k)$ cae fuera del conjunto, y debemos proyectar al borde. Los posibles puntos entonces se encuentran en la curva marcada en rojo.

Como punto final, y adelantándonos a una discusión de un capítulo posterior, vamos a comentar que este método se puede acelerar "à la Nesterov", es decir, combinar con el método de gradiente acelerado:

$$\begin{cases} y_k = x_k + \beta_k (x_k - x_{k-1}) \\ x_{k+1} = P_{\mathscr{X}} \left( y_k - s_k \nabla f(y_k) \right). \end{cases}$$

con las mismas elecciones de $\beta_k$ que discutimos en el capítulo anterior.

# Bibliografía


[1] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[2] Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[3] Dimitri P Bertsekas. *Convex optimization algorithms*. Athena Scientific Belmont, 2015.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[5] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

[6] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[7] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[8] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[9] Daniel P Palomar and Yonina C Eldar. *Convex optimization in signal processing and communications*. Cambridge university press, 2010.

[10] Jonathan Richard Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. 1994.

[11] D.J. Wilde. *Optimum Seeking Methods*. Prentice-Hall international series in the physical and chemical engineering sciences. Prentice-Hall, 1964.

[12] Stephen Wright, Jorge Nocedal, et al. Numerical Optimization. *Springer Science*, 35(67-68):7, 1999.