Apuntes de la asignatura

Programación Matemática

Curso 2020-2021

Prof. Julio González Díaz

Índice general

Prólogo

Estos apuntes han sido preparados como apoyo para la asignatura de Programación Matemática del Máster Interuniversitario en Técnicas Estadísticas en el que participan las tres universidades gallegas. Se presupone que el alumno ya tiene conocimientos básicos de programación lineal.

Para la elaboración de este material se han usado distintas referencias, entre las que destacan el libro "Nonlinear Programming: Theory and Algorithms" (Bazaraa y otros, 2006) siendo algunas secciones de estas notas poco más que meras traducciones del mismo. También se ha usado el libro "Nonlinear Optimization" (Ruszczyński, 2006) y material de distintos cursos de optimización impartidos por profesores invitados en la Facultad de Matemáticas de la Universidad de Santiago de Compostela.

En una de sus formulaciones más habituales, un problema de programación matemática se puede escribir del siguiente modo

$$\begin{array}{rll} \text{minimizar} & f(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 & i = 1, \ldots, m \\ & h_j(\boldsymbol{x}) = 0 & j = 1, \ldots, l. \end{array}$$

El principal objetivo de este curso será que el alumno se familiarice con los principales conceptos teóricos y prácticos necesarios para abordar exitosamente este tipo de problemas en el caso más general, en el que las funciones f, g_i y h_j pueden ser funciones no lineales y no convexas.

Es importante destacar que, incluso para aquellos estudiantes con una clara orientación hacia la estadística, puede resultar de gran utilidad tener una sólida base en optimización. La gran mayoría de las técnicas de estimación estadística acaban teniendo que hacer de una manera un ajuste de parámetros, intentando minimizar algún criterio: minimizar errores, maximizar verosimilitud,... El caso más sencillo es el de la regresión lineal, cuya resolución pasa por un resolver un problema de mínimos cuadrados cuya solución se puede encontrar de forma explícita. El caso de la regresión no lineal ya resulta más complejo, y los problemas de mínimos cuadrados no lineales resultantes pasa por el uso de algoritmos específicos de optimización, entre los que destaca el de Levenberg-Marquardt.

A lo largo de estas notas se irán proponiendo distintos ejercicios. Como indicación acerca de la dificultad de los mismos se usará "Ejercicio" para ejercicios muy sencillos, "•Ejercicio" para ejercicios sencillos, "••Ejercicio" para ejercicios intermedios y, por último, "•••Ejercicio" para ejercicios cuya resolución podría llegar a requerir consultar material adicional.

Como consideración relativa a la notación mencionar que, salvo que se diga lo contrario, a lo largo de estas notas todo vector en $\mathbb{R}^n$ se pensará como vector columna $(n \times 1)$ y que tanto **vectores** como **matrices** se denotarán mediante letras **negritas**. Las componentes de

los vectores se denotarán con subíndices, y se usarán superíndices para denotar los elementos de una sucesión. Para minimizar posibles confusiones con la notación, la letra t se usará únicamente como índice de los elementos de una sucesión o de las iteraciones de un determinado algoritmo.

Se agradecerá que cualquier comunicación relativa a erratas o errores de estos apuntes a través de la siguiente dirección de correo electrónico: julio.gonzalez@usc.es.

. . . . . .

Prof. Julio González Díaz

Tema 1
Introducción al Análisis Convexo

Contenidos

1.1 Preliminares

La optimización de funciones convexas es un elemento muy importante de la programación no lineal. No sólo porque muchos problemas de optimización son en efecto problemas definidos sobre funciones y regiones convexas, sino también porque muchos métodos en optimización no convexa se basan en trabajar con sucesiones de problemas convexos que sean aproximaciones del problema original.

En este primer tema veremos una pequeña muestra de los resultados clásicos en análisis convexo que resultan fundamentales en el estudio de problemas de optimización no lineal, centrándonos especialmente en aquellos que serán más relevantes durante el resto del curso.

Antes de empezar a hablar de convexidad, presentamos una observación geométrica que usaremos frecuentemente y un breve recordatorio de topología. El producto escalar de dos vectores $\boldsymbol{x}$ e $\boldsymbol{y}$, $\boldsymbol{y}^\intercal\boldsymbol{x}$, se puede expresar como el producto de las normas de ambos vectores por el coseno del ángulo que forman. Por tanto, una condición de la forma $\boldsymbol{y}^\intercal\boldsymbol{x} \leq 0$ quiere decir que el ángulo que forman el vector $\boldsymbol{y}$ y el vector $\boldsymbol{x}$ es de al menos 90°. Análogamente, $\boldsymbol{y}^\intercal\boldsymbol{x} \geq 0$ quiere decir que dicho ángulo es, a lo sumo, 90°.

Breve recordatorio de los conceptos básicos de topología

Definimos la *bola* o *ε-entorno* de centro $\boldsymbol{x} \in \mathbb{R}^n$ y radio $\varepsilon \in \mathbb{R}$, $B(\boldsymbol{x}, \varepsilon)$, como el conjunto de los puntos que distan de $\boldsymbol{x}$ menos de ε: $B(\boldsymbol{x}, \varepsilon) = \{\boldsymbol{y} \in \mathbb{R}^n : \|\boldsymbol{y} - \boldsymbol{x}\| < \varepsilon\}$. Una bola no es más que la generalización de los conceptos de círculo en $\mathbb{R}^2$ y esfera en $\mathbb{R}^3$ a cualquier dimensión. Ahora definiremos clausura, interior y frontera de un conjunto en $\mathbb{R}^n$ apoyándonos en este tipo de entornos.

Clausura (o adherencia). Un punto $\boldsymbol{x}$ está en la *clausura* de S, $\bar{S}$, si cualquier entorno de $\boldsymbol{x}$ interseca a S. Equivalentemente, para todo $\varepsilon > 0$, $S \cap B(\boldsymbol{x}, \varepsilon) \neq \emptyset$.

Intuitivamente, la clausura de un conjunto S está formada por aquellos puntos del espacio que están "pegados" a S.

Interior. Un punto $\boldsymbol{x}$ está en el *interior* de S, $\mathring{S}$, si existe algún entorno de $\boldsymbol{x}$ que esté contenido en S. Equivalentemente, existe $\varepsilon > 0$, tal que $B(\boldsymbol{x}, \varepsilon) \subseteq S$.

Intuitivamente, un punto está en el interior de un conjunto S si está completamente rodeado de puntos del conjunto.

Frontera. Un punto $\boldsymbol{x}$ está en la *frontera* de S, ∂S, si todo entorno de $\boldsymbol{x}$ contiene puntos de S y del complementario de S. Equivalentemente, para todo $\varepsilon > 0$, $B(\boldsymbol{x}, \varepsilon) \cap S \neq \emptyset$ y $B(\boldsymbol{x}, \varepsilon) \cap \mathbb{R}^n \backslash S \neq \emptyset$.

Ahora podemos apoyarnos en estos conceptos para definir distintas propiedades topológicas de conjuntos.

Conjunto cerrado. Un conjunto $S \subseteq \mathbb{R}^n$ es *cerrado* si $S = \bar{S}$.

Conjunto abierto. Un conjunto $S \subseteq \mathbb{R}^n$ es *abierto* si $S = \mathring{S}$.

Conjunto acotado. Un conjunto $S \subseteq \mathbb{R}^n$ es *acotado* si existe $r > 0$ tal que $S \subseteq B(\boldsymbol{x}, r)$.

Conjunto compacto. Un conjunto $S \subseteq \mathbb{R}^n$ es *compacto* si es cerrado y acotado.[1]

Para terminar este recordatorio presentamos, sin demostración, una serie de propiedades topológicas elementales. El alumno que no esté familiarizado con la topología puede intentar demostrar estas propiedades por sí mismo o consultar algún manual básico topología.

Proposición 1.1. *Dado un conjunto S, las siguientes relaciones se cumplen:*

(i) $S \subseteq \bar{S}$, $\mathring{S} \subseteq S$ *y* $\partial S = \bar{S} \backslash \mathring{S}$.

(ii) *Un conjunto es cerrado si y sólo si contiene contiene a todos los puntos de su frontera.*

(iii) *Un conjunto es abierto si y sólo si contiene a ningún punto de su frontera.*

(iv) *La clausura de un conjunto S es el conjunto cerrado más pequeño que lo contiene.*

(v) *El interior de un conjunto S es el conjunto abierto más grande contenido en él.*

•**Ejercicio 1.1.** Identifica la clausura, el interior y la frontera de cada uno de los siguientes conjuntos convexos:

(i) $\{\boldsymbol{x} \in \mathbb{R}^3 : x_1^2 + x_3^2 \leq x_2\}$.

(ii) $\{\boldsymbol{x} \in \mathbb{R}^2 : 2 \leq x_1 \leq 5, x_2 = 4\}$.

(iii) $\{\boldsymbol{x} \in \mathbb{R}^3 : \boldsymbol{x} \geq \boldsymbol{0},\ x_1 + x_2 \leq 5,\ -x_1 + x_2 + x_3 \leq 7\}$.

(iv) $\{\boldsymbol{x} \in \mathbb{R}^3 : x_1 + x_2 = 5,\ x_1 + x_2 + x_3 \leq 8\}$.

(v) $\{\boldsymbol{x} \in \mathbb{R}^3 : x_1^2 + x_2^2 + x_3^2 \leq 9,\ x_1 + x_3 \leq 2\}$. ◁

1.2 Conjuntos convexos y propiedades

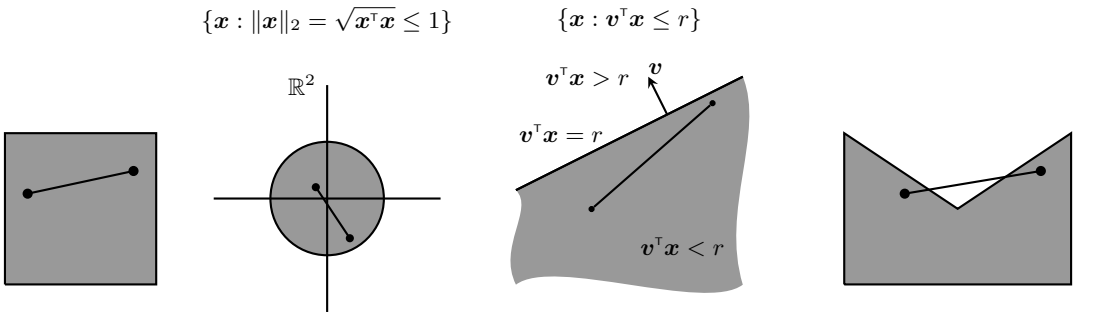
1.2.1 Definiciones básicas

A continuación presentamos, a modo de recordatorio, una serie de conceptos básicos que serán empleados de modo rutinario durante estas notas.

Dados dos puntos $\boldsymbol{x}$ e $\boldsymbol{y}$ en $\mathbb{R}^n$ y dado $\lambda \in [0, 1]$, el punto $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}$ es una *combinación convexa* de $\boldsymbol{x}$ e $\boldsymbol{y}$. El conjunto de todas las combinaciones convexas de $\boldsymbol{x}$ e $\boldsymbol{y}$ coincide con el segmento que une estos dos puntos. Un conjunto $S \subseteq \mathbb{R}^n$ se dice que es *convexo* si cualquier combinación convexa de puntos de S también pertenece a S. En general, dados puntos $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k$, una combinación convexa de estos puntos es de la forma $\sum_{i=1}^k \lambda_i \boldsymbol{x}^i$, donde los λ_i son no negativos y $\sum_{i=1}^k \lambda_i = 1$. Dado un conjunto cualquiera S, la *envoltura convexa* de S, $\text{conv}(S)$, se define como el conjunto de todas las combinaciones convexas de puntos de S. La Figura 1.1(a) muestra varios conjuntos convexos y la Figura 1.1(b) uno no convexo.

A continuación presentamos un ejemplo ilustrativo de cómo probar la convexidad de un conjunto dado.

[1] Realmente la definición de conjunto compacto es más general, pero en el caso particular de espacios euclídeos es equivalente a la que acabamos de presentar.



(a) Conjuntos convexos. (b) Conjunto no convexo.

Figura 1.1: Ilustrando el concepto de convexidad.

hola

Tabla 1.1: af

Ejemplo 1.1. Vamos a demostrar que, dados tres vectores $\boldsymbol{z}^1$, $\boldsymbol{z}^2$ y $\boldsymbol{z}^3$ en $\mathbb{R}^n$, entonces el conjunto $S = \{\alpha_1\boldsymbol{z}^1 + \alpha_2\boldsymbol{z}^2 + \alpha_3\boldsymbol{z}^3 : \alpha_1 \geq 0,\ \alpha_2 \geq 0,\ \alpha_3 \geq 0 \text{ y } \alpha_1 + \alpha_2 + \alpha_3 = 1\}$ es un conjunto convexo. El conjunto S aparece representado en la Figura 1.2.

Tomemos dos puntos $\boldsymbol{x} \in S$ e $\boldsymbol{y} \in S$ y $\lambda \in (0,1)$. Para probar que el conjunto S es convexo basta probar que el punto $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}$ pertenece a S. Como $\boldsymbol{x}$ e $\boldsymbol{y}$ pertenecen a S, existen números no negativos α_1, α_2, α_3, β_1, β_2 y β_3 con $\alpha_1 + \alpha_2 + \alpha_3 = 1$ y $\beta_1 + \beta_2 + \beta_3 = 1$ tales que

$$\begin{aligned}\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y} &= \lambda(\alpha_1\boldsymbol{z}^1 + \alpha_2\boldsymbol{z}^2 + \alpha_3\boldsymbol{z}^3) + (1-\lambda)(\beta_1\boldsymbol{z}^1 + \beta_2\boldsymbol{z}^2 + \beta_3\boldsymbol{z}^3)\\ &= (\lambda\alpha_1 + (1-\lambda)\beta_1)\boldsymbol{z}^1 + (\lambda\alpha_2 + (1-\lambda)\beta_2)\boldsymbol{z}^2 + (\lambda\alpha_3 + (1-\lambda)\beta_3)\boldsymbol{z}^3\\ &= \gamma_1\boldsymbol{z}^1 + \gamma_2\boldsymbol{z}^2 + \gamma_3\boldsymbol{z}^3.\end{aligned}$$

Claramente, como $\lambda \in (0,1)$ y los α_i y β_j son no negativos, tenemos que γ_1, γ_2 y γ_3 son no negativos. Además,

$$\begin{aligned}\gamma_1 + \gamma_2 + \gamma_3 &= (\lambda\alpha_1 + 1 - (\lambda)\beta_1) + (\lambda\alpha_2 + 1 - (\lambda)\beta_2) + (\lambda\alpha_3 + 1 - (\lambda)\beta_3)\\ &= \lambda(\alpha_1 + \alpha_2 + \alpha_3) + (1-\lambda)(\beta_1 + \beta_2 + \beta_3).\end{aligned}$$

Como $\alpha_1 + \alpha_2 + \alpha_3 = 1$ y $\beta_1 + \beta_2 + \beta_3 = 1$, la expresión anterior se reduce a $\lambda \cdot 1 + (1-\lambda)\cdot 1 = 1$, con lo que $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in S$. ◇

Ejercicio 1.2. Dado un conjunto cualquiera $S \subseteq \mathbb{R}^n$, demuestra que $\text{conv}(S)$ es un conjunto convexo. ◁

•**Ejercicio 1.3.** Dado un conjunto cualquiera $S \subseteq \mathbb{R}^n$, demuestra lo siguiente:

(i) $\text{conv}(S)$ es el conjunto convexo más pequeño que contiene a S.

(ii) $\text{conv}(S)$ es la intersección de todos los conjuntos convexos que contienen a S. ◁

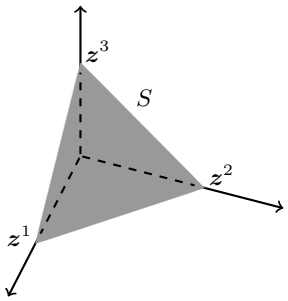


Figura 1.2: Representación gráfica del conjunto del Ejemplo 1.1.

A continuación presentamos algunos ejemplos de conjuntos convexos que nos serán de utilidad más adelante:

Hiperplano. El conjunto $S = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : 3x_1 + 2x_2 - x_3 = 4\}$ es un plano en $\mathbb{R}^3$. En general, un *hiperplano* en $\mathbb{R}^n$ se define mediante un vector no nulo $\boldsymbol{v} \in \mathbb{R}^n$ y un escalar $r \in \mathbb{R}$ como el conjunto $H(\boldsymbol{v}, r) = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{v}^\intercal \boldsymbol{x} = r\} = \{\boldsymbol{x} \in \mathbb{R}^n : \sum_{i=1}^n v_i x_i = r\}$.

El vector $\boldsymbol{v}$ se denomina *vector normal* por ser ortogonal al hiperplano que define. Esto se ve fácilmente en los ejes de coordenadas. En $\mathbb{R}^2$ el eje Y se corresponde con el hiperplano $H((1,0),0) = \{\boldsymbol{x} \in \mathbb{R}^2 : x_1 = 0\}$. El vector $(1,0)$ se llama vector normal porque es la única dirección ortogonal al hiperplano que define. En general, dado $\bar{\boldsymbol{x}} \in H(\boldsymbol{v}, r) = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{v}^\intercal \boldsymbol{x} = r\}$, como $\boldsymbol{v}^\intercal \bar{\boldsymbol{x}} = r$, $H(\boldsymbol{v}, r)$ también se puede definir como $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{v}^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}) = 0\}$ con lo que, efectivamente, $\boldsymbol{v}$ es perpendicular al hiperplano que define.

Semiespacio (cerrado). $S = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : 3x_1 + 2x_2 - x_3 \leq 4\}$ se corresponde con los puntos que se encuentran a un lado del hiperplano definido arriba. En general, un *semiespacio* en $\mathbb{R}^n$ es un conjunto de la forma $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{v}^\intercal \boldsymbol{x} \leq r\}$.

Dado un hiperplano $H(\boldsymbol{v}, r) = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{v}^\intercal \boldsymbol{x} = r\}$, se definen los *semiespacios superior e inferior* como $H(\boldsymbol{v}, r)^+ = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{v}^\intercal \boldsymbol{x} \geq r\}$ y $H(\boldsymbol{v}, r)^- = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{v}^\intercal \boldsymbol{x} \leq r\}$. Utilizando ahora que, dado $\bar{\boldsymbol{x}} \in H(\boldsymbol{v}, r)$, $H(\boldsymbol{v}, r)$ se puede escribir como $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{v}^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}) = 0\}$, tendremos también que $H(\boldsymbol{v}, r)^+ = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{v}^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}) \geq 0\}$ y $H(\boldsymbol{v}, r)^- = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{v}^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq 0\}$. Es decir, $H(\boldsymbol{v}, r)^+$ está compuesto por los puntos $\boldsymbol{x}$ tales que $\boldsymbol{x} - \bar{\boldsymbol{x}}$ forma con $\boldsymbol{v}$ un ángulo de 90° o menos (sea quien sea $\bar{\boldsymbol{x}} \in H(\boldsymbol{v}, r)$). Análogamente, para los puntos de $H(\boldsymbol{v}, r)^-$ tendremos un ángulo de 90° o más.

Poliedro. $S = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : 3x_1 + 2x_2 - x_3 \leq 4 \text{ y } 2x_1 + 4x_2 + x_3 \leq 1\}$. Un *poliedro* se define como una intersección finita de semiespacios. En general, dada una matriz $\boldsymbol{A}_{m \times n}$ y un vector $\boldsymbol{b} \in \mathbb{R}^m$, el conjunto $\{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$ es un poliedro convexo.

Politopo. $\text{conv}(\{(1,3,4),(3,5,2),(2,2,2)\})$. En general, un *politopo* se define como la envoltura convexa de una cantidad finita de puntos. Por definición, todo politopo es un conjunto acotado y es fácil demostrar que todo poliedro acotado es un politopo y viceversa. Los

dos dibujos de la izquierda en la Figura 1.3 muestran el mismo politopo, en un caso enfatizando los puntos extremos cuya envoltura convexa dan lugar al politopo y en el otro mediante los vectores normales a los semiespacios que lo definen.

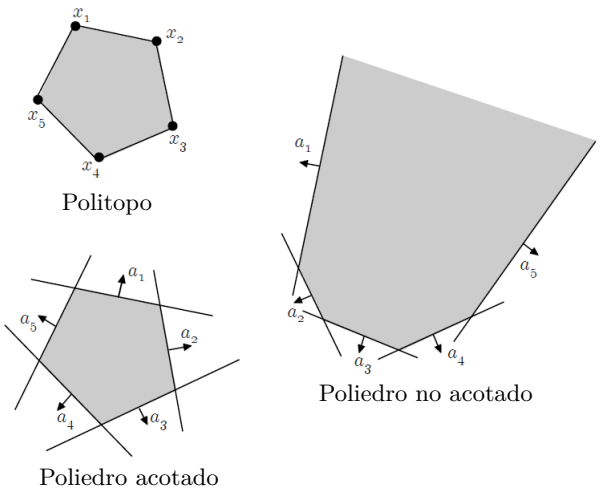


Figura 1.3: Ilustración de los conceptos de poliedro y politopo.

Una de las propiedades más importantes de los conjuntos convexos es que cualquier intersección de conjuntos convexos es un conjunto convexo, como el siguiente ejercicio pide demostrar junto con otras dos afirmaciones acerca de la conservación de la convexidad mediante ciertas operaciones entre conjuntos.

•**Ejercicio 1.4.** Dados dos conjuntos convexos $S_1 \in \mathbb{R}^n$ y $S_2 \in \mathbb{R}^n$, demuestra las siguientes afirmaciones:

(i) El conjunto $S_1 \cap S_2$ es convexo.

(ii) El conjunto $S_1 - S_2 = \{\boldsymbol{x} - \boldsymbol{y} : \boldsymbol{x} \in S_1,\ \boldsymbol{y} \in S_2\}$ es convexo.

(iii) El conjunto $S_1 + S_2 = \{\boldsymbol{x} + \boldsymbol{y} : \boldsymbol{x} \in S_1,\ \boldsymbol{y} \in S_2\}$ es convexo. ◁

Además, desde el punto de vista topológico tenemos que tanto el interior como la clausura de un conjunto convexo son conjuntos convexos.

••**Ejercicio 1.5.** Demuestra que, dado un conjunto convexo $S \subseteq \mathbb{R}^n$, los conjuntos $\mathring{S}$ y $\bar{S}$ son convexos. ◁

Para terminar este apartado proponemos otro ejercicio en el que se plantea una primera conexión entre el concepto de convexidad y la optimización.

•**Ejercicio 1.6.** Sea P el siguiente problema de programación lineal:

$$\begin{aligned} \text{minimizar} \quad & \boldsymbol{c}^\intercal \boldsymbol{x} \\ \text{sujeto a} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq 0, \end{aligned}$$

y sea $S = \{\boldsymbol{x} : \boldsymbol{x} \text{ solución óptima de } P\}$, demuestra que S es un conjunto convexo. ◁

1.2.2 Conos convexos

Este apartado está dedicado a los conos convexos, que resultan muy útiles para estudiar problemas de programación matemática en general y de programación lineal en particular. Antes de eso definimos el concepto de combinación cónica (la razón del nombre quedará clara con la definición formal de cono convexo). Dada una cantidad finita de vectores $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k$, una *combinación cónica* de estos puntos es de la forma $\sum_{i=1}^{k} \lambda_i \boldsymbol{x}^i$, donde los λ_i son no negativos. Obsérvese que la única diferencia con respecto a las combinaciones convexas es que no se pide que $\sum_{i=1}^{k} \lambda_i = 1$.

Definición 1.1. Un conjunto $C \subseteq \mathbb{R}^n$ es un *cono* si, para todo $\boldsymbol{x} \in C$ y todo $\lambda \geq 0$, el punto $\lambda \boldsymbol{x}$ pertenece a C. Un conjunto $C \subseteq \mathbb{R}^n$ es un *cono convexo* si es un cono y además es convexo.

Es fácil ver que la definición de cono convexo es equivalente a que cualquier combinación cónica de puntos de C pertenezca a C.

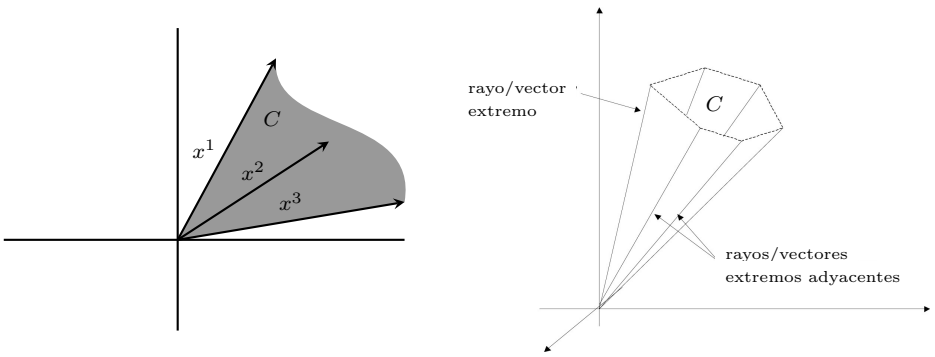


Figura 1.4: Ilustración del concepto de cono convexo.

Obsérvese que un cono convexo siempre contendrá al vector $\mathbf{0}$. Un ejemplo sencillo de cono convexo lo tenemos considerando los ejes cartesianos. Por ejemplo, el primer cuadrante $C = \{\boldsymbol{x} : x_1 \geq 0 \text{ y } x_2 \geq 0\}$ es un cono convexo. Siguiendo el paralelismo con los conjuntos convexos, también se puede definir la *envoltura cónica* de un conjunto S como el conjunto de todas las combinaciones cónicas de elementos de S. En el dibujo de la izquierda en la Figura 1.4 tenemos representado el cono convexo C, que es la envoltura cónica del conjunto $\{\boldsymbol{x}^1, \boldsymbol{x}^2, \boldsymbol{x}^3\}$, y también del conjunto $\{\boldsymbol{x}^1, \boldsymbol{x}^3\}$, ya que sólo los rayos/vectores extremos son necesarios para caracterizar un cono convexo. En el dibujo del lado derecho tenemos un cono convexo en tres dimensiones.

•• **Ejercicio 1.7.** Dado un conjunto cualquiera $S \subseteq \mathbb{R}^n$, demuestra lo siguiente:

(i) La envoltura cónica de S es el cono convexo más pequeño que contiene a S.

(ii) La envoltura cónica de S es la intersección de todos los conos convexos que contienen a S. ◁

El siguiente ejercicio pide demostrar que los conjuntos definidos como intersección de una cantidad finita de semiespacios tal que los hiperplanos que los definen pasan por el origen son conos convexos.

••**Ejercicio 1.8.** Demuestra que todo conjunto de la forma $C = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} \leq 0\}$ es un cono convexo. ◁

Un tipo especial de conos convexos son los denominados conos polares, que definimos a continuación.

Definición 1.2. Dado un conjunto no vacío $S \in \mathbb{R}^n$ definimos el *cono polar* asociado a S como el conjunto $S^* = \{\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{v}^\intercal \boldsymbol{x} \leq 0 \text{ para todo } \boldsymbol{x} \in S\}$.

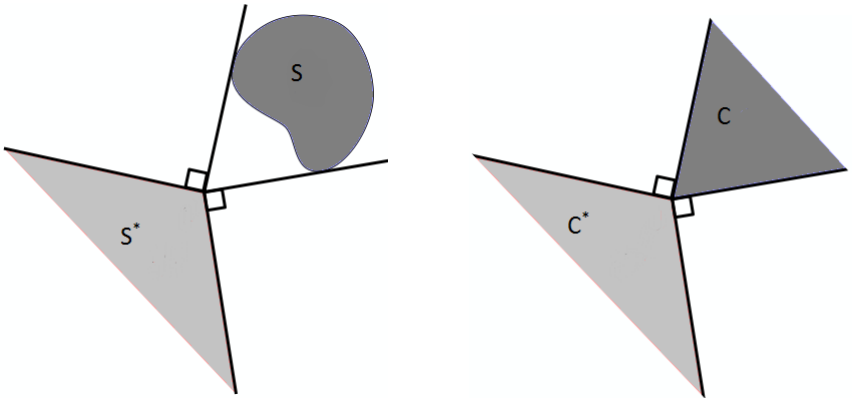


Figura 1.5: Ilustración del concepto de cono polar.

El cono polar asociado a un conjunto S admite una interpretación geométrica muy sencilla: un vector $\boldsymbol{v}$ está en el cono polar asociado a un conjunto S si forma un ángulo de al menos 90° con todos los elementos del conjunto S. Esto lo ilustramos en la Figura 1.5, donde vemos el cono polar de un conjunto cualquiera S y el cono polar de un cono convexo C.

A la vista de la Figura 1.5, resulta bastante claro que un cono polar será un cono convexo. El Ejercicio 1.10 nos pide probar que esto es cierto, entre otras afirmaciones.

Ejercicio 1.9. Representa gráficamente los siguientes conjuntos y sus conos polares:

(i) $\{(x_1, x_2) : 0 \leq x_2 \leq 2x_1\}$.

(ii) $\{(x_1, x_2) : x_2 \leq -3|x_1|\}$.

◁

•**Ejercicio 1.10.** Demuestra que, dados S, S_1 y S_2 conjuntos no vacíos en $\mathbb{R}^n$, entonces las siguientes afirmaciones son ciertas:

(i) S^* es un cono convexo y cerrado.

(ii) $S \subseteq S^{**}$, donde S^{**} es el cono polar de S^*.

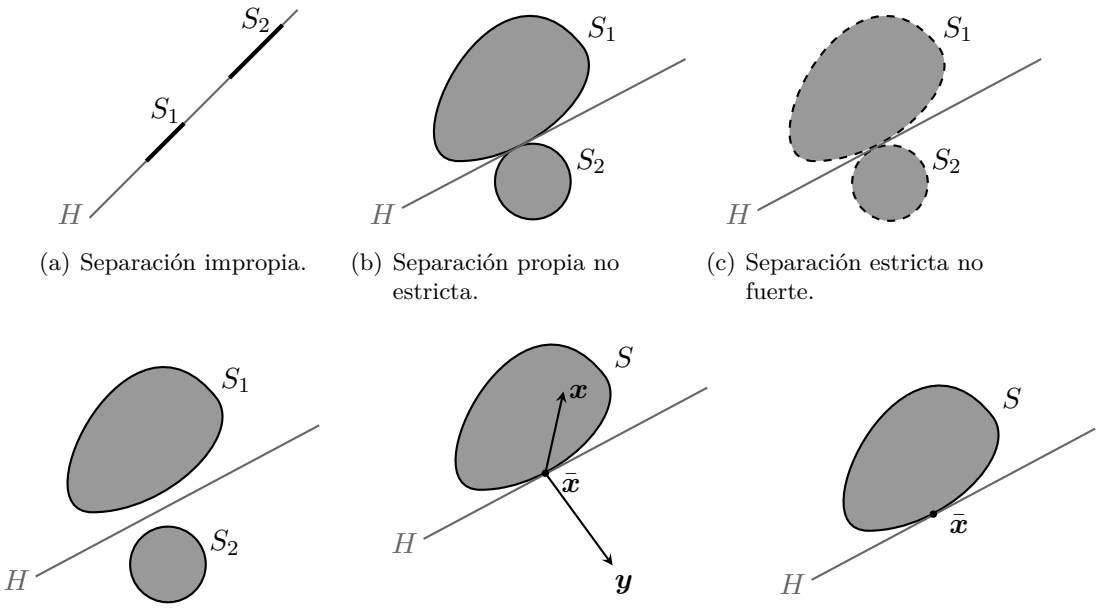
(iii) Si $S_1 \subseteq S_2$, entonces $S_2^* \subseteq S_1^*$.

◁

1.2.3 Teoremas de separación y de hiperplano soporte

A continuación presentamos una serie de resultados, conocidos habitualmente como teoremas de separación, que son de gran importancia a la hora de trabajar con la geometría subyacente a problemas de optimización. En particular, suponen el camino habitual para probar el Lema de Farkas que veremos en la Sección 1.2.4.

Dados dos conjuntos S_1 y S_2, decimos que $H(\boldsymbol{v}, r)$ es un *hiperplano de separación* para S_1 y S_2 si $S_1 \subseteq H(\boldsymbol{v}, r)^+ = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{v}^\top \boldsymbol{x} \geq r\}$ y $S_2 \subseteq H(\boldsymbol{v}, r)^- = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{v}^\top \boldsymbol{x} \leq r\}$. Si para todo $\boldsymbol{x} \in S_1$ se tiene $\boldsymbol{v}^\top \boldsymbol{x} > r$ y para todo $\boldsymbol{x} \in S_2$ se tiene $\boldsymbol{v}^\top \boldsymbol{x} < r$, tenemos *separación estricta*. Si además existe $\varepsilon > 0$ tal que para todo $\boldsymbol{x} \in S_1$ se tiene $\boldsymbol{v}^\top \boldsymbol{x} > r + \varepsilon$ y para todo $\boldsymbol{x} \in S_2$ se tiene $\boldsymbol{v}^\top \boldsymbol{x} \leq r$, tenemos *separación fuerte*. Resumiendo, separación implica que S_1 está en uno de los semiespacios definidos por el hiperplano y S_2 está en el otro; en particular, podemos tener que tanto S_1 como S_2 están contenidos en el hiperplano, lo que se conoce como *separación impropia*. Separación estricta añade que ninguno de los dos conjuntos tiene puntos en el propio hiperplano lo que en particular implica que $S_1 \cap S_2 = \emptyset$. Separación fuerte implica además que la distancia entre los dos conjuntos es mayor que 0. La Figura 1.6 muestra distintos tipos de separación, ordenados de menos fuerte a más fuerte.



(a) Separación impropia. (b) Separación propia no estricta. (c) Separación estricta no fuerte.

(d) Separación fuerte. (e) Illustración del Teorema 1.2. (f) Hiperplano soporte.

Figura 1.6: Hiperplanos de separación e hiperplanos soporte.

Teorema 1.2 (Separación de un punto y un conjunto convexo)**.** *Sean $S \subseteq \mathbb{R}^n$ un conjunto no vacío, convexo y cerrado y sea $\boldsymbol{y} \in \mathbb{R}^n \backslash S$. Entonces existe un hiperplano que separa fuertemente $\{\boldsymbol{y}\}$ de S.*

Demostración. Consideremos el siguiente problema de minimización: $\inf_{\boldsymbol{x} \in S} \|\boldsymbol{y} - \boldsymbol{x}\|$. Entonces,

existe $\bar{\boldsymbol{x}} \in S$ tal que $\|\boldsymbol{y} - \bar{\boldsymbol{x}}\| = \inf_{\boldsymbol{x}\in S}\|\boldsymbol{y} - \boldsymbol{x}\|$.[2]

Como $\bar{\boldsymbol{x}} \in S$, $\bar{\boldsymbol{x}} \neq \boldsymbol{y}$. Dado que S es convexo, para todo $\boldsymbol{x} \in S$ y todo $\lambda \in (0, 1]$, tenemos $\lambda\boldsymbol{x}+(1-\lambda)\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}+\lambda(\boldsymbol{x}-\bar{\boldsymbol{x}}) \in S$. Por tanto, $\|\boldsymbol{y}-\bar{\boldsymbol{x}}\| \leq \|\boldsymbol{y}-(\bar{\boldsymbol{x}}+\lambda(\boldsymbol{x}-\bar{\boldsymbol{x}}))\| = \|(\boldsymbol{y}-\bar{\boldsymbol{x}})-\lambda(\boldsymbol{x}-\bar{\boldsymbol{x}})\|$. Ya que para cada $\boldsymbol{v} \in \mathbb{R}^n$, $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^\intercal\boldsymbol{v}$, tenemos $(\boldsymbol{y}-\bar{\boldsymbol{x}})^\intercal(\boldsymbol{y}-\bar{\boldsymbol{x}}) \leq ((\boldsymbol{y}-\bar{\boldsymbol{x}})-\lambda(\boldsymbol{x}-\bar{\boldsymbol{x}}))^\intercal((\boldsymbol{y}-\bar{\boldsymbol{x}})-\lambda(\boldsymbol{x}-\bar{\boldsymbol{x}}))$. Entonces, $2\lambda(\boldsymbol{y}-\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x}-\bar{\boldsymbol{x}}) \leq \lambda^2(\boldsymbol{x}-\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x}-\bar{\boldsymbol{x}})$. Ahora, dividiendo por λ y haciendo tender λ a 0 tenemos que la desigualdad $(\boldsymbol{y} - \bar{\boldsymbol{x}})^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq 0$ se cumple para todo $\boldsymbol{x} \in S$; equivalentemente, $(\boldsymbol{y}-\bar{\boldsymbol{x}})$ y $(\boldsymbol{x}-\bar{\boldsymbol{x}})$ forman un ángulo de al menos 90° (ver Figura 1.6(e)). Por tanto, $\bar{\boldsymbol{x}}^\intercal(\boldsymbol{y}-\bar{\boldsymbol{x}}) \geq \boldsymbol{x}^\intercal(\boldsymbol{y}-\bar{\boldsymbol{x}})$ y, además, como $\bar{\boldsymbol{x}} \neq \boldsymbol{y}$, $\|\boldsymbol{y}-\bar{\boldsymbol{x}}\| = (\boldsymbol{y}-\bar{\boldsymbol{x}})^\intercal(\boldsymbol{y}-\bar{\boldsymbol{x}}) > 0$, tenemos que $\bar{\boldsymbol{x}}^\intercal(\boldsymbol{y} - \bar{\boldsymbol{x}}) < \boldsymbol{y}^\intercal(\boldsymbol{y} - \bar{\boldsymbol{x}})$.

Definamos ahora $\boldsymbol{v} = (\boldsymbol{y} - \bar{\boldsymbol{x}}) \neq \boldsymbol{0}$ y $r = \bar{\boldsymbol{x}}^\intercal(\boldsymbol{y} - \bar{\boldsymbol{x}})$. De las desigualdades anteriores se sigue que, para cada $\boldsymbol{x} \in S$, $\boldsymbol{v}^\intercal\boldsymbol{x} \leq r < \boldsymbol{v}^\intercal\boldsymbol{y}$, con lo que $H(\boldsymbol{v}, r)$ separa fuertemente $\{\boldsymbol{y}\}$ y S.[3] □

Corolario 1.3. *Si $S \subseteq \mathbb{R}^n$ es un conjunto convexo y cerrado, entonces S es la intersección de todos los semiespacios que lo contienen.*

•• **Ejercicio 1.11.** Demuestra el Corolario 1.3. ◁

A continuación presentamos la noción de hiperplano soporte y el Teorema del hiperplano soporte, que es una consecuencia prácticamente inmediata del teorema anterior. Dados $S \subseteq \mathbb{R}^n$ y $\bar{\boldsymbol{x}} \in \partial S$, decimos que un hiperplano $H(\boldsymbol{v}, r)$ es un hiperplano soporte de S en $\bar{\boldsymbol{x}}$ si $\bar{\boldsymbol{x}} \in H(\boldsymbol{v}, r)$ y $S \subseteq H(\boldsymbol{v}, r)^+$ o $S \subseteq H(\boldsymbol{v}, r)^-$. En particular, dicho hiperplano soporte separa $\bar{\boldsymbol{x}}$ y $\bar{S}$ (aunque no estrictamente).

Teorema 1.4 (Teorema del hiperplano soporte)**.** *Dado un conjunto no vacío y convexo $S \subseteq \mathbb{R}^n$ y $\bar{\boldsymbol{x}} \in \partial S$, entonces existe un hiperplano soporte de S en $\bar{\boldsymbol{x}}$.*

Demostración. Como $\bar{\boldsymbol{x}} \in \partial S$, existe una sucesión $\{\boldsymbol{x}^t\} \subseteq \mathbb{R}^n\backslash\bar{S}$ tal que $\{\boldsymbol{x}^t\} \to \bar{\boldsymbol{x}}$. Como $\bar{S}$ es no vacío, cerrado y convexo, para cada $t \in \mathbb{N}$ podemos aplicar el Teorema 1.2. Entonces, existen $\{\boldsymbol{v}^t\} \subseteq \mathbb{R}^n\backslash\boldsymbol{0}$ y $\{r^t\} \subseteq \mathbb{R}$ tales que, para cada $t \in \mathbb{N}$, $H(\boldsymbol{v}^t, r^t)$ separa fuertemente $\{\boldsymbol{x}^t\}$ y $\bar{S}$. Sin pérdida de generalidad podemos tomar $\{\boldsymbol{v}^t\}$ tal que, para todo $t \in \mathbb{N}$, $\|\boldsymbol{v}^t\| = 1$. Entonces, $\{\boldsymbol{v}^t\} \subseteq \mathbb{R}^n$ tiene una sucesión convergente con límite $\boldsymbol{v}$, con $\|\boldsymbol{v}\| = 1$. Para cada t de la subsucesión y cada $\boldsymbol{x} \in \bar{S}$ tenemos que $(\boldsymbol{v}^t)^\intercal\boldsymbol{x}^t > (\boldsymbol{v}^t)^\intercal\boldsymbol{x}$. Fijando un $\boldsymbol{x} \in \bar{S}$ y llevando la subsucesión al límite tenemos que $\boldsymbol{v}^\intercal\bar{\boldsymbol{x}} \geq \boldsymbol{v}^\intercal\boldsymbol{x}$. Como esto es cierto para todo $\boldsymbol{x} \in \bar{S}$ podemos tomar $r = \boldsymbol{v}^\intercal\bar{\boldsymbol{x}}$ tenemos que $H(\boldsymbol{v}, r)$ es un hiperplano soporte de S en $\bar{\boldsymbol{x}}$. □

Corolario 1.5. *Dado un conjunto no vacío y convexo $S \subseteq \mathbb{R}^n$ y $\bar{\boldsymbol{x}} \notin \mathring{S}$, entonces existe un hiperplano separando $\bar{\boldsymbol{x}}$ y $\bar{S}$.*

Demostración. Si $\bar{\boldsymbol{x}} \notin \bar{S}$, el corolario se sigue del Teorema 1.2. por otro lado, si $\bar{\boldsymbol{x}} \in \partial S$, el corolario se sigue del Teorema 1.4. □

[2] Siendo formales, la existencia de $\bar{\boldsymbol{x}}$ se sigue del Teorema de Weierstrass, que nos asegura que toda función continua definida sobre un compacto alcanza sus valores máximo y mínimo. Como S no tiene por qué ser un compacto, bastaría con tomar una bola cerrada B centrada en $\boldsymbol{y}$ y que interseque a S. Entonces, el conjunto $S \cap B$ es compacto y los puntos de S que no pertenecen a este compacto están a mayor distancia de $\boldsymbol{y}$ que cualquier punto de B. Entonces, $\inf_{\boldsymbol{x}\in S}\|\boldsymbol{y} - \boldsymbol{x}\| = \inf_{\boldsymbol{x}\in S\cap B}\|\boldsymbol{y} - \boldsymbol{x}\|$.

[3] Para la separación fuerte podemos tomar $\varepsilon = (r + \boldsymbol{v}^\intercal\boldsymbol{y})/2$, con lo que $\boldsymbol{v}^\intercal\boldsymbol{x} < r + \varepsilon < \boldsymbol{v}^\intercal\boldsymbol{y}$.

Teorema 1.6 (Separación de conjuntos convexos)**.** *Dados $S_1 \subseteq \mathbb{R}^n$ y $S_2 \subseteq \mathbb{R}^n$ dos conjuntos no vacíos, disjuntos y convexos, entonces existe un hiperplano separando S_1 y S_2.*

Demostración. Sea $S = S_1 - S_2 = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \boldsymbol{x}^1 - \boldsymbol{x}^2, \text{ con } \boldsymbol{x}^1 \in S \text{ y } \boldsymbol{x}^2 \in S_2\}$. La convexidad de S_1 y S_2 implica la convexidad de S (véase Ejercicio 1.4). Como $S_1 \cap S_2 = \emptyset$, $\mathbf{0} \notin S$. Entonces, por el Corolario 1.5, existen $\boldsymbol{v} \in \mathbb{R}^n \backslash \{0\}$ y $r \in \mathbb{R}$ tales que $H(\boldsymbol{v}, r)$ separa $\mathbf{0}$ y S. Por tanto, para cada $\boldsymbol{x}^1 \in S_1$ y cada $\boldsymbol{x}^2 \in S_2$, $\boldsymbol{v}^\intercal(\boldsymbol{x}^1 - \boldsymbol{x}^2) \geq r \geq \boldsymbol{v}^\intercal \mathbf{0} = \mathbf{0}$, con lo que $\boldsymbol{v}^\intercal \boldsymbol{x}^1 \geq \boldsymbol{v}^\intercal \boldsymbol{x}^2$.

Tomando ahora $r = \text{ínf}\{\boldsymbol{v}^\intercal \boldsymbol{x} : \boldsymbol{x} \in S_1\} \geq \sup\{\boldsymbol{v}^\intercal \boldsymbol{x} : \boldsymbol{x} \in S_2\}$ tenemos que $H(\boldsymbol{v}, r)$ separa S_1 y S_2. □

1.2.4 Lema de Farkas

A continuación presentamos el Lema de Farkas, también conocido como el Teorema de alternativa. Este resultado y sus múltiples variantes son de gran relevancia para el estudio teórico de problemas de programación matemática, especialmente a la hora de derivar condiciones de optimalidad. En particular, el Teorema de dualidad fuerte en programación lineal se puede obtener de manera relativamente sencilla a partir de distintas variantes del Lema de Farkas.

Previamente introducimos un par de resultados auxiliares que serán de utilidad a la hora de interpretar geométricamente el Lema de Farkas.

Proposición 1.7. *Si C es un cono convexo, cerrado y no vacío, entonces $C = C^{**}$.*

Demostración. La inclusión $C \subseteq C^{**}$ es prácticamente inmediata y es cierta sea o no sea C un cono convexo (véase el Ejercicio 1.10).

Supongamos ahora que existe $\boldsymbol{x} \in C^{**} \backslash C$. Como $\boldsymbol{x} \notin C$, por el Teorema 1.2, existe un hiperplano $H(\boldsymbol{v}, r)$ que separa fuertemente $\{\boldsymbol{x}\}$ de C. Por tanto, $\boldsymbol{v}^\intercal \boldsymbol{x} > r$ y, para todo $\boldsymbol{y} \in C$, $\boldsymbol{v}^\intercal \boldsymbol{y} \leq r$. Como $\boldsymbol{y} = \mathbf{0} \in C$, $r \geq 0$ y $\boldsymbol{v}^\intercal \boldsymbol{x} > 0$.

Veamos que $\boldsymbol{v} \in C^*$. En caso contrario existiría $\bar{\boldsymbol{y}} \in C$ tal que $\boldsymbol{v}^\intercal \bar{\boldsymbol{y}} > 0$. Como C es un cono convexo, $\lambda \bar{\boldsymbol{y}} \in C$ para todo $\lambda > 0$, con lo que podemos hacer $\boldsymbol{v}^\intercal(\lambda \bar{\boldsymbol{y}})$ arbitrariamente grande. Esto contradice que $\boldsymbol{v}^\intercal \boldsymbol{y} \leq r$ para todo $\boldsymbol{y} \in C$. Por tanto, $\boldsymbol{v} \in C^*$.

Entonces, como $\boldsymbol{x} \in C^{**} = \{\boldsymbol{z} : \boldsymbol{z}^\intercal \boldsymbol{w} \leq 0 \text{ para todo } \boldsymbol{w} \in C^*\}$, $\boldsymbol{x}^\intercal \boldsymbol{v} = \boldsymbol{v}^\intercal \boldsymbol{x} \leq 0$, lo que contradice que $\boldsymbol{v}^\intercal \boldsymbol{x} > 0$. Por tanto, no es posible que $\boldsymbol{x} \notin C$. □

Proposición 1.8. *Tomemos $\boldsymbol{A}_{m \times n}$. Dado el cono $C = \{\boldsymbol{A}^\intercal \boldsymbol{y} : \boldsymbol{y} \in \mathbb{R}^m, \boldsymbol{y} \geq \mathbf{0}\}$, entonces el cono polar de C, C^*, viene dado por:*

$$\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} \leq \mathbf{0}\}.$$

Demostración. Tomemos $\boldsymbol{x} \in \mathbb{R}^n$ e $\boldsymbol{y} \geq \mathbf{0}$. Entonces, $\boldsymbol{v} = \boldsymbol{A}^\intercal \boldsymbol{y} \in C$ y tenemos $\boldsymbol{x}^\intercal \boldsymbol{v} = \boldsymbol{x}^\intercal \boldsymbol{A}^\intercal \boldsymbol{y} = \boldsymbol{y}^\intercal \boldsymbol{A}\boldsymbol{x}$. Ahora pueden pasar dos cosas:

- Si $\boldsymbol{A}\boldsymbol{x} \leq \mathbf{0}$ entonces, para todo $\boldsymbol{y} \geq \mathbf{0}$ o, equivalentemente, para todo $\boldsymbol{v} \in C$, tenemos que $\boldsymbol{x}^\intercal \boldsymbol{v} = \boldsymbol{x}^\intercal \boldsymbol{A}^\intercal \boldsymbol{y} = \boldsymbol{y}^\intercal \boldsymbol{A}\boldsymbol{x} \leq 0$ y $\boldsymbol{x} \in C^*$.
- Si existe i tal que $(\boldsymbol{A}\boldsymbol{x})_i > 0$, entonces podemos tomar $\boldsymbol{y} \geq \mathbf{0}$ con y_i tan grande como sea preciso para obtener $\boldsymbol{x}^\intercal \boldsymbol{v} = \boldsymbol{x}^\intercal \boldsymbol{A}^\intercal \boldsymbol{y} = \boldsymbol{y}^\intercal \boldsymbol{A}\boldsymbol{x} > 0$, con lo que $\boldsymbol{x} \notin C^*$. □

Teorema 1.9 (Lema de Farkas)**.** *Dada una matriz $\boldsymbol{A}_{m\times n}$ y un vector $\boldsymbol{c}\in\mathbb{R}^n$, entonces uno y sólo uno de los dos siguientes sistemas tiene solución:*

Sistema 1. $\boldsymbol{Ax}\leq\boldsymbol{0}$ *y* $\boldsymbol{c}^\intercal\boldsymbol{x}>0$ *para algún* $\boldsymbol{x}\in\mathbb{R}^n$.

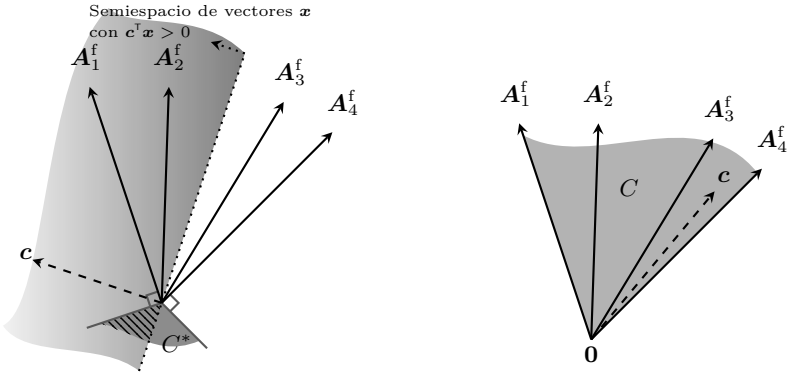
Sistema 2. $\boldsymbol{A}^\intercal\boldsymbol{y}=\boldsymbol{c}$ *e* $\boldsymbol{y}\geq\boldsymbol{0}$ *para algún* $\boldsymbol{y}\in\mathbb{R}^m$.

Antes de demostrar el Lema de Farkas vamos a presentar su **interpretación geométrica**, ilustrada gráficamente en la Figura 1.7:

Sistema 2. Si $\boldsymbol{A}_1^{\mathrm{f}},\ldots,\boldsymbol{A}_m^{\mathrm{f}}$ son las filas de $\boldsymbol{A}$ (columnas de $\boldsymbol{A}^\intercal$), el Sistema 2 tiene solución si $\boldsymbol{c}$ pertenece al cono convexo, C, generado por dichas filas (su envoltura cónica).

Sistema 1. El Sistema 1 tiene solución si existe algún $\boldsymbol{x}$ que forme un ángulo mayor o igual de 90° con todos los vectores que forman las filas de $\boldsymbol{A}$ y, simultáneamente, un ángulo menor de 90° con el vector $\boldsymbol{c}$.

Equivalentemente, apoyándonos en la Proposición 1.8, tenemos que el Sistema 1 tiene solución si existe un vector $\boldsymbol{x}$ en el cono polar de C que forme un ángulo menor de 90° con el vector $\boldsymbol{c}$.



(a) El Sistema 1 tiene solución. (b) El Sistema 2 tiene solución.

Figura 1.7: Ilustración de los dos sistemas del enunciado del Lema de Farkas.

Es importante destacar que de esta última interpretación prácticamente nos proporciona una demostración del Lema de Farkas. Dicho lema se puede reformular de la siguiente manera: "O $\boldsymbol{c}$ está en el cono C o $\boldsymbol{c}$ forma un ángulo de menos de 90° con algún elemento de C^*. Además, sólo una de las dos situaciones anteriores es posible." Pero con esta formulación el resultado es casi inmediato. Si $\boldsymbol{c}\in C$, por definición de C^*, $\boldsymbol{c}$ formará un ángulo mayor o igual a 90° con cualquier $\boldsymbol{x}\in C^*$. Por otro lado, si $\boldsymbol{c}\notin C$, como $C=C^{**}$ (Proposición 1.7), entonces $\boldsymbol{c}\notin C^{**}$ y por tanto $\boldsymbol{c}$ formará un ángulo de menos de 90° con algún elemento de C^*.

Nótese que el argumento anterior se apoya implícitamente en los teoremas de separación, ya que la Proposición 1.7 se demostró usando el Teorema 1.2. A continuación presentamos una demostración directa, que no se sustenta en las nociones de cono convexo y cono polar, pero quizá un poco más difícil de seguir geométricamente.

Demostración del Lema de Farkas. Para empezar, supongamos que el Sistema 2 tiene solución. Entonces, existe $\boldsymbol{y} \in \mathbb{R}^m$ tal que $\boldsymbol{A}^\intercal \boldsymbol{y} = \boldsymbol{c}$ e $\boldsymbol{y} \geq \boldsymbol{0}$. Tomemos ahora $\boldsymbol{x} \in \mathbb{R}^n$ tal que $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{0}$. Entonces, $\boldsymbol{c}^\intercal \boldsymbol{x} = \boldsymbol{y}^\intercal \boldsymbol{A}\boldsymbol{x} \leq 0$, pues $\boldsymbol{y} \geq \boldsymbol{0}$ y $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{0}$. Por tanto, el Sistema 1 no tiene solución.

Supongamos que el Sistema 2 no tiene solución y probemos que entonces el Sistema 1 sí la tiene. Tomemos el cono convexo generado por las columnas de $\boldsymbol{A}^\intercal$, $C = \{\boldsymbol{x} : \boldsymbol{x} = \boldsymbol{A}^\intercal \boldsymbol{y}, \boldsymbol{y} \geq \boldsymbol{0}\}$. C es convexo y cerrado y $\boldsymbol{c} \notin C$. Por el Teorema 1.2, existe un hiperplano $H(\boldsymbol{v}, r)$ que separa fuertemente $\{\boldsymbol{c}\}$ de C. Por tanto, $\boldsymbol{v}^\intercal \boldsymbol{c} > r$ y, para todo $\boldsymbol{x} \in C$, $\boldsymbol{v}^\intercal \boldsymbol{x} \leq r$. Como $\boldsymbol{0} \in C$ y $\boldsymbol{v}^\intercal \boldsymbol{0} = 0$, tenemos que $r \geq 0$, con lo que $\boldsymbol{v}^\intercal \boldsymbol{c} > 0$. Además, representando cada $\boldsymbol{x} \in C$ mediante la combinación cónica de la que procede tenemos que, para todo $\boldsymbol{y} \geq \boldsymbol{0}$, $r \geq \boldsymbol{v}^\intercal \boldsymbol{A}^\intercal \boldsymbol{y} = \boldsymbol{y}^\intercal \boldsymbol{A}\boldsymbol{v}$. Como las componentes de $\boldsymbol{y}$ se pueden hacer arbitrariamente grandes, la desigualdad anterior implica que $\boldsymbol{A}\boldsymbol{v} \leq 0$. Como $\boldsymbol{A}\boldsymbol{v} \leq 0$ y $\boldsymbol{v}^\intercal \boldsymbol{c} > 0$, $\boldsymbol{v} \in \mathbb{R}^n$ es una solución del Sistema 1. □

Si miramos nuevamente el enunciado del Lema de Farkas, no debería sorprendernos que tenga implicaciones en el estudio de la dualidad en programación lineal. El Sistema 1 recuerda a la región factible del primal con una condición sobre la función objetivo. Por otro lado, el Sistema 2 usa la traspuesta de la matriz del Sistema 1 y aparece el "vector de costes" del primal en el lado derecho, lo que claramente apunta al dual. El siguiente ejercicio pide que se exploten estas conexiones para demostrar el Teorema de dualidad fuerte de la programación lineal. En el Tema 5 nos apoyaremos en una variante del Lema de Farkas para probar formalmente el Teorema de dualidad fuerte en programación no lineal, que en particular implica el resultado para programación lineal.

•••**Ejercicio 1.12.** Usa el Lema de Farkas o alguna variante del mismo para probar el Teorema de dualidad fuerte de la programación lineal, que dice lo siguiente:

Dados un par primal-dual de problemas de programación lineal, P y D, si alguno de los dos tiene un óptimo finito, entonces también el otro lo tiene. Además, los valores óptimos de ambos problemas coinciden y tanto P como D tienen soluciones óptimas. ◁

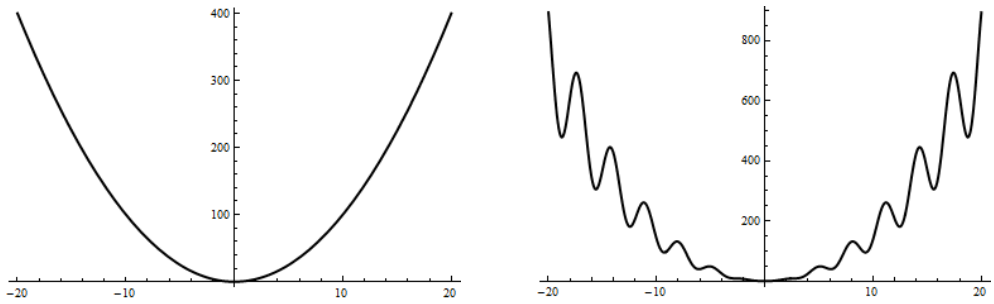
1.3 Funciones convexas y propiedades

1.3.1 Discusión previa

Supongamos que queremos minimizar una función $f : \mathbb{R}^n \to \mathbb{R}$. Pensemos, por ejemplo, en la función real $f(x) = x^2$. ¿Qué opciones tendríamos para buscar el mínimo?

- Podríamos representar la función gráficamente y encontrar así el mínimo. Aunque esto serviría con la función $f(x) = x^2$, este método no sería factible para funciones definidas sobre $\mathbb{R}^n$ con $n > 3$.
- Otra opción sería buscar un punto donde se anule la derivada. Nuevamente, este método serviría para una función como $f(x) = x^2$, pero este enfoque para funciones más generales tiene importantes limitaciones:
  - Buscar puntos donde se anula la derivada (o el gradiente) parten de asumir que la función es diferenciable. Por ejemplo, no nos permitiría encontrar el mínimo de la función $f(x) = |x|$.

– Además, incluso en el caso de funciones diferenciables, las condiciones sobre el gradiente son simplemente condiciones locales, y podemos encontrarnos con mínimos locales con propiedades muy malas. Por ejemplo, la función $f(x) = x^2\big(\text{sen}(x)^2 + \frac{x^2}{100} + 1\big)$, representada en la Figura 1.8(b), tiene infinitos *mínimos locales* y sólo uno de ellos, $x = 0$, es un *mínimo global*.



(a) Función $f(x) = x^2$. (b) Función $f(x) = x^2\big(\text{sen}(x)^2 + \frac{x^2}{100} + 1\big)$.

Figura 1.8: Unicidad y multiplicidad de óptimos locales.

- Si queremos un método que no necesite trabajar con funciones diferenciables también podemos hacer un mallado de la región factible, evaluar la función en todos los puntos de esa malla, y quedarnos con el mejor. El mayor problema de este natural método es que dicha malla sería imposible de definir en problemas donde el dominio de definición de la función F es un conjunto no acotado y, aún siendo dicho dominio un conjunto acotado, estos mallados resultan imposibles de llevar a cabo cuando la dimensión crece. Por ejemplo, para hacer un mallado en el conjunto $[-1, 1]^{10}$ de una precisión comparable a lo que se obtiene al tomar 100 puntos sobre el intervalo $[-1, 1]$, deberíamos tener una malla con 100^{10} puntos. Es decir, el número de puntos necesarios en la malla crece exponencialmente con la dimensión del espacio de definición de la función, lo que hace que, en general, la optimización en base a mallados sea prohibitiva computacionalmente.

- Otra opción sería trabajar con el concepto de *dirección de descenso*. La idea sería partir de un punto cualquiera $\boldsymbol{x}$ del dominio de definición de la función f y buscar alguna dirección $\boldsymbol{d}$ en la que la función decrezca. Esto nos permite definir un método iterativo que consistiría en moverse dentro del dominio de f siguiendo direcciones de descenso hasta que encontremos algún punto para el que no existen direcciones de descenso, lo que aseguraría la optimalidad (local) de dicho punto.

  Aunque la idea de este enfoque es muy natural, la búsqueda de buenas direcciones de descenso también puede resultar una tarea compleja. Una herramienta muy conveniente para llevarla a cabo es el concepto de *derivada direccional*. Por supuesto, cualquier función diferenciable tendrá las derivadas direccionales bien definidas, pero hay muchas funciones no diferenciables para las cuales también están bien definidas. Por ejemplo, la función $f(x) = |x|$ no es diferenciable pero sí tiene derivadas direccionales.

Como resumen de esta discusión podemos decir que, a la ahora de minimizar una función, hay que tener cuidado con la existencia de óptimos locales que no sean óptimos globales y, por otro lado, la existencia de derivadas direccionales puede resultar de gran utilidad. A continuación presentamos la clase de las *funciones convexas* y demostraremos que dicha clase tiene un buen comportamiento con respecto a los dos aspectos que acabamos de comentar.

1.3.2 Definiciones básicas

En esta sección desarrollaremos algunas propiedades fundamentales de las funciones convexas y de las funciones cóncavas.

Dado un conjunto no vacío y convexo $S \subseteq \mathbb{R}^n$ y una función $f : S \to \mathbb{R}$, la función f es *convexa* en S si, para todo $\boldsymbol{x} \in S$ e $\boldsymbol{y} \in S$ y para todo $\lambda \in (0, 1)$, se tiene

$$f(\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1 - \lambda) f(\boldsymbol{y}). \tag{1.1}$$

Además, si la anterior desigualdad es estricta siempre que $\boldsymbol{x} \neq \boldsymbol{y}$, entonces f es *estrictamente convexa*. Una función $f : S \to \mathbb{R}$ es cóncava (estrictamente cóncava) si $-f$ es convexa (estrictamente convexa) en S; equivalentemente, una función es cóncava si en la Ecuación (1.1) reemplazamos "$\leq$" por "$\geq$". Una función que es al mismo tiempo convexa y cóncava se llama *afín* y se puede expresar como $f(\boldsymbol{x}) = a\boldsymbol{x} + b$ con $a \in \mathbb{R}$ y $b \in \mathbb{R}$.

Geométricamente, que una función sea convexa simplemente nos dice que, dados dos puntos de S, el grafo de la función entre dichos puntos está por debajo de la recta que une sus imágenes (por encima de dicha recta en el caso de funciones cóncavas). La Figura 1.9 ilustra esta idea. Nótese que es importante que la función f esté definida sobre un conjunto convexo, pues de otro modo no tendríamos asegurado que el punto $\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y}$ pertenece a S. Por otro lado, hay funciones que pueden ser convexas sólo sobre una parte de su dominio de definición. Por ejemplo $f(x) = x^3$ no es convexa sobre $\mathbb{R}$ pero sí sobre $S = \{x : x \geq 0\}$.

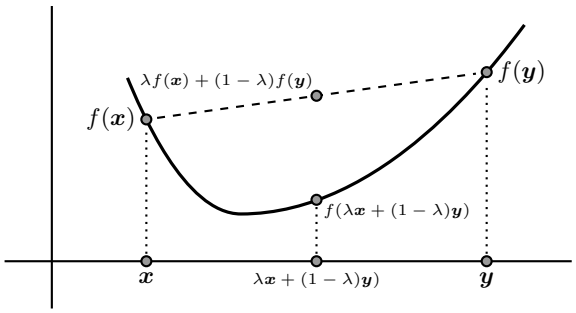


Figura 1.9: Una función convexa.

••Ejercicio 1.13. Demuestra que, dado $S \subseteq \mathbb{R}^n$, si una función $f : S \to \mathbb{R}$ es a la vez cóncava y convexa, entonces existen $\boldsymbol{a} \in \mathbb{R}^n$ y $b \in \mathbb{R}$ de manera que, para todo $\boldsymbol{x} \in S$, $f(\boldsymbol{x}) = \boldsymbol{a}^\intercal \boldsymbol{x} + b$. ◁

••Ejercicio 1.14. Demuestra que las siguientes funciones son convexas:

- $f(x) = 3x + 4$, con $x \in \mathbb{R}$.
- $f(x) = |x|$, con $x \in \mathbb{R}$.
- $f(x) = x^2 - x$, con $x \in \mathbb{R}$.
- $f(x) = -x^{1/3}$, con $x \geq 0$.
- $f(\boldsymbol{x}) = 2x_1^2 + x_2^2 - 2x_1x_2$ con $\boldsymbol{x} \in \mathbb{R}^2$.
- $f(\boldsymbol{x}) = x_1^4 + 2x_2^2 + 3x_3^2 - 4x_1 - 4x_2x_3$ con $\boldsymbol{x} \in \mathbb{R}^3$. ◁

El siguiente resultado presenta una primera lista de propiedades de las funciones convexas.

Proposición 1.10.

(i) *Sean $f_1, f_2, \ldots, f_k : \mathbb{R}^n \to \mathbb{R}$ funciones convexas y sean $\alpha_1, \alpha_2, \ldots, \alpha_k$ coeficientes no negativos, entonces la siguiente función es convexa:*

$$f(\boldsymbol{x}) = \sum_{i=1}^{k} \alpha_i f_i(\boldsymbol{x}).$$

(ii) *Sean $f_1, f_2, \ldots, f_k : \mathbb{R}^n \to \mathbb{R}$ funciones convexas, entonces la siguiente función es convexa:*

$$f(\boldsymbol{x}) = \text{máx}\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x})\}.$$

(iii) *Sea $g : \mathbb{R}^n \to \mathbb{R}$ una función cóncava y sea $S = \{\boldsymbol{x} : g(\boldsymbol{x}) > 0\}$ un conjunto convexo, definamos $f : S \to \mathbb{R}$ como $f(\boldsymbol{x}) = \frac{1}{g(\boldsymbol{x})}$. Entonces f es convexa sobre S.*

(iv) *Sea $g : \mathbb{R} \to \mathbb{R}$ una función no decreciente y convexa y sea $h : \mathbb{R}^n \to \mathbb{R}$ una función convexa. Entonces la función compuesta $f : \mathbb{R}^n \to \mathbb{R}$ definida como $f(\boldsymbol{x}) = g(h(\boldsymbol{x}))$ es una función convexa.*

(v) *Sea $g : \mathbb{R}^m \to \mathbb{R}$ una función convexa y sea $h : \mathbb{R}^n \to \mathbb{R}^m$ una función afín de la forma $h(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$, con $\boldsymbol{A}$ una matrix $m \times n$ y $\boldsymbol{b} \in \mathbb{R}^m$. Entonces la función compuesta $f : \mathbb{R}^n \to \mathbb{R}$ definida como $f(\boldsymbol{x}) = g(h(\boldsymbol{x}))$ es una función convexa.*

Demostración. A continuación demostramos los apartados (II) y (IV), los otros apartados quedan como ejercicio.

- **(II).** Consideremos la función f y un punto $\boldsymbol{z} = \lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}$, con $\lambda \in (0,1)$. Entonces, $f(\boldsymbol{z}) = \text{máx}\{f_1(\boldsymbol{z}), f_2(\boldsymbol{z}), \ldots f_k(\boldsymbol{z})\} = f_i(\boldsymbol{z})$, para algún $i \in \{1, 2, \ldots, k\}$. Como f_i es convexa tenemos $f_i(\boldsymbol{z}) \leq \lambda f_i(\boldsymbol{x}) + (1-\lambda)f_i(\boldsymbol{y})$. Entonces, como $f_i(\boldsymbol{x}) \leq f(\boldsymbol{x})$ y $f_i(\boldsymbol{y}) \leq f(\boldsymbol{y})$,

  $$f(\boldsymbol{z}) = f_i(\boldsymbol{z}) \leq \lambda f_i(\boldsymbol{x}) + (1-\lambda)f_i(\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y}).$$

  Con lo que la función f es convexa.

- **(IV).** Consideremos la función f y un punto $\boldsymbol{z} = \lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}$. Entonces,

$$\begin{aligned} f(\boldsymbol{z}) = g(h(\boldsymbol{z})) &\leq g(\lambda h(\boldsymbol{x}) + (1-\lambda)h(\boldsymbol{y})) \\ &\leq \lambda g(h(\boldsymbol{x})) + (1-\lambda)g(h(\boldsymbol{y})) \\ &= \lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y}). \end{aligned}$$

  Donde la primera desigualdad se sigue de la convexidad de h y el carácter no decreciente de g y la segunda desigualdad se sigue de la convexidad de g. □

•**Ejercicio 1.15.** Demuestra los apartados (I), (III) y (V) de la Proposición 1.10. ◁

El siguiente ejercicio pide demostrar que si en el Apartado (IV) de la Proposición 1.10 eliminamos el supuesto de que la función g sea no decreciente, entonces el resultado deja de ser cierto.

•**Ejercicio 1.16.** Demuestra mediante un ejemplo que la composición de funciones convexas no tiene por qué ser una función convexa. ◁

En la práctica es muy habitual encontrarse funciones de la forma de las que tenemos en la Proposición 1.10: sumas, máximos y ciertas composiciones de funciones convexas. Por tanto, es importante tener presentes las relaciones establecidas por este resultado.

A partir de aquí nos centraremos únicamente en funciones convexas, y los resultados para funciones cóncavas se pueden obtener sin más que tener en cuenta que una función f es cóncava si y sólo si $-f$ es convexa.

Dada una función $f : S \to \mathbb{R}$, se define el *conjunto de nivel inferior* asociado a f y a un valor $r \in \mathbb{R}$ como $S_r = \{\boldsymbol{x} \in S : f(\boldsymbol{x}) \leq r\}$. Análogamente se podría definir el conjunto de nivel superior cambiando "$\leq$" por "$\geq$".

En la Figura 1.10 presentamos una ilustración del conjunto de nivel inferior para una función convexa y una función no convexa. En dicha figura puede verse que S_r es convexo cuando f es convexa y S_r es no convexo cuando f es no convexa. El siguiente resultado muestra que esto no ha sido casualidad: el conjunto de nivel inferior asociado a una función convexa es siempre un conjunto convexo (con la propiedad análoga siendo cierta para funciones cóncavas y el conjunto de nivel superior).

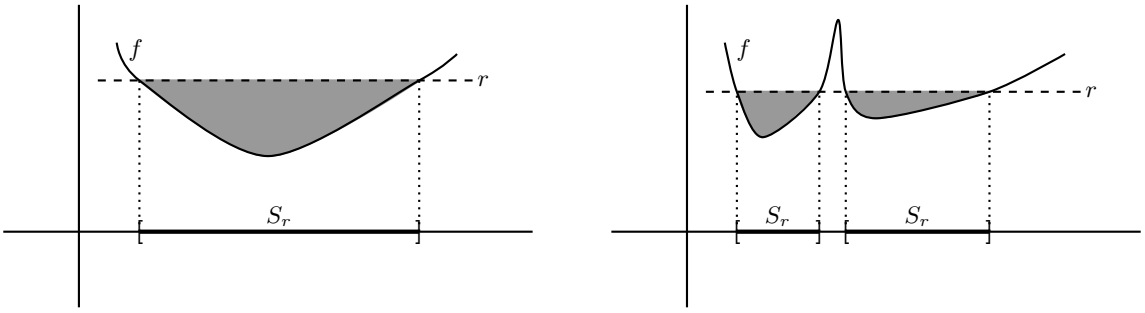
Proposición 1.11. *Dados un conjunto convexo $S \subseteq \mathbb{R}^n$ y una función convexa $f : S \to \mathbb{R}$, entonces, para todo $r \in \mathbb{R}$, el conjunto de nivel inferior $S_r = \{\boldsymbol{x} \in S : f(\boldsymbol{x}) \leq r\}$ es un conjunto convexo.*

Demostración. Tomemos $\boldsymbol{x} \in S_r$ e $\boldsymbol{y} \in S_r$. Entonces, $\boldsymbol{x} \in S$, $\boldsymbol{y} \in S$, $f(\boldsymbol{x}) \leq r$ y $f(\boldsymbol{y}) \leq r$. Tomemos ahora $\lambda \in (0,1)$ y $\boldsymbol{z} = \lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}$. Por la convexidad de S, $\boldsymbol{z} \in S$ y, además, por la convexidad de f:

$$f(\boldsymbol{z}) = f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y}) \leq \lambda r + (1-\lambda)r = r,$$

entonces, como $f(\boldsymbol{z}) \leq r$, tenemos que $\boldsymbol{z} \in S_r$. Por tanto, S_r es convexo. □

Nótese que la proposición anterior no dice que los conjuntos de nivel inferiores de las funciones no convexas vayan a ser siempre no convexos. De hecho, algunos serán convexos y otros no convexos. Basta notar que si tomamos $r = \text{máx}\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\}$, entonces $S_r = S$; con lo que, para este nivel r, S_r será convexo siempre que S lo sea.



(a) Función convexa. S_r convexo.

(b) Función no convexa. S_r no convexo.

Figura 1.10: Ilustrando la relación entre conjuntos de nivel y convexidad de funciones.

1.3.3 Continuidad y derivadas direccionales de funciones convexas

Una propiedad importante de las funciones convexas (y también de las cóncavas) es que son continuas en el interior de su dominio.

Teorema 1.12. *Dado un conjunto convexo $S \subseteq \mathbb{R}^n$ y una función convexa $f : S \to \mathbb{R}$, entonces f es continua en el interior de S.*

La demostración de este resultado es el objeto del siguiente ejercicio.

•••**Ejercicio 1.17.** Demuestra el Teorema 1.12.[4] ◁

A continuación presentamos el concepto de derivada direccional de una función en un punto. Como veremos, pedir que una función tenga derivadas direccionales es mucho menos exigente que pedir que sea diferenciable. Como ya comentamos al inicio de esta sección, muchos algoritmos para minimizar problemas no lineales se basan en encontrar direcciones en las que la función decrece, para lo cual el concepto de derivada direccional es de gran utilidad.

Definición 1.3. Sea $S \subseteq \mathbb{R}^n$ un conjunto no vacío y sea $f : S \to \mathbb{R}$. Sea $\bar{\boldsymbol{x}} \in S$ y sea $\boldsymbol{d} \in \mathbb{R}^n$ un vector no nulo tal que existe $\bar{\lambda} > 0$ con $\bar{\boldsymbol{x}} + \lambda \boldsymbol{d} \in S$ para todo $\lambda \in [0, \bar{\lambda}]$. Entonces, la *derivada direccional* de f en $\bar{\boldsymbol{x}}$ en la dirección $\boldsymbol{d}$, que denotamos $f'(\bar{\boldsymbol{x}}, \boldsymbol{d})$, viene dada por el siguiente límite si éste existe:

$$f'(\bar{\boldsymbol{x}}, \boldsymbol{d}) = \lim_{\lambda \to 0^+} \frac{f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda}.$$

Nótese que, una vez que hemos fijado $\bar{\boldsymbol{x}}$ y $\boldsymbol{d}$, $f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d})$ sólo depende de λ y podemos pensar en f como una función de $\mathbb{R}$ en $\mathbb{R}$. En este sentido, la definición que acabamos de presentar es similar a la definición de derivada de una función en un punto. Sin embargo, ni siquiera en este sentido es equivalente, como se deduce del siguiente ejercicio.

Ejercicio 1.18. Dada la función $f : \mathbb{R} \to \mathbb{R}$ definida como $\mathrm{f}(x) = |x|$, ¿están bien definidas las derivadas direccionales de f en $x = 0$? ¿Es f diferenciable en $x = 0$? ◁

[4] He encontrado varias demostraciones de este resultado en distintos libros y, aunque no son especialmente complicadas, son bastante tediosas. Mi intuición me dice que tiene que haber algún argumento más sencillo, pero yo no lo he encontrado.

El siguiente resultado establece que uno no tiene que tener especial cuidado a la hora de asegurar la existencia de derivadas direccionales de funciones convexas.

Proposición 1.13. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función convexa. Sea $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ y sea $\boldsymbol{d} \in \mathbb{R}^n$ un vector no nulo. Entonces, la derivada direccional $f'(\bar{\boldsymbol{x}}, \boldsymbol{d})$ existe.*

Demostración. Tomemos $\lambda_2 > \lambda_1 > 0$. Por la convexidad de f tenemos que

$$f(\bar{\boldsymbol{x}} + \lambda_1 \boldsymbol{d}) = f\Big(\frac{\lambda_1}{\lambda_2}(\bar{\boldsymbol{x}} + \lambda_2 \boldsymbol{d}) + (1 - \frac{\lambda_1}{\lambda_2})\bar{\boldsymbol{x}}\Big) \leq \frac{\lambda_1}{\lambda_2} f(\bar{\boldsymbol{x}} + \lambda_2 \boldsymbol{d}) + (1 - \frac{\lambda_1}{\lambda_2}) f(\bar{\boldsymbol{x}}),$$

de donde obtenemos que

$$\frac{f(\bar{\boldsymbol{x}} + \lambda_1 \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda_1} \leq \frac{f(\bar{\boldsymbol{x}} + \lambda_2 \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda_2}.$$

Esta última desigualdad contiene una propiedad importante de las funciones convexas: "cada vez crecen más rápido" (en funciones reales diferenciables esto se traduce en que la derivada segunda es siempre positiva). En otras palabras, el cociente $\frac{f(\bar{\boldsymbol{x}}+\lambda \boldsymbol{d})-f(\bar{\boldsymbol{x}})}{\lambda}$ es monótono decreciente a medida que $\lambda \to 0^+$.

Ahora, dado $\lambda \geq 0$ tenemos, por la convexidad de f, que

$$f(\bar{\boldsymbol{x}}) = f\Big(\frac{\lambda}{1+\lambda}(\bar{\boldsymbol{x}} - \boldsymbol{d}) + \frac{1}{1+\lambda}(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d})\Big) \leq \frac{\lambda}{1+\lambda} f(\bar{\boldsymbol{x}} - \boldsymbol{d}) + \frac{1}{1+\lambda} f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}),$$

de donde

$$\frac{f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda} \geq f(\bar{\boldsymbol{x}}) - f(\bar{\boldsymbol{x}} - \boldsymbol{d}).$$

Por tanto, a medida que $\lambda \to 0^+$, tenemos que $\frac{f(\bar{\boldsymbol{x}}+\lambda \boldsymbol{d})-f(\bar{\boldsymbol{x}})}{\lambda}$ es una sucesión decreciente de valores acotados inferiormente por la constante $f(\bar{\boldsymbol{x}}) - f(\bar{\boldsymbol{x}} - \boldsymbol{d})$. Esto implica que el límite existe y viene dado por

$$\lim_{\lambda \to 0^+} \frac{f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda} = \inf_{\lambda > 0} \frac{f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda}. \qquad \square$$

Nótese que el resultado anterior se enunció para funciones $f : \mathbb{R}^n \to \mathbb{R}$, pero si tomamos un conjunto convexo $S \subseteq \mathbb{R}^n$ en vez de $\mathbb{R}^n$, el resultado sólo aplica a los puntos del interior de S (un problema potencial en puntos de la frontera tiene que ver con la posibilidad de que en ellos haya asíntotas verticales). El siguiente ejercicio nos pide construir un ejemplo de función convexa para la cual no exista alguna derivada direccional.

•**Ejercicio 1.19.** Presenta un ejemplo de una función convexa f definida sobre un conjunto convexo S tal que existe algún punto $\bar{\boldsymbol{x}} \in \partial S$ en el que alguna derivada direccional no existe. $\triangleleft$

. . . . . .

Prof. Julio González Díaz

1.3.4 Subgradientes de funciones convexas

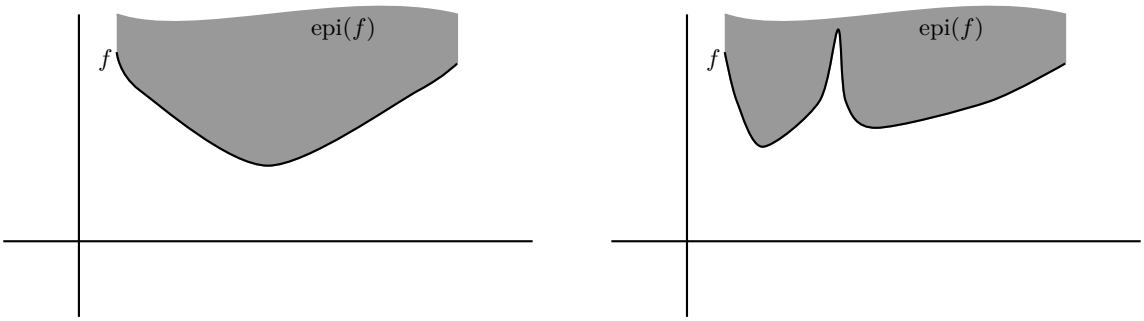
El hecho de que, para toda función convexa f definida sobre un conjunto convexo S, la existencia de derivadas direccionales esté asegurada para cualquier punto del interior de S tiene bastantes implicaciones. Una de ellas es que la diferenciabilidad de f ya no es tan crítica, pues la existencia de estas derivadas direccionales también nos permite trabajar con las propiedades locales de f en cualquier punto $\bar{\boldsymbol{x}} \in S$.

En este apartado veremos un concepto que generaliza el concepto de gradiente para funciones convexas (diferenciables o no) y que tiene gran importancia en el campo de la optimización: el subgradiente. Para ello hemos de definir primero lo que es el *epigrafo* de una función convexa.

Una función $f : S \subseteq \mathbb{R}^n \to \mathbb{R}$ queda completamente caracterizada por su *grafo*, el conjunto $\{(\boldsymbol{x}, f(\boldsymbol{x})) : \boldsymbol{x} \in S\} \subseteq \mathbb{R}^{n+1}$. El epigrafo de f viene dado por los puntos de $\mathbb{R}^{n+1}$ que se encuentran por encima del grafo.

Definición 1.4. Dado un conjunto no vacío $S \subseteq \mathbb{R}^n$ y dada $f : S \to \mathbb{R}$, el *epigrafo* de f, denotado $\text{epi}(f) \subseteq \mathbb{R}^{n+1}$ se define como

$$\text{epi}(f) = \{(\boldsymbol{x}, y) : \boldsymbol{x} \in S,\ y \in \mathbb{R} \text{ e } y \geq f(\boldsymbol{x})\}.$$



(a) Función convexa y su epigrafo.

(b) Función no convexa y su epigrafo.

Figura 1.11: Ilustrando el concepto de epigrafo.

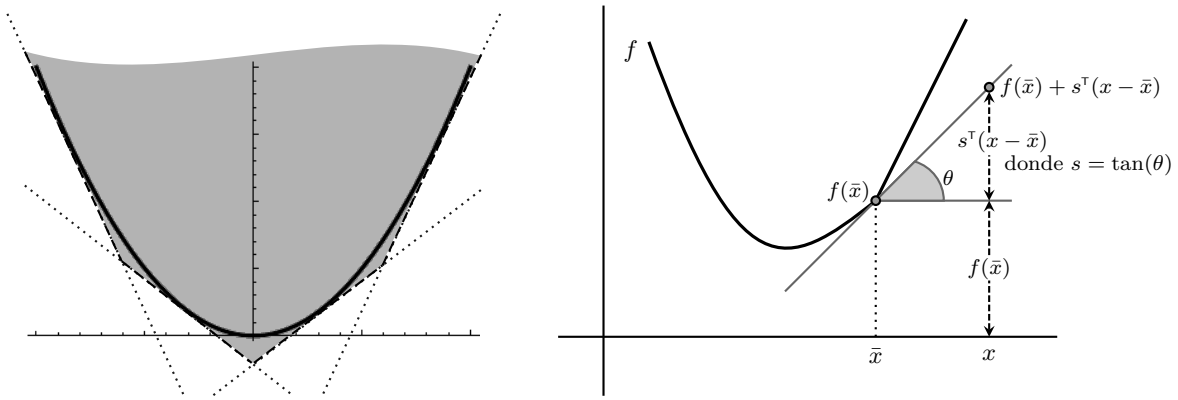
En la Figura 1.11 tenemos representados los epigrafos de dos funciones, una convexa y otra no convexa. Al igual que pasaba cuando estudiamos los conjuntos de nivel, nos encontramos con que el epigrafo de una función convexa es un conjunto convexo, mientras que esto no tiene por qué ser cierto para funciones no convexas. A continuación presentamos este resultado. La demostración es sencilla y queda como ejercicio.

Teorema 1.14. *Sea $S \subseteq \mathbb{R}^n$ un conjunto no vacío y convexo y sea $f : S \to \mathbb{R}$, entonces f es convexa si y solo si su epigrafo es un conjunto convexo.*

•**Ejercicio 1.20.** Demuestra el Teorema 1.14. ◁

El Teorema 1.14 es de gran importancia en el estudio de funciones convexas. Como el epigrafo, $\text{epi}(f)$, de una función convexa es un conjunto convexo, el Teorema del hiperplano

soporte (Teorema 1.4) nos asegura que, en cada punto de la frontera de epi(f) habrá (al menos) un hiperplano soporte. Además, del Corolario 1.3 se sigue que el conjunto epi(f) se puede expresar como la intersección de semiespacios asociados a sus hiperplanos soporte. Desde el punto de vista de la optimización este resultado es de gran importancia, ya que nos permite aproximar el grafo de f (la frontera de epi(f)), como la frontera del conjunto obtenido como la intersección de hiperplanos soporte. Cada una de estas aproximaciones se puede ver como una *linealización exterior* de la función f (ver Figura 1.12(a)).



(a) Linealización exterior de una función.

(b) Subgradiente de una función $f : \mathbb{R} \to \mathbb{R}$.

Figura 1.12: Representación de un subgradiente de una función convexa.

El concepto de hiperplano soporte al epigrafo de una función en un punto da lugar de modo natural al concepto que definimos a continuación: el subgradiente de una función en un punto.

Definición 1.5. Dados un conjunto convexo $S \subseteq \mathbb{R}^n$ y una función convexa $f : S \to \mathbb{R}$, entonces, un vector $\boldsymbol{s} \in \mathbb{R}^n$ es un *subgradiente* de f en $\bar{\boldsymbol{x}} \in S$ si, para todo $\boldsymbol{x} \in S$,

$$f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}) + \boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}).$$

El conjunto de todos los subgradientes de f en $\bar{\boldsymbol{x}}$ se llama *subdiferencial* de f en $\bar{\boldsymbol{x}}$.

En la Figura 1.12(b) se puede ver claramente la relación entre un subgradiente y el concepto de hiperplano soporte. Podemos ver que s se corresponde con la pendiente del hiperplano soporte.

Geométricamente, el subdiferencial de una función convexa f en $\bar{\boldsymbol{x}}$ está formado por todos los hiperplanos en $\mathbb{R}^{n+1}$ que pasan por el punto $(\bar{\boldsymbol{x}}, f(\bar{\boldsymbol{x}}))$ y que están "bajo" el epigrafo de la función. Por otro lado, del mismo modo que la derivada (y el gradiente) de una función diferenciable en un punto dado marca la velocidad de crecimiento/decrecimiento de la función en ese punto, el subdiferencial también contiene información de los rangos entre los que se mueve dicha velocidad según hacia donde nos desplacemos. Esto se puede ver muy bien en la función convexa $f(x) = |x|$. Como es una función definida sobre $\mathbb{R}$, su subdiferencial en 0 está contenido en $\mathbb{R}$ y, más concretamente, viene dado por el intervalo $[-1, 1]$ (Ejercicio 1.22). En este caso, -1 es la derivada por la izquierda de dicha función, es decir, la velocidad a la

que crece la función a la izquierda de 0. De modo análogo, 1 es la velocidad a la que crece la función a la derecha de 0. Cualquier velocidad intermedia es también un subgradiente. Aunque esta interpretación es algo más difícil de ver en dimensiones más altas, la intuición sigue siendo válida.

•**Ejercicio 1.21.** Demuestra que el subdiferencial de una función convexa en un punto es un conjunto convexo. ◁

•**Ejercicio 1.22.** Demuestra que el subdiferencial en el punto $x = 0$ de la función $f : \mathbb{R} \to \mathbb{R}$ definida por $f(x) = |x|$ es el intervalo $[-1, 1]$. ◁

En el siguiente resultado se prueba que toda función convexa tiene al menos un subgradiente en cada punto del interior de su dominio. Este resultado se apoya en la relación entre los conceptos de subgradiente e hiperplano soporte.

Teorema 1.15. *Sea $S \subseteq \mathbb{R}^n$ un conjunto no vacío y convexo y sea $f : S \to \mathbb{R}$ una función convexa, entonces, para todo $\bar{\boldsymbol{x}}$ en el interior de S, existe al menos un subgradiente $\boldsymbol{s}$ de f en $\bar{\boldsymbol{x}}$.*

Además, para todo subgradiente $\boldsymbol{s}$, el hiperplano $H = \{(\boldsymbol{x}, y) \in \mathbb{R}^{n+1} : y = f(\bar{\boldsymbol{x}}) + \boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}})\}$ es un hiperplano soporte de epi(f) *en el punto $(\bar{\boldsymbol{x}}, f(\bar{\boldsymbol{x}}))$.*

Demostración. El Teorema 1.14 nos asegura que epi(f) es un conjunto convexo. Además, como para todo $\bar{\boldsymbol{x}} \in S$ y todo $\varepsilon > 0$, $(\bar{\boldsymbol{x}}, f(\bar{\boldsymbol{x}}) - \varepsilon)$ no pertenece a epi(f), el punto $(\bar{\boldsymbol{x}}, f(\bar{\boldsymbol{x}}))$ pertenece a la frontera de epi(f). Entonces, por el Teorema del hiperplano soporte (Teorema 1.4), existe un vector no nulo $(\boldsymbol{s}_0, \mu) \in \mathbb{R}^{n+1}$ tal que, para todo $(\boldsymbol{x}, y) \in$ epi(f),

$$(\boldsymbol{s}_0, \mu)^\intercal(\boldsymbol{x}, y) \leq (\boldsymbol{s}_0, \mu)^\intercal(\bar{\boldsymbol{x}}, f(\bar{\boldsymbol{x}})), \quad \text{o, equivalentemente,} \quad \boldsymbol{s}_0^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) + \mu(y - f(\bar{\boldsymbol{x}})) \leq 0. \tag{1.2}$$

Además, μ no puede ser positivo, pues entonces podríamos contradecir esta desigualdad tomando $y \in \mathbb{R}$ lo suficientemente grande. A continuación probamos que $\mu < 0$ viendo que no puede ser cero. Supongamos que $\mu = 0$. Entonces, para todo $\boldsymbol{x} \in S$, $\boldsymbol{s}_0^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq 0$. Ahora bien, como $\bar{\boldsymbol{x}}$ pertenece al interior de S, existe $\lambda > 0$ tal que $\boldsymbol{z} = \bar{\boldsymbol{x}} + \lambda \boldsymbol{s}_0 \in S$ y tenemos que $0 \geq \boldsymbol{s}_0^\intercal(\boldsymbol{z} - \bar{\boldsymbol{x}}) = \lambda \boldsymbol{s}_0^\intercal \boldsymbol{s}_0 \geq 0$. Pero esto implicaría que $\boldsymbol{s}_0 = \boldsymbol{0}$ y esto contradice que $(\boldsymbol{s}_0, \mu)$ sea un vector no nulo.

Hemos probado que $\mu < 0$. Sea $\boldsymbol{s} = \frac{\boldsymbol{s}_0}{|\mu|}$. Dividiendo por $|\mu|$ la segunda expresión en la Ecuación (1.2) obtenemos que, para todo $(\boldsymbol{x}, y) \in$ epi(f),

$$\boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) - y + f(\bar{\boldsymbol{x}}) \leq 0.$$

Sea $\boldsymbol{v} = (\boldsymbol{s}, -1) \in \mathbb{R}^{n+1}$ y $r = \boldsymbol{s}^\intercal \bar{\boldsymbol{x}} - f(\bar{\boldsymbol{x}}) \in \mathbb{R}$. Entonces, el hiperplano

$$\begin{aligned} H(\boldsymbol{v}, r) &= \{(\boldsymbol{x}, y) \in \mathbb{R}^{n+1} : \boldsymbol{v}^\intercal(\boldsymbol{x}, y) = r\} \\ &= \{(\boldsymbol{x}, y) \in \mathbb{R}^{n+1} : (\boldsymbol{s}, -1)^\intercal(\boldsymbol{x}, y) = \boldsymbol{s}^\intercal \bar{\boldsymbol{x}} - f(\bar{\boldsymbol{x}})\} \\ &= \{(\boldsymbol{x}, y) \in \mathbb{R}^{n+1} : y = f(\bar{\boldsymbol{x}}) + \boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}})\}, \end{aligned}$$

soporta a epi(f) en $(\bar{\boldsymbol{x}}, f(\bar{\boldsymbol{x}}))$. Por último, si en la expresión $\boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) - y + f(\bar{\boldsymbol{x}}) \leq 0$ tomamos $y = f(\boldsymbol{x})$ tenemos que, para todo $\boldsymbol{x} \in S$, $f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}) + \boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}})$, con lo que $\boldsymbol{s}$ es un subgradiente de f en $\bar{\boldsymbol{x}}$. □

Corolario 1.16. *Sea $S \subseteq \mathbb{R}^n$ un conjunto no vacío y convexo y sea $f : S \to \mathbb{R}$ una función estrictamente convexa, entonces, para todo $\bar{\boldsymbol{x}}$ en el interior de S, existe un vector $\boldsymbol{s}$ tal que, para todo $\boldsymbol{x} \in S$, $\boldsymbol{x} \neq \bar{\boldsymbol{x}}$,*

$$f(\boldsymbol{x}) > f(\bar{\boldsymbol{x}}) + \boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}).$$

••Ejercicio 1.23. Demuestra el Corolario 1.16. ◁

Para terminar presentamos un resultado que viene siendo un recíproco del Teorema 1.15.

Teorema 1.17. *Sea $S \subseteq \mathbb{R}^n$ un conjunto no vacío y convexo y sea $f : S \to \mathbb{R}$. Supongamos que para cada $\bar{\boldsymbol{x}}$ en el interior de S existe un subgradiente de f en $\bar{\boldsymbol{x}}$. Entonces, f es convexa en el interior de S.*

Demostración. Sean $\boldsymbol{x}$ e $\boldsymbol{y}$ dos puntos en el interior de S y sea $\lambda \in (0, 1)$. Como el interior de un conjunto convexo es un conjunto convexo, $\bar{\boldsymbol{x}} = \lambda\boldsymbol{x} + (1 - \lambda)\boldsymbol{y} \in S$. Entonces, existe un subgradiente $\boldsymbol{s}$ de f en $\bar{\boldsymbol{x}}$. Por tanto, como $\boldsymbol{x} \in S$ e $\boldsymbol{y} \in S$,

$$f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}) + \boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) = f(\bar{\boldsymbol{x}}) + \boldsymbol{s}^\intercal(\boldsymbol{x} - (\lambda\boldsymbol{x} + (1 - \lambda)\boldsymbol{y})) = f(\bar{\boldsymbol{x}}) + (1 - \lambda)\boldsymbol{s}^\intercal(\boldsymbol{x} - \boldsymbol{y}),$$
$$f(\boldsymbol{y}) \geq f(\bar{\boldsymbol{x}}) + \boldsymbol{s}^\intercal(\boldsymbol{y} - \bar{\boldsymbol{x}}) = f(\bar{\boldsymbol{x}}) + \boldsymbol{s}^\intercal(\boldsymbol{y} - (\lambda\boldsymbol{x} + (1 - \lambda)\boldsymbol{y})) = f(\bar{\boldsymbol{x}}) + \lambda\boldsymbol{s}^\intercal(\boldsymbol{y} - \boldsymbol{x}).$$

Multiplicando estas desigualdades por λ y $(1 - \lambda)$, respectivamente, y después sumándolas obtenemos

$$\lambda f(\boldsymbol{x}) + (1 - \lambda)f(\boldsymbol{y}) \geq f(\bar{\boldsymbol{x}}) = f(\lambda\boldsymbol{x} + (1 - \lambda)\boldsymbol{y}),$$

de donde se sigue la convexidad de f en el interior de S. □

••Ejercicio 1.24. Demuestra mediante un ejemplo que una función en las hipótesis del Teorema 1.17 no tiene por qué ser convexa en todo el dominio S. ◁

1.3.5 Funciones convexas y diferenciabilidad

En esta sección veremos qué propiedades tienen las funciones convexas cuando además son diferenciables. Antes de nada, es conveniente definir formalmente la noción de diferenciabilidad de funciones de $\mathbb{R}^n$ en $\mathbb{R}$.

Recordemos que una función $f : \mathbb{R} \to \mathbb{R}$ es diferenciable en un punto $\bar{x} \in \mathbb{R}$ si el siguiente límite existe:

$$\lim_{h \to 0} \frac{f(\bar{x} + h) - f(\bar{x})}{h}.$$

En caso afirmativo, la derivada de f en $\bar{x}$, $f'(\bar{x})$, se define precisamente como dicho límite. Sin embargo, esta definición no es la más cómoda para trabajar con funciones en $\mathbb{R}^n$ y es más cómodo apoyarse en que lo anterior implica que, si una función $f : \mathbb{R} \to \mathbb{R}$ es diferenciable en $\bar{x}$, entonces existe una función $\varphi : \mathbb{R} \to \mathbb{R}$ tal que

$$\lim_{h \to 0} \varphi(h) = 0 \quad \text{y} \quad f(\bar{x} + h) = f(\bar{x}) + f'(\bar{x})h + \varphi(h)h.$$

Claramente, bastaría con definir $\varphi(h) = \frac{f(\bar{x}+h)-f(\bar{x})}{h} - f'(\bar{x})$ y la diferenciabilidad de f en $\bar{x}$ implica que $\varphi(h)$ cumple las propiedades requeridas. Equivalentemente, uno puede hablar de la existencia de $\varphi : \mathbb{R} \to \mathbb{R}$ tal que

$$\lim_{x \to \bar{x}} \varphi(x - \bar{x}) = 0 \quad \text{y, para todo } x \in \mathbb{R}, \quad f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + \varphi(x - \bar{x})(x - \bar{x}).$$

Es justamente esta última formulación la que da lugar a una de las definiciones más habituales del concepto de diferenciabilidad de funciones definidas sobre $\mathbb{R}^n$.

Definición 1.6. Sea $S \subseteq \mathbb{R}^n$ un conjunto no vacío y sea $f : S \to \mathbb{R}$. Entonces decimos que f es *diferenciable* en $\bar{\boldsymbol{x}} \in \mathring{S}$ si existen $\nabla f(\bar{\boldsymbol{x}}) \in \mathbb{R}^n$, llamado *gradiente* de f en $\bar{\boldsymbol{x}}$, y una función $\varphi : \mathbb{R}^n \to \mathbb{R}$ tales que

$$\lim_{\boldsymbol{x}\to\bar{\boldsymbol{x}}} \varphi(\boldsymbol{x}-\bar{\boldsymbol{x}}) = 0 \quad \text{y, para todo } \boldsymbol{x} \in S, \quad f(\boldsymbol{x}) = f(\bar{\boldsymbol{x}})+\nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x}-\bar{\boldsymbol{x}})+\varphi(\boldsymbol{x}-\bar{\boldsymbol{x}})\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|. \quad (1.3)$$

Una función f es diferenciable en un conjunto abierto $S' \subseteq S$ si es diferenciable en todos los puntos de S'. Si en la representación de f de la Ecuación (1.3) no incluimos el sumando que contiene a la función φ, dicha representación se conoce como la *aproximación de Taylor de primer orden* de f en $\bar{\boldsymbol{x}}$.

No es difícil comprobar que si una función f es diferenciable en un punto $\bar{\boldsymbol{x}}$, entonces el gradiente ha de ser único y viene dado por

$$\nabla f(\bar{\boldsymbol{x}}) = \Big(\frac{\partial f(\bar{\boldsymbol{x}})}{\partial x_1}, \frac{\partial f(\bar{\boldsymbol{x}})}{\partial x_2}, \ldots, \frac{\partial f(\bar{\boldsymbol{x}})}{\partial x_n}\Big),$$

donde, para cada $i \in \{1, 2, \ldots, n\}$, si $\boldsymbol{e}^i$ denota el i-ésimo vector de la base canónica, entonces

$$\frac{\partial f(\bar{\boldsymbol{x}})}{\partial x_i} = \lim_{\lambda\to 0} \frac{f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{e}^i) - f(\boldsymbol{x})}{\lambda}.$$

••Ejercicio 1.25. Demuestra que el gradiente asociado a una función diferenciable f en un punto $\bar{\boldsymbol{x}} \in \mathring{S}$ es único. ◁

••Ejercicio 1.26. Demuestra que el gradiente asociado a una función diferenciable f en un punto $\bar{\boldsymbol{x}} \in \mathring{S}$ viene dado por

$$\nabla f(\bar{\boldsymbol{x}}) = \Big(\frac{\partial f(\bar{\boldsymbol{x}})}{\partial x_1}, \frac{\partial f(\bar{\boldsymbol{x}})}{\partial x_2}, \ldots, \frac{\partial f(\bar{\boldsymbol{x}})}{\partial x_n}\Big).$$

◁

Es importante destacar también que el cálculo de derivadas direccionales para funciones diferenciables puede hacerse de modo sencillo a través del gradiente. Recordemos que

$$f'(\bar{\boldsymbol{x}}, \boldsymbol{d}) = \lim_{\lambda\to 0^+} \frac{f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda}.$$

Ahora, tomando $\boldsymbol{x} = \bar{\boldsymbol{x}} + \lambda \boldsymbol{d}$ en la definición de diferenciabilidad, tenemos

$$f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) = f(\bar{\boldsymbol{x}}) + \lambda \nabla f(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} + \lambda\varphi(\lambda \boldsymbol{d})\|\boldsymbol{d}\|,$$

dividiendo por $\lambda > 0$ y reordenando los términos obtenemos

$$\frac{f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda} = \nabla f(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} + \varphi(\lambda \boldsymbol{d})\|\boldsymbol{d}\|.$$

Y si ahora hacemos tender λ a 0 llegamos a

$$f'(\bar{\boldsymbol{x}}, \boldsymbol{d}) = \lim_{\lambda \to 0^+} \frac{f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda} = \nabla f(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d}. \tag{1.4}$$

En el siguiente resultado demostramos que, en el caso de funciones convexas diferenciables, el subdiferencial en un punto $\bar{\boldsymbol{x}}$, que recordemos es el conjunto de todos los subgradientes en $\bar{\boldsymbol{x}}$, tiene un único elemento: el gradiente. En particular esto implica que las consideraciones hechas en la subsección anterior para los subgradientes se pueden aplicar inmediatamente para el gradiente.

Proposición 1.18. *Sea $S \subseteq \mathbb{R}^n$ un conjunto no vacío y convexo y sea $f : S \to \mathbb{R}$ una función convexa. Supongamos que f es diferenciable en $\bar{\boldsymbol{x}} \in \mathring{S}$. Entonces, el gradiente de f en $\bar{\boldsymbol{x}}$, $\nabla f(\bar{\boldsymbol{x}})$, es el único subgradiente de f en $\bar{\boldsymbol{x}}$.*

Demostración. El Teorema 1.15 nos asegura que existe al menos un subgradiente $\boldsymbol{s}$ de f en $\bar{\boldsymbol{x}}$. Como f es diferenciable, para todo $\boldsymbol{d} \in \mathbb{R}^n$, tomando $\boldsymbol{x} = \bar{\boldsymbol{x}} + \lambda \boldsymbol{d}$ tenemos

$$\begin{aligned} f(\boldsymbol{x}) &= f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) \geq f(\bar{\boldsymbol{x}}) + \lambda \boldsymbol{s}^\intercal \boldsymbol{d}, \text{ y} \\ f(\boldsymbol{x}) &= f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) = f(\bar{\boldsymbol{x}}) + \lambda \nabla f(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} + \lambda \varphi(\lambda \boldsymbol{d}) \|\boldsymbol{d}\|. \end{aligned}$$

Restando ambas expresiones obtenemos

$$0 \geq \lambda \Big(\boldsymbol{s} - \nabla f(\bar{\boldsymbol{x}}) \Big)^\intercal \boldsymbol{d} - \lambda \varphi(\lambda \boldsymbol{d}) \|\boldsymbol{d}\|.$$

Dividiendo por $\lambda > 0$ y tomando $\lambda \to 0^+$ obtenemos que $\big(\boldsymbol{s} - \nabla f(\bar{\boldsymbol{x}})\big)^\intercal \boldsymbol{d} \leq 0$. Como $\bar{\boldsymbol{x}} \in \mathring{S}$ podemos tomar $\boldsymbol{d} = \boldsymbol{s} - \nabla f(\bar{\boldsymbol{x}})$ y entonces $\boldsymbol{x} = \bar{\boldsymbol{x}} + \lambda \boldsymbol{d}$ pertenecerá a S si λ es lo suficientemente pequeño. Por tanto, $\big(\boldsymbol{s} - \nabla f(\bar{\boldsymbol{x}})\big)^\intercal \big(\boldsymbol{s} - \nabla f(\bar{\boldsymbol{x}})\big) \leq 0$, lo que implica que $\boldsymbol{s} = \nabla f(\bar{\boldsymbol{x}})$. □

La Proposición 1.18 tiene una importante implicación: las nociones de subgradiente y subdiferencial son una extensión, para funciones no diferenciables, del concepto de gradiente. Además, su existencia para funciones convexas está garantizada por el Teorema 1.15.

Ahora podemos apoyarnos en los resultados que demostramos para el subgradiente y obtener de modo inmediato el siguiente resultado.

Teorema 1.19. *Sea $S \subseteq \mathbb{R}^n$ un conjunto no vacío, abierto y convexo y sea $f : S \to \mathbb{R}$ una función diferenciable. Entonces,*

$$\begin{array}{rcll} f \text{ es convexa} & \iff & \text{dados } \bar{\boldsymbol{x}}, \boldsymbol{x} \in S, & f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}). \\ f \text{ es estr. convexa} & \iff & \text{dados } \bar{\boldsymbol{x}}, \boldsymbol{x} \in S, \text{ con } \boldsymbol{x} \neq \bar{\boldsymbol{x}}, & f(\boldsymbol{x}) > f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}). \end{array}$$

•**Ejercicio 1.27.** Demuestra el Teorema 1.19 apoyándote en los resultados ya demostrados relativos a gradiente y subgradientes de funciones convexas. ◁

Aunque ya mencionamos algo similar al estudiar los subgradientes, es conveniente recordar por qué resultados como el Teorema 1.19 pueden ser útiles a la hora de diseñar algoritmos de minimización de un problema de la forma "minimizar $f(\boldsymbol{x})$ sujeto a $\boldsymbol{x} \in S$". Fijado $\bar{\boldsymbol{x}}$, la función $g(\boldsymbol{x}) = f(\bar{\boldsymbol{x}}) + \boldsymbol{s}^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}})$ es una función afín que acota f inferiormente. Por tanto,

si minimizamos g en S o en una relajación del conjunto S obtendremos una cota inferior del mínimo del problema original. Si hacemos esto sucesivamente para distintos puntos de S, y tomamos el máximo de dichas funciones, lo que hacemos es construir una linealización exterior del grafo de f, como ya ilustramos en la Figura 1.12(a). Este tipo de enfoques son comunes en algoritmos denominados de *aproximación sucesiva*.

A continuación presentamos una condición necesaria y suficiente para que una función diferenciable sea convexa. En el caso de funciones de una variable la intuición, que ya mencionamos anteriormente, se reduce a pedir que la pendiente de la función sea creciente.

Teorema 1.20. *Sea $S \subseteq \mathbb{R}^n$ un conjunto no vacío, abierto y convexo sea $f : S \to \mathbb{R}$ una función diferenciable en S. Entonces, f es convexa si y sólo si, para cada par de puntos $\boldsymbol{x} \in S$ e $\boldsymbol{y} \in S$, se tiene*

$$\Big(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\Big)^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}) \geq 0.$$

Análogamente, f es estrictamente convexa si y sólo si, para cada par de puntos distintos $\boldsymbol{x} \in S$ e $\boldsymbol{y} \in S$, se tiene

$$\Big(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\Big)^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}) > 0.$$

Demostración. Presentamos la prueba para el caso convexo, siendo análoga la correspondiente al caso estrictamente convexo.

"$\Rightarrow$" Supongamos que f es convexa y tomemos $\boldsymbol{x} \in S$ e $\boldsymbol{y} \in S$. Entonces, por el Teorema 1.19, tenemos que

$$\begin{aligned} f(\boldsymbol{x}) &\geq f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{y}), \text{ y} \\ f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}). \end{aligned}$$

Sumando ambas igualdades obtenemos que $\big(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\big)^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}) \geq 0$.

"$\Leftarrow$" Tomemos nuevamente $\boldsymbol{x} \in S$ e $\boldsymbol{y} \in S$. El Teorema del valor medio nos asegura que existe $\lambda \in (0, 1)$ tal que $\boldsymbol{z} = \lambda\boldsymbol{x} + (1 - \lambda)\boldsymbol{y}$ cumple que

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \nabla f(\boldsymbol{z})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}). \tag{1.5}$$

Por hipótesis tenemos que $\big(\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x})\big)^{\mathsf{T}}(\boldsymbol{z} - \boldsymbol{x}) \geq 0$ y, como $\boldsymbol{z} - \boldsymbol{x} = (1 - \lambda)(\boldsymbol{y} - \boldsymbol{x})$, entonces, $(1 - \lambda)\big(\nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x})\big)^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}) \geq 0$. Por tanto, $\nabla f(\boldsymbol{z})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}) \geq \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x})$. Ahora, aplicando la Ecuación (1.5), tenemos

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \nabla f(\boldsymbol{z})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}) \geq \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}).$$

Con lo que el Teorema 1.19 nos asegura que f es convexa. □

Aunque los teoremas 1.19 y 1.20 dan sendas caracterizaciones de la convexidad, estas condiciones no son fáciles de verificar en la práctica. A continuación veremos que, cuando tenemos funciones que son dos veces diferenciables, se pueden conseguir condiciones más manejables.

Definición 1.7. Sea $S \subseteq \mathbb{R}^n$ un conjunto no vacío y sea $f : S \to \mathbb{R}$. Entonces decimos que f es *dos veces diferenciable* en $\bar{\boldsymbol{x}} \in \mathring{S}$ si existen $\nabla f(\bar{\boldsymbol{x}}) \in \mathbb{R}^n$, una matriz simétrica $\boldsymbol{H}(\bar{\boldsymbol{x}}) \in \mathbb{R}^{n \times n}$,

llamada *matriz hessiana* y una función $\varphi : \mathbb{R}^n \to \mathbb{R}$ tales que $\lim_{\boldsymbol{x}\to\bar{\boldsymbol{x}}} \varphi(\boldsymbol{x}-\bar{\boldsymbol{x}}) = 0$ y, para todo $\boldsymbol{x} \in S$,

$$f(\boldsymbol{x}) = f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x}-\bar{\boldsymbol{x}}) + \frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}})^\intercal \boldsymbol{H}(\bar{\boldsymbol{x}})(\boldsymbol{x}-\bar{\boldsymbol{x}}) + \varphi(\boldsymbol{x}-\bar{\boldsymbol{x}})\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|^2. \tag{1.6}$$

Una función f es dos veces diferenciable en un conjunto abierto $S' \subseteq S$ si es dos veces diferenciable en todos los puntos de S'. Si en la representación de f de la Ecuación (1.6) no incluimos el sumando que contiene a la función φ, dicha representación se conoce como la *aproximación de Taylor de segundo orden* de f en $\bar{\boldsymbol{x}}$.

Nuevamente, no sería difícil de comprobar que, dada una función dos veces diferenciable, su matriz hessiana viene dada por las derivadas segundas con respecto a las distintas direcciones coordenadas:

$$\boldsymbol{H}(\bar{\boldsymbol{x}}) = \begin{pmatrix} \frac{\partial^2 f(\bar{\boldsymbol{x}})}{\partial x_1 \partial x_1} & \frac{\partial^2 f(\bar{\boldsymbol{x}})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\bar{\boldsymbol{x}})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\bar{\boldsymbol{x}})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\bar{\boldsymbol{x}})}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(\bar{\boldsymbol{x}})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\bar{\boldsymbol{x}})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\bar{\boldsymbol{x}})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\bar{\boldsymbol{x}})}{\partial x_n \partial x_n} \end{pmatrix}.$$

A continuación presentamos una importante caracterización de la convexidad en base a las propiedades de la matriz hessiana. Este resultado puede verse como una continuación natural de la idea detrás del Teorema 1.20: si interpretamos $\big(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\big)^\intercal(\boldsymbol{y}-\boldsymbol{x}) \geq 0$ como que la convexidad es equivalente a que a que "el gradiente crece" a medida que "avanzamos" dentro del dominio, entonces podemos pensar que una función será convexa si su derivada segunda es no negativa. Para funciones de $\mathbb{R}^n$ en $\mathbb{R}$ el papel de la derivada segunda lo asume la matriz hessiana y la idea de que sea no negativa la captura el concepto de matriz semidefinida positiva.

Una matrix simétrica $\boldsymbol{A}_{n\times n}$ es *semidefinida positiva* si, para todo $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x} \geq 0$ y es *definida positiva* si, para todo $\boldsymbol{x} \in \mathbb{R}^n \backslash \{0\}$, $\boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x} > 0$. Análogamente, se pueden definir los conceptos de *semidefinida negativa* y *definida negativa* sin más que cambiar el sentido de las desigualdades. Como toda matriz simétrica es diagonalizable, también se podrían definir los conceptos de matriz definida positiva, semidefinida positiva, definida negativa y semidefinida negativa como aquella que tiene todos sus autovalores positivos, no negativos, negativos y no positivos, respectivamente. Esta definición alternativa es útil a la hora de probar resultados matemáticos, pero en la práctica hay métodos más eficientes para verificar si una matriz pertenece a alguna de las clases que acabamos de definir.

Teorema 1.21. *Sea $S \subseteq \mathbb{R}^n$ un conjunto no vacío, abierto y convexo sea $f : S \to \mathbb{R}$ una función dos veces diferenciable en S. Entonces, f es convexa si y sólo si la matriz hessiana es semidefinida positiva en todo punto de S.*

Demostración. "⇒" Supongamos que f es convexa y tomemos $\bar{\boldsymbol{x}} \in S$. Tenemos que probar que, para todo $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x}^\intercal \boldsymbol{H}(\bar{\boldsymbol{x}})\boldsymbol{x} \geq 0$. Como S es abierto, para cada $\boldsymbol{x} \in \mathbb{R}^n$, tenemos que $\bar{\boldsymbol{x}} + \lambda\boldsymbol{x} \in S$ si $|\lambda| \neq 0$ es suficientemente pequeño. Apoyándonos en el Teorema 1.19 y en el

hecho de que f es dos veces diferenciable tenemos

$$f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}) + \lambda \nabla f(\bar{\boldsymbol{x}})^\intercal \boldsymbol{x}, \text{ y}$$
$$f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{x}) = f(\bar{\boldsymbol{x}}) + \lambda \nabla f(\bar{\boldsymbol{x}})^\intercal \boldsymbol{x} + \lambda^2 \frac{1}{2} \boldsymbol{x}^\intercal \boldsymbol{H}(\bar{\boldsymbol{x}}) \boldsymbol{x} + \lambda^2 \varphi(\lambda \boldsymbol{x}) \|\boldsymbol{x}\|^2.$$

Restando estas dos ecuaciones obtenemos

$$\lambda^2 \frac{1}{2} \boldsymbol{x}^\intercal \boldsymbol{H}(\bar{\boldsymbol{x}}) \boldsymbol{x} + \lambda^2 \varphi(\lambda \boldsymbol{x}) \|\boldsymbol{x}\|^2 \geq 0.$$

Dividiendo por λ^2 y haciendo tender λ a 0 tenemos que $\boldsymbol{x}^\intercal \boldsymbol{H}(\bar{\boldsymbol{x}}) \boldsymbol{x} \geq 0$.

"$\Leftarrow$" Supongamos que la matriz hessiana es semidefinida positiva en todo punto de S. Tomemos $\boldsymbol{x} \in S$ e $\boldsymbol{y} \in X$. El Teorema del valor medio nos asegura que existe $\lambda \in (0, 1)$ tal que $\boldsymbol{z} = \lambda \boldsymbol{x} + (1 - \lambda) \boldsymbol{y}$ cumple que

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \nabla f(\boldsymbol{x})^\intercal (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2} (\boldsymbol{y} - \boldsymbol{x})^\intercal \boldsymbol{H}(\boldsymbol{z}) (\boldsymbol{y} - \boldsymbol{x}).$$

Como $\boldsymbol{z} \in S$, por hipótesis sabemos que $\boldsymbol{H}(\boldsymbol{z})$ es semidefinida positiva. Entonces, $(\boldsymbol{y} - \boldsymbol{x})^\intercal \boldsymbol{H}(\boldsymbol{z})(\boldsymbol{y} - \boldsymbol{x}) \geq 0$, con lo que

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\intercal (\boldsymbol{y} - \boldsymbol{x}).$$

Al ser esto cierto para todo $\boldsymbol{x} \in S$ e $\boldsymbol{y} \in X$, el Teorema 1.19 implica que f es convexa. □

El resultado anterior es de gran utilidad a la hora de verificar la convexidad de una función diferenciable en un punto dado. Además, si la función es cuadrática, entonces esta verificación es independiente del punto elegido pues la matriz hessiana es constante.

De modo similar podría demostrarse que si la matriz es definida positiva en todo punto del dominio entonces es estrictamente convexa. Sin embargo, el recíproco no es del todo cierto en este caso. Basta considerar la función $f(x) = x^4$, que es estrictamente convexa y, sin embargo, la matriz hessiana es $\boldsymbol{H}(x) = 12x^2$, que es definida positiva en todo $\mathbb{R}$ salvo en el 0, donde es semidefinida positiva. Al igual que pasa en el caso real, donde en este caso uno puede apoyarse en las derivadas de orden superior para verificar la convexidad de la función f (en este caso, bastaría notar que la derivada tercera es $24x$ y la derivada cuarta $24 > 0$), también para funciones de $\mathbb{R}^n$ en $\mathbb{R}$ se puede conseguir una caracterización completa de las funciones convexas utilizando la misma idea, pero aplicándola a las derivadas direccionales de orden superior.

1.4 Ejercicios adicionales

•**Ejercicio 1.28.** Dados dos conjuntos $S_1 \subseteq \mathbb{R}^n$ y $S_2 \subseteq \mathbb{R}^n$, demuestra que $\text{conv}(S_1 \cap S_2) \subseteq \text{conv}(S_1) \cap \text{conv}(S_2)$. ¿Es cierta la otra inclusión? En caso negativo, da un contraejemplo. ◁

•**Ejercicio 1.29.** Demuestra que un politopo es un conjunto cerrado y acotado. ◁

••**Ejercicio 1.30.** Demuestra, mediante un ejemplo, que la envoltura convexa de un conjunto cerrado no tiene por qué ser un conjunto cerrado. ◁

••**Ejercicio 1.31.** Demuestra que, dado un cono convexo no vacío $C \subseteq \mathbb{R}^n$ y su cono polar C^*, $C + C^* = \mathbb{R}^n$, es decir, dado $\boldsymbol{x} \in \mathbb{R}^n$ existen $\boldsymbol{x}^1 \in C$ y $\boldsymbol{x}^2 \in C^*$ tales que $\boldsymbol{x} = \boldsymbol{x}^1 + \boldsymbol{x}^2$. ◁

•**Ejercicio 1.32.** Dado el conjunto convexo $S = \{\boldsymbol{x} \in \mathbb{R}^3 : x_1^2 + x_2^2 + x_3^2 \leq 4,\ x_1^2 - 4x_2 \leq 0\}$ y el punto $\boldsymbol{y} = (1, 0, 2)$, encuentra la mínima distancia de $\boldsymbol{y}$ a S, un punto de S que se encuentre a dicha distancia y un hiperplano separando $\{\boldsymbol{y}\}$ de S. ◁

•**Ejercicio 1.33.** Demuestra que el subdiferencial de una función en un punto es un conjunto cerrado. ◁

•**Ejercicio 1.34.** Estudia la concavidad y convexidad de las siguientes funciones (apoyándote en la matriz hessiana):

- $f(x_1, x_2) = 2x_1^2 - 4x_1x_2 - 8x_1 + 3x_2$.
- $f(x_1, x_2) = x_1 e^{-(x_1+3x_2)}$.
- $f(x_1, x_2) = -x_1 - 3x_2^2 + 4x_1x_2 + 10x_1 - 10x_2$.
- $f(x_1, x_2, x_3) = 2x_1x_2 + 2x_1^2 + x_2^2 + 2x_3^2 - 5x_1x_3$.
- $f(x_1, x_2, x_3) = -2x_1^2 - 3x_2^2 - 2x_3^2 + 8x_1x_2 + 3x_1x_3 + 4x_2x_3$. ◁

•**Ejercicio 1.35.** Dada la función $f(x_1, x_2) = e^{2x_1^2 - x_2^2} - 3x_1 + 5x_2$, ¿cuál es la aproximación de Taylor de primer orden de f en $(1, 1)$? ¿y la aproximación de segundo orden? Argumenta si estas aproximaciones son cóncavas o convexas. ◁

Tema 2
Optimización Convexa

Contenidos

2.1 Mínimos y máximos en programación convexa

En el tema anterior vimos que las funciones convexas tienen una gran cantidad de propiedades. Además, ya comentamos que muchas de ellas son especialmente útiles a la hora de resolver problemas de optimización. En este tema empezaremos a trabajar de modo más concreto en las conexiones entre convexidad y optimización, y que hacen de la convexidad de funciones un objeto central al estudio de problemas de optimización. Por otro lado, se trata de una propiedad bastante menor en el resto de los campos de las matemáticas, donde la distinción fundamental suele ser entre lineal y no lineal. A la hora de resolver problemas de optimización la distinción entre problemas convexos y no convexos tiene al menos la misma importancia que la distinción entre problemas lineales y no lineales.

En este tema vamos a estudiar el problema de minimizar funciones convexas definidas sobre conjuntos convexos. En el caso de funciones cóncavas, los resultados que presentamos en esta sección siguen siendo ciertos pero para problemas de maximizar en vez de minimizar. Esto es inmediato, pues minimizar $f(\boldsymbol{x})$ es lo mismo que maximizar $-f(\boldsymbol{x})$ y $f(\boldsymbol{x})$ es convexa si y sólo si $-f(\boldsymbol{x})$ es cóncava.

En su formulación más general, un *problema de minimización* se puede expresar como

$$\begin{aligned} \text{minimizar} \quad & f(\boldsymbol{x}) \\ \text{sujeto a} \quad & \boldsymbol{x} \in S. \end{aligned} \tag{2.1}$$

En este caso, cualquier punto perteneciente al conjunto S se llama solución factible. Los problemas en los que tanto la función objetivo f como el conjunto factible S son convexos se denominan *problemas de programación convexa.*

Dado un problema de minimización como el expresado en la Ecuación (2.1), decimos que un punto $\bar{\boldsymbol{x}} \in S$ es un *óptimo global* o una *solución óptima* si, para todo $\boldsymbol{x} \in S$, $f(\bar{\boldsymbol{x}}) \leq f(\boldsymbol{x})$. Decimos que un punto $\bar{\boldsymbol{x}} \in S$ es un *óptimo local* si existe $\varepsilon > 0$ tal que, para todo $\boldsymbol{x} \in B(\bar{\boldsymbol{x}}, \varepsilon) \cap S$, $f(\bar{\boldsymbol{x}}) \leq f(\boldsymbol{x})$. Cuando las desigualdades son estrictas, hablamos de óptimos globales estrictos y óptimos locales estrictos. En la Figura 2.1 ilustramos los conceptos de optimalidad local y optimalidad global. Los puntos C, D y E de la figura son óptimos locales estrictos. El punto A es un óptimo local no estricto, al igual que cualquier punto del interior del segmento que une A y B (no son estrictos pues en todo el segmento la función objetivo toma el mismo valor). El punto F es el único óptimo global que, por tanto, es estricto.

En la práctica, la mayor parte de los problemas de programación matemática se nos presentan con una descripción explícita del conjunto S, típicamente de la forma:

$$\begin{aligned} \text{minimizar} \quad & f(\boldsymbol{x}) & & \\ \text{sujeto a} \quad & g_i(\boldsymbol{x}) \leq 0 & & i = 1, \ldots, m \\ & h_j(\boldsymbol{x}) = 0 & & j = 1, \ldots, l. \end{aligned} \tag{2.2}$$

A continuación presentamos una condición suficiente para que la región factible asociada a este tipo de formulación sea un conjunto convexo.

Proposición 2.1. *Dado un problema de programación matemática como el formulado en la Ecuación (2.2), si la función f es convexa y las funciones g_i son convexas y las h_j son afines, entonces se trata de un problema de programación convexa.*

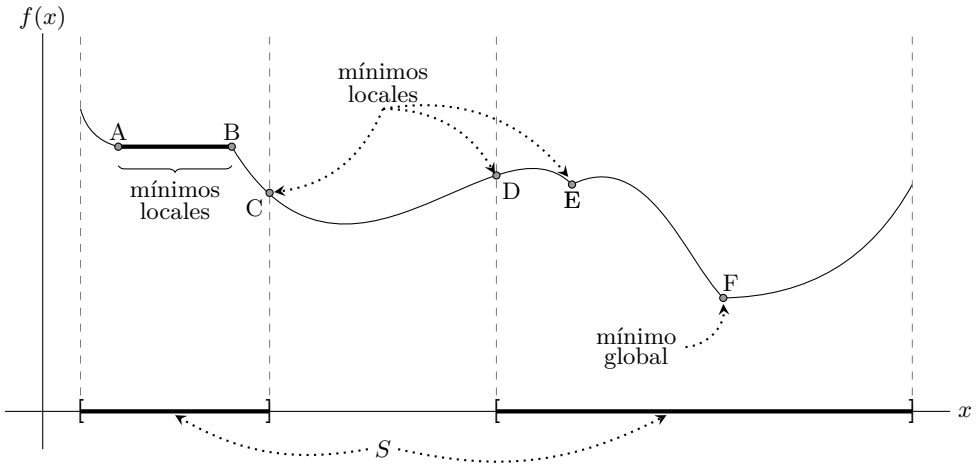


Figura 2.1: Ilustración de los conceptos de óptimo local y óptimo global.

Demostración. Como f es convexa, basta probar que el conjunto factible S asociado a dicho problema de programación matemático es un conjunto convexo. Tomemos $\boldsymbol{x}, \boldsymbol{y} \in S$ y $\lambda \in [0,1]$. Tenemos que probar que $\boldsymbol{z} = \lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}$ pertenece a S. Dada una función h_j tenemos

$$h_j(\boldsymbol{z}) = h_j(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \overset{\text{linealidad}}{=} \lambda h_j(\boldsymbol{x}) + (1-\lambda)h_j(\boldsymbol{y}) \overset{\boldsymbol{x},\boldsymbol{y}\in S}{=} \lambda \cdot 0 + (1-\lambda)\cdot 0 = 0,$$

por tanto, $\boldsymbol{z}$ satisface las restricciones dadas por las funciones h_j. Dada una función g_i tenemos

$$g_i(\boldsymbol{z}) = g_i(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \overset{\text{convexidad}}{\leq} \lambda g_i(\boldsymbol{x}) + (1-\lambda)g_i(\boldsymbol{y}) \overset{\boldsymbol{x},\boldsymbol{y}\in S}{\leq} \lambda \cdot 0 + (1-\lambda)\cdot 0 = 0,$$

por tanto, $\boldsymbol{z}$ también satisface las restricciones dadas por las funciones g_i. Entonces, hemos probado que $\boldsymbol{z} \in S$, con lo que el conjunto S es un conjunto convexo. □

Como toda función afín es convexa, el resultado anterior implica que la programación lineal es un caso particular de la programación convexa.[1] El siguiente resultado nos garantiza que, en cualquier problema de programación convexa, todo óptimo local es un óptimo global.

Teorema 2.2. *Sea $S \subseteq \mathbb{R}^n$ un conjunto no vacío y convexo y sea $f : S \to \mathbb{R}$ una función convexa en S. Si $\bar{\boldsymbol{x}}$ es un óptimo local del problema* "minimizar $f(\boldsymbol{x})$ *sujeto a* $\boldsymbol{x} \in S$", *entonces*

(i) *El punto $\bar{\boldsymbol{x}}$ es un óptimo global.*

(ii) *Si $\bar{\boldsymbol{x}}$ es un óptimo local estricto o f es estrictamente convexa, entonces $\bar{\boldsymbol{x}}$ es el único óptimo global.*

Demostración. Supongamos que $\bar{\boldsymbol{x}}$ es un óptimo local que no es un óptimo global. Entonces,

[1] Nótese que los lados derechos de las restricciones de un problema de programación lineal se pueden ver como funciones lineales si las vemos de la forma $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ o afines si las pensamos como $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \leq \boldsymbol{0}$. Esto hace que sea habitual referirse a las restricciones de estos problemas como lineales sea cual sea la representación, aunque en una de ellas esto no sea del todo preciso.

- Existe $\varepsilon > 0$ tal que, para todo $\boldsymbol{x} \in B(\boldsymbol{x}, \varepsilon) \cap S$, $f(\bar{\boldsymbol{x}}) \leq f(\boldsymbol{x})$.
- Existe $\boldsymbol{y} \in S$ tal que $f(\boldsymbol{y}) < f(\bar{\boldsymbol{x}})$.

Como S es convexo, para todo $\lambda \in [0,1]$, $\boldsymbol{z}_\lambda = \lambda\bar{\boldsymbol{x}} + (1-\lambda)\boldsymbol{y} \in S$. Además, para todo $\lambda \in [0,1)$,

$$f(\boldsymbol{z}_\lambda) = f(\lambda\bar{\boldsymbol{x}} + (1-\lambda)\boldsymbol{y}) \overset{\text{convexidad}}{\leq} \lambda f(\bar{\boldsymbol{x}}) + (1-\lambda) f(\boldsymbol{y}) \overset{f(\boldsymbol{y})<f(\bar{\boldsymbol{x}})}{<} \lambda f(\bar{\boldsymbol{x}}) + (1-\lambda) f(\bar{\boldsymbol{x}}) = f(\bar{\boldsymbol{x}}).$$

Entonces, si tomamos λ suficientemente próximo a 1 tendremos que $\boldsymbol{z}_\lambda \in B(\bar{\boldsymbol{x}}, \varepsilon) \cap S$ y que $f(\boldsymbol{z}_\lambda) < f(\bar{\boldsymbol{x}})$, lo que contradice que $\bar{\boldsymbol{x}}$ sea un mínimo local. □

•**Ejercicio 2.1.** Demuestra el segundo apartado del Teorema 2.2. ◁

El Teorema 2.2, a pesar de su relativa sencillez, es uno de los resultados más relevantes a la hora de valorar la dificultad de un problema de programación matemática. Para problemas de programación matemática en general, existen muchos algoritmos eficientes para buscar óptimos locales; en particular, una técnica habitual es moverse por la región factible buscando puntos donde se anule el gradiente. El hecho de que optimalidad local implique optimalidad global hace que estos algoritmos sean suficientes para resolver completamente problemas de programación convexa.

En el resultado siguiente presentamos una condición necesaria y suficiente para que un punto factible sea un mínimo global en un problema de programación convexa. En caso de que este mínimo global no exista pueden suceder dos cosas:

- $\inf\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\}$ es un número real pero no se alcanza dentro de S. Por ejemplo, $f(x) = x$ y $S = (0,1) \subseteq \mathbb{R}$.
- $\inf\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\} = -\infty$. Por ejemplo, $f(x) = x$ y $S = \mathbb{R}$.

Teorema 2.3. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función convexa y sea $S \subseteq \mathbb{R}^n$ un conjunto no vacío y convexo. Consideremos el siguiente problema de optimización:*

$$\begin{aligned} \text{minimizar} \quad & f(\boldsymbol{x}) \\ \text{sujeto a} \quad & \boldsymbol{x} \in S. \end{aligned}$$

El punto $\bar{\boldsymbol{x}} \in S$ es un mínimo global de este problema si y sólo si f tiene un subgradiente $\boldsymbol{s}$ en $\bar{\boldsymbol{x}}$ tal que

$$\boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \geq 0, \quad \textit{para todo } \boldsymbol{x} \in S.$$

Demostración. "⇐" Supongamos que, para todo $\boldsymbol{x} \in S$, $\boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \geq 0$, donde $\boldsymbol{s}$ es un subgradiente f en $\bar{\boldsymbol{x}}$. Entonces, por la definición de subgradiente, para todo $\boldsymbol{x} \in S$

$$f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}) + \boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \geq f(\bar{\boldsymbol{x}}).$$

Y $\bar{\boldsymbol{x}}$ es un mínimo global de f en S.

"⇒" Supongamos ahora que $\bar{\boldsymbol{x}}$ es un mínimo global y definamos los dos siguientes subconjuntos en $\mathbb{R}^{n+1}$:

$$\begin{aligned} F_1 &= \{(\boldsymbol{x} - \bar{\boldsymbol{x}}, y) : \boldsymbol{x} \in \mathbb{R}^n, y > f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}})\}, \text{ y} \\ F_2 &= \{(\boldsymbol{x} - \bar{\boldsymbol{x}}, y) : \boldsymbol{x} \in S, y \leq 0\}. \end{aligned}$$

Es fácil comprobar que F_1 y F_2 son conjuntos convexos. Además, $F_1 \cap F_2 = \emptyset$ pues, de otro modo, existiría un punto $(\boldsymbol{x}, y) \in \mathbb{R}^{n+1}$ tal que $\boldsymbol{x} \in S$ y $0 \geq y > f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}})$, lo que es imposible pues $\bar{\boldsymbol{x}}$ es un mínimo global de f en S. Por el Teorema 1.6, existe un hiperplano separando F_1 y F_2. Más concretamente, existen un vector no nulo $(\boldsymbol{s}_0, \mu) \in \mathbb{R}^{n+1}$ y un escalar $r \in \mathbb{R}$ tales que

$$\begin{aligned}
&\boldsymbol{s}_0^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) + \mu y \leq r, \text{para todo } (\boldsymbol{x} - \bar{\boldsymbol{x}}, y) \in F_1, \text{es decir, para todo } \boldsymbol{x} \in \mathbb{R}^n \text{ e } y > f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}).\\
&\boldsymbol{s}_0^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) + \mu y \geq r, \text{para todo } (\boldsymbol{x} - \bar{\boldsymbol{x}}, y) \in F_2, \text{es decir, para todo } \boldsymbol{x} \in S \text{ e } y \leq 0.
\end{aligned}$$

Tomando $\boldsymbol{x} = \bar{\boldsymbol{x}}$ e $y = 0$ en la segunda desigualdad obtenemos $r \leq 0$. Ahora, tomando $\boldsymbol{x} = \bar{\boldsymbol{x}}$ e $y = \varepsilon \geq 0$ en la primera desigualdad obtenemos que, para todo $\varepsilon > 0$, $\mu\varepsilon \leq r$. Entonces, necesariamente $\mu \leq 0$. Además, r no puede ser negativo, pues tomando ε suficientemente próximo a cero llegaríamos a una contradicción con $\mu\varepsilon \leq r$. Por tanto, $r = 0$.

Si $\mu = 0$, entonces la primera desigualdad implica que, para todo $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{s}_0^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq 0$. Pero ahora bastaría tomar $\boldsymbol{x} = \bar{\boldsymbol{x}} + \boldsymbol{s}_0$ para obtener que $\boldsymbol{s}_0^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) = \boldsymbol{s}_0^\intercal \boldsymbol{s}_0 \leq 0$, lo que implica que $\boldsymbol{s}_0 = \boldsymbol{0}$, contradiciendo que $(\boldsymbol{s}_0, \mu)$ es no nulo.

Entonces, $\mu < 0$. Si ahora dividimos ambas desigualdades por $-\mu$, tomamos $\boldsymbol{s} = \frac{\boldsymbol{s}_0}{-\mu}$ y tenemos en cuenta que $r = 0$, obtenemos:

$$\begin{aligned}
&\boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) - y \leq 0, \quad \text{para todo } \boldsymbol{x} \in \mathbb{R}^n \text{ e } y > f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}).\\
&\boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) - y \geq 0, \quad \text{para todo } \boldsymbol{x} \in S \text{ e } y \leq 0.
\end{aligned}$$

Tomando $y = 0$ en la segunda desigualdad tenemos que, para todo $\boldsymbol{x} \in S$, $\boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \geq 0$. Además, si en la primera desigualdad, fijado $\boldsymbol{x}$, tomamos una sucesión de valores de y que converja a $f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}})$ se sigue que, para todo $\boldsymbol{x} \in \mathbb{R}^n$,

$$f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}) + \boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}).$$

Por tanto, $\boldsymbol{s}$ es un subgradiente de f en $\bar{\boldsymbol{x}}$ tal que, para todo $\boldsymbol{x} \in S$, $\boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \geq 0$. □

En la Sección 2.3 presentamos una discusión detallada de las intuiciones e implicaciones de este resultado, presentando además una interpretación geométrica del mismo (ver Figura 2.4).

•**Ejercicio 2.2.** Demuestra la convexidad de los conjuntos F_1 y F_2 construidos en la demostración del Teorema 2.3. ◁

Corolario 2.4. *Bajo las hipótesis del Teorema 2.3, si S es abierto, entonces $\bar{\boldsymbol{x}}$ es un mínimo global del problema si y sólo si el vector $\boldsymbol{0}$ es un subgradiente de f en $\bar{\boldsymbol{x}}$.*

Demostración. Aplicando el Teorema 2.3 tenemos que $\bar{\boldsymbol{x}}$ es un mínimo global si y sólo si existe un subgradiente $\boldsymbol{s}$ de f en $\bar{\boldsymbol{x}}$ tal que, para todo $\boldsymbol{x} \in S$, $\boldsymbol{s}^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \geq 0$. Como S es abierto, si tomamos $\lambda > 0$ suficientemente pequeño tenemos que $\boldsymbol{x} = \bar{\boldsymbol{x}} - \lambda\boldsymbol{s}$ pertenece a S. Entonces, $-\lambda\boldsymbol{s}^\intercal\boldsymbol{s} \geq 0$ y, por tanto, $\boldsymbol{s} = \boldsymbol{0}$. □

Por último, podemos ver a qué se reducen estos resultados en el caso diferenciable (ver Figura 2.4(b)). En particular, vemos que recuperamos la condición clásica de optimalidad, que nos dice que el gradiente se ha de anular en los puntos óptimos (del interior del dominio).

Corolario 2.5. *Bajo las hipótesis del Teorema 2.3, si f es diferenciable tenemos que, $\bar{\boldsymbol{x}} \in S$ es un mínimo global del problema si y sólo si, para todo $\boldsymbol{x} \in S$, $\nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \geq 0$. Además, si S es abierto, $\bar{\boldsymbol{x}}$ es un mínimo global del problema si y sólo si $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$.*

Demostración. Inmediato a partir de la Proposición 1.18, que nos asegura que el gradiente es el único subgradiente de una función diferenciable. □

A continuación presentamos un resultado relativo al conjunto de óptimos globales de un problema de programación convexa.

Proposición 2.6. *El conjunto de óptimos globales de un problema de programación convexa es un conjunto convexo.*

•**Ejercicio 2.3.** Demuestra la Proposición 2.6. ◁

Máximos en problemas de programación convexa

Recordemos que las funciones convexas son funciones que, en cada dirección, o bien "aceleran" su crecimiento o "desaceleran" su decrecimiento. Dado esto, parece natural que para encontrar máximos de estas funciones tengamos que buscar en la frontera de dominio de definición. Esto es justamente lo que nos dice el siguiente resultado que, por tanto, aplica también a los mínimos en problemas con funciones cóncavas.

Teorema 2.7. *Sea $S \subseteq \mathbb{R}^n$ un poliedro acotado (politopo) no vacío y sea $f : S \to \mathbb{R}$ una función continua y convexa en S. Entonces, al menos un óptimo del problema "*maximizar $f(\boldsymbol{x})$ *sujeto a $\boldsymbol{x} \in S$" es un punto extremo de S.*

Demostración. Dado que f es continua y S compacto, la función f tiene al menos un máximo global $\bar{\boldsymbol{x}} \in S$. Si $\bar{\boldsymbol{x}}$ es un punto extremo de S ya tendríamos el resultado. En caso contrario, como S es un poliedro acotado, el Teorema de Carathéodory nos asegura que existen $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m$ puntos extremos de S, con $m > 1$, tales que $\bar{\boldsymbol{x}} = \sum_{k=1}^m \lambda_k \boldsymbol{x}^k$ con $\sum_{k=1}^m \lambda_k = 1$ y $\lambda_k > 0$ para todo $k \in \{1, \ldots, m\}$.[2] Por la convexidad de f tenemos que

$$f(\bar{\boldsymbol{x}}) = f(\sum_{k=1}^m \lambda_k \boldsymbol{x}^k) \leq \sum_{k=1}^m \lambda_k f(\boldsymbol{x}^k).$$

Dado que, para todo $k \in \{1, \ldots, m\}$, $f(\bar{\boldsymbol{x}}) \geq f(\boldsymbol{x}^k)$, la anterior desigualdad implica que, para todo $k \in \{1, \ldots, m\}$, $f(\bar{\boldsymbol{x}}) = f(\boldsymbol{x}^k)$. Por tanto, los puntos extremos $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m$ son también máximos globales de f en S. □

El Teorema 2.7 también es cierto si S es un conjunto convexo compacto general, no necesariamente poliédrico. Asimismo, no es necesario asumir la continuidad de S pero, dado que el Teorema 1.12 sólo garantiza la continuidad de f en el interior de S, habría que tener algo de cuidado con los puntos de la frontera del conjunto.

Nótese que, aplicado a problemas de programación lineal, cuya función objetivo es tanto cóncava como convexa, el resultado anterior nos dice algo que ya conocíamos: al menos un punto extremo es óptimo (pues estamos minimizando una función cóncava, que es lo mismo que maximizar una función convexa).

[2] Los puntos extremos de S para los que el coeficiente sea 0 ya no los ponemos en la representación de $\bar{\boldsymbol{x}}$.

2.2 Direcciones de descenso y direcciones factibles

En esta sección presentamos formalmente otro enfoque para encontrar condiciones de optimalidad. Aunque los resultados obtenidos son esencialmente los mismos, este enfoque resultará muy conveniente cuando queramos generalizar los resultados más allá de problemas de programación convexa. Supongamos entonces que tenemos un problema de optimización de la forma $\text{minimizar}_{\boldsymbol{x}\in S} f(\boldsymbol{x})$, con $f : \mathbb{R}^n \to \mathbb{R}$. A continuación presentamos la definición de los conceptos de dirección de descenso y dirección factible.

Definición 2.1. El conjunto de *direcciones de descenso* de la función f en el punto $\boldsymbol{x}$, $D^D(f, \boldsymbol{x})$, viene dado por

$$D^D(f, \boldsymbol{x}) = \{\boldsymbol{d} \in \mathbb{R}^n : f(\boldsymbol{x} + \lambda \boldsymbol{d}) < f(\boldsymbol{x}) \text{ para todo } \lambda \in (0, \delta), \text{para un cierto } \delta > 0\}.$$

Definición 2.2. El conjunto de *direcciones factibles* del conjunto S en el punto $\boldsymbol{x}$, $D^F(S, \boldsymbol{x})$, viene dado por

$$D^F(S, \boldsymbol{x}) = \{\boldsymbol{d} \in \mathbb{R}^n : \boldsymbol{x} + \lambda \boldsymbol{d} \in S \text{ para todo } \lambda \in (0, \delta), \text{para un cierto } \delta > 0\}.$$

•**Ejercicio 2.4.** Demuestra que los conjuntos $D^D(f, \boldsymbol{x}) \cup \{\boldsymbol{0}\}$ y $D^F(S, \boldsymbol{x})$ son conos convexos cuando f es convexa y S es convexo, respectivamente. ◁

Para facilitar la notación, cuando no haya confusión denotaremos estos dos conjuntos simplemente como D^D y D^F y nos referiremos a sus elementos como direcciones de descenso y direcciones factibles, respectivamente. El siguiente resultado relaciona el concepto de dirección de descenso con el de derivada direccional. Recordemos que, para funciones diferenciables, la Ecuación (1.4) nos dice que $f'(\bar{\boldsymbol{x}}, \boldsymbol{d}) = \nabla f(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d}$.

Proposición 2.8. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función diferenciable. Dados un punto $\boldsymbol{x} \in \mathbb{R}^n$ y una dirección $\boldsymbol{d} \in \mathbb{R}^n$. Si $\nabla f(\boldsymbol{x})^\intercal \boldsymbol{d} < 0$, entonces $\boldsymbol{d}$ es una dirección de descenso en $\boldsymbol{x}$. Si además f es convexa entonces el recíproco también es cierto.*

Demostración. "⇒" Supongamos que $\nabla f(\boldsymbol{x})^\intercal \boldsymbol{d} < 0$. Por la definición de gradiente tenemos

$$f(\boldsymbol{x} + \lambda \boldsymbol{d}) = f(\boldsymbol{x}) + \lambda \nabla f(\boldsymbol{x})^\intercal \boldsymbol{d} + \lambda \varphi(\lambda \boldsymbol{d}) \|\boldsymbol{d}\|,$$

con $\lim_{\lambda \to 0} \varphi(\lambda \boldsymbol{d}) = 0$. Reordenando los términos y dividiendo por $\lambda \neq 0$ obtenemos

$$\frac{f(\boldsymbol{x} + \lambda \boldsymbol{d}) - f(\boldsymbol{x})}{\lambda} = \nabla f(\boldsymbol{x})^\intercal \boldsymbol{d} + \varphi(\lambda \boldsymbol{d}) \|\boldsymbol{d}\|.$$

Si ahora hacemos tender λ a 0 y usamos que $\nabla f(\boldsymbol{x})^\intercal \boldsymbol{d} < 0$ y $\lim_{\lambda \to 0} \varphi(\lambda \boldsymbol{d}) = 0$ tenemos que existe $\delta > 0$ tal que $\nabla f(\boldsymbol{x})^\intercal \boldsymbol{d} + \varphi(\lambda \boldsymbol{d}) \|\boldsymbol{d}\| < 0$ para todo $\lambda \in (0, \delta)$, de donde se sigue el resultado.

"⇐" Supongamos ahora que $\boldsymbol{d}$ es una dirección de descenso en $\boldsymbol{x}$ y que además la función f es convexa. Por el Teorema 1.19, la convexidad de f implica que, dado $\lambda > 0$, $f(\boldsymbol{x} + \lambda \boldsymbol{d}) - f(\boldsymbol{x}) \geq \lambda \nabla f(\boldsymbol{x})^\intercal \boldsymbol{d}$, de donde se deduce inmediatamente que $\nabla f(\boldsymbol{x})^\intercal \boldsymbol{d} < 0$ por ser $\boldsymbol{d}$ una dirección de descenso en $\boldsymbol{x}$. □

Nótese que el recíproco en el resultado anterior no es cierto si la función no es convexa pues, para la función $f(x) = x^3$, la dirección $d = -1$ es una dirección de descenso en $x = 0$ y, sin embargo, $f'(0, -1) = \nabla f(0)(-1) = 0$. De modo más general, cuando en una función diferenciable nos situamos en un máximo o un punto de silla, tendremos que hay direcciones de descenso y, sin embargo, no podemos apoyarnos en las derivadas direccionales o el gradiente para identificarlas.

El siguiente resultado caracteriza los óptimos globales de un problema de programación convexa como aquellos puntos en los que no hay ninguna dirección que sea simultáneamente factible y de descenso.

Teorema 2.9. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función convexa y sea $S \subseteq \mathbb{R}$ un conjunto no vacío y convexo. Consideremos el siguiente problema de optimización:*

$$\begin{array}{rl} \text{minimizar} & f(\boldsymbol{x}) \\ \text{sujeto a} & \boldsymbol{x} \in S. \end{array}$$

El punto $\bar{\boldsymbol{x}} \in S$ es un mínimo local (y global) de este problema si y sólo si $D^D \cap D^F = \emptyset$.

Demostración. "$\Rightarrow$" La existencia de una dirección $\boldsymbol{d} \in D^D \cap D^F$ implica inmediatamente que $\bar{\boldsymbol{x}}$ no es una solución óptima.

"$\Leftarrow$" Supongamos que un punto $\bar{\boldsymbol{x}}$ no es un óptimo global. Entonces, existe $\boldsymbol{x} \in S$ con $f(\boldsymbol{x}) < f(\bar{\boldsymbol{x}})$. Definamos $\boldsymbol{d} = (\boldsymbol{x} - \bar{\boldsymbol{x}})$. Tenemos que, para todo $\lambda \in (0, 1)$,

$$\bar{\boldsymbol{x}} + \lambda \boldsymbol{d} = \lambda(\bar{\boldsymbol{x}} + \boldsymbol{d}) + (1 - \lambda)\bar{\boldsymbol{x}} = \lambda \boldsymbol{x} + (1 - \lambda)\bar{\boldsymbol{x}},$$

que pertenece a S pues S es convexo. Por tanto $\boldsymbol{d} \in D^F$. Además, por la convexidad de f,para todo $\lambda \in (0, 1)$,

$$f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) = f(\lambda \boldsymbol{x} + (1 - \lambda)\bar{\boldsymbol{x}}) \leq \lambda f(\boldsymbol{x}) + (1 - \lambda) f(\bar{\boldsymbol{x}}) < f(\bar{\boldsymbol{x}}),$$

con lo que $\boldsymbol{d} \in D^D$ y, por tanto, $\boldsymbol{d} \in D^D \cap D^F$ y $D^D \cap D^F \neq \emptyset$. □

Claramente, $D^D \cap D^F = \emptyset$ siempre será una condición necesaria para tener optimalidad local, independientemente de las propiedades de f y S. Sin embargo, la convexidad es crucial para la otra implicación. Consideremos la función $f : \mathbb{R} \to \mathbb{R}$ definida por

$$f(x) = \begin{cases} 0 & \text{si } x = 0, \\ \sin(\frac{1}{x}) & \text{en otro caso.} \end{cases}$$

La representación de la función f la podemos ver en la Figura 2.2. Claramente, en $x = 0$ tenemos $D^D = \emptyset$, con lo que $D^D \cap D^F = \emptyset$. Sin embargo, $x = 0$ no es un óptimo local. El problema de esta función, en comparación con las funciones convexas, es que ni siquiera tiene definidas las derivadas direccionales en $x = 0$. Por otro lado, nótese que el problema no convexo $\text{minimizar}_{x \in \mathbb{R}} f(x) = x^3$ en $x = 0$ no sirve como contraejemplo del Teorema 2.9. En este caso tenemos que la dirección $d = -1$ pertenece a $D^D \cap D^F$, con lo que 0 no puede ser un óptimo local.

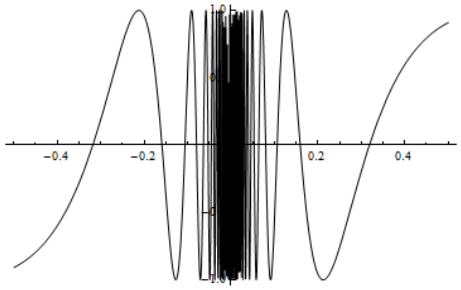
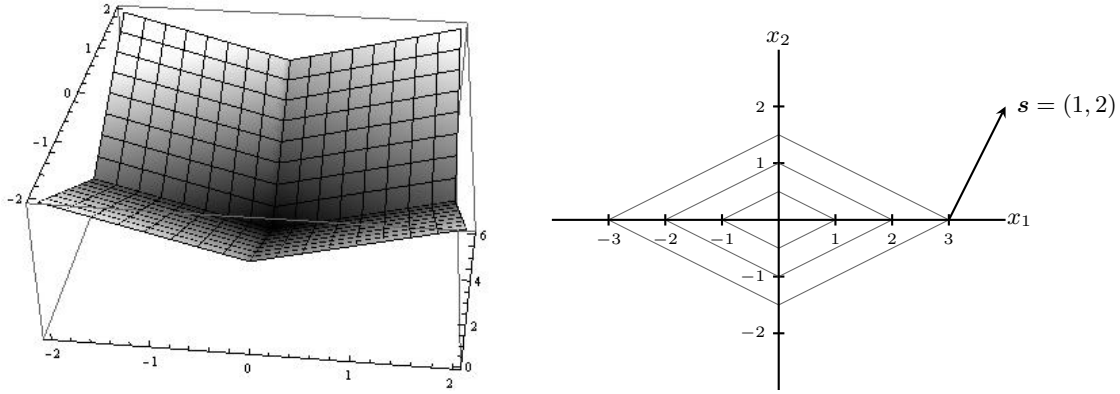


Figura 2.2: Función $f(x) = 0$ si $x = 0$ y $f(x) = \sin(\frac{1}{x})$ en otro caso.



(a) Representación de $f(x_1, x_2) = |x_1| + 2|x_2|$.

(b) Curvas de nivel de $f(x_1, x_2) = |x_1| + 2|x_2|$.

Figura 2.3: Ilustración de la función del Ejemplo 2.1.

Terminamos esta sección con una pequeña discusión relativa a la búsqueda de direcciones de descenso. En el caso diferenciable la Proposición 2.8 implica que, siempre que $\nabla f(\boldsymbol{x}) \neq \mathbf{0}$, tenemos que $-\nabla f(\boldsymbol{x})$ es una dirección de descenso en $\boldsymbol{x}$ pues $-\nabla f(\boldsymbol{x})^\intercal \nabla f(\boldsymbol{x}) = -\|\nabla f(\boldsymbol{x})\| < 0$. De hecho, en la Sección 4.8.1 se demuestra que $-\nabla f(\boldsymbol{x})$ es la dirección de *máximo descenso* de la función f en $\boldsymbol{x}$ (Proposición 4.7).

En el caso no diferenciable la identificación de direcciones de descenso no es tan inmediata, ya que el resultado anterior para el gradiente no aplica necesariamente a los subgradientes. Más concretamente, dado un subgradiente $\boldsymbol{s}$ de f en $\bar{\boldsymbol{x}}$, el vector $-\boldsymbol{s}$ no tiene por qué ser una dirección de descenso. Un ejemplo trivial en el que se puede ver este comportamiento es en la función $f(x) = |x|$. El punto $x = 0$ es un mínimo del problema y, claramente, en él no existe ninguna dirección de descenso. Sin embargo, en el punto $x = 0$ existen infinitos subgradientes no nulos. Si este comportamiento sucediese únicamente cuando nos encontramos en mínimos del problema, entonces no sería muy problemático. Desafortunadamente, como mostramos en el siguiente ejemplo, la situación anterior se puede extender de modo sencillo a situaciones en las que el punto en estudio no cumple ninguna condición de optimalidad.

Ejemplo 2.1. Considérese la función $f : \mathbb{R}^2 \to \mathbb{R}$ dada por $f(x_1, x_2) = |x_1| + 2|x_2|$, representada en la Figura 2.3(a), y cuyas curvas de nivel pueden verse en la Figura 2.3(b).

Es fácil ver que el vector $\boldsymbol{s} = (1, 2)$ es un subgradiente en $\bar{\boldsymbol{x}} = (3, 0)$. Como f es una función totalmente separable en ambas variables, es fácil ver que su subgradiente en un punto (x_1, x_2) se puede expresar como el producto cartesiano del subdiferencial de $|x|$ en x_1 y del subdiferencial de $2|x|$ en x_2. Por tanto, el subdiferencial en $(3, 0)$ se corresponde con $\{1\} \times [-2, 2]$ (nótese que $|x|$ es diferenciable en $x = 3$). En el caso de $\boldsymbol{s} = (1, 2)$ tenemos que, para todo $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2$,

$$f(\boldsymbol{x}) = |x_1| + 2|x_2| \geq \Big(|3| + 1 \cdot (x_1 - 3)\Big) + \Big(2 \cdot |0| + 2 \cdot (x_1 - 0)\Big) = f(\bar{\boldsymbol{x}}) + \boldsymbol{s}^{\mathsf{T}}(\boldsymbol{x} - \bar{\boldsymbol{x}}),$$

con lo que $\boldsymbol{s} = (1, 2)$ es un subgradiente en $\bar{\boldsymbol{x}} = (3, 0)$. Por otro lado, tenemos que $f(3, 0) = 3$ y que, para todo $\lambda \in (0, 3)$, $f(\bar{\boldsymbol{x}} - \lambda \boldsymbol{s}) = f(3 - \lambda, -2\lambda) = 3 - \lambda + 4\lambda = 3 + 3\lambda > 3$, con lo que $-\boldsymbol{s} = -(1, 2)$ no es una dirección de descenso.

Una consecuencia de la discusión anterior relativa a la búsqueda de direcciones de descenso es que no será inmediato adaptar algoritmos basados en el uso de gradientes a problemas no diferenciables usando subgradientes. Por un lado, estos últimos ni siquiera tienen por qué devolver direcciones de descenso y, por otro, pueden no ser fáciles de calcular. En la Sección 4.9.1 se presenta un método de subgradiente como adaptación del método de máximo descenso a problemas no diferenciables y la Proposición 4.17 muestra que, bajo ciertas condiciones, a pesar de que no tiene por qué seguir direcciones de descenso en todas las iteraciones, este algoritmo puede tener un buen comportamiento en la práctica. Por otro lado, en el estudio de la dualidad y, más concretamente, de las formas de resolver el problema dual llevado a cabo en la Sección 5.4.6, se tiene una situación en la que el cálculo de subgradientes no resulta complicado.

2.3 Ejemplo ilustrativo y discusión

En esta sección vamos a discutir e interpretar los resultados de la sección anterior, incluyendo sus implicaciones a la hora de diseñar algoritmos de optimización. Para ello nos vamos a centrar en el caso diferenciable, estudiando el gradiente. Si bien es cierto que algunos de los argumentos que vamos a hacer para el gradiente en un punto $\bar{\boldsymbol{x}}$ se pueden adaptar para los subgradientes en el caso no diferenciable, hay que tener bastante cuidado ya que, como comentamos al final del apartado anterior, un subgradiente no tiene por qué darnos una dirección de descenso.

El Teorema 2.3 nos da una condición necesaria y suficiente de optimalidad global para problemas de programación convexa. Además, en el caso de funciones diferenciables sobre dominios abiertos, esta condición se reduce a la habitual de que se anule el gradiente (y, por tanto, todas las derivadas direccionales).

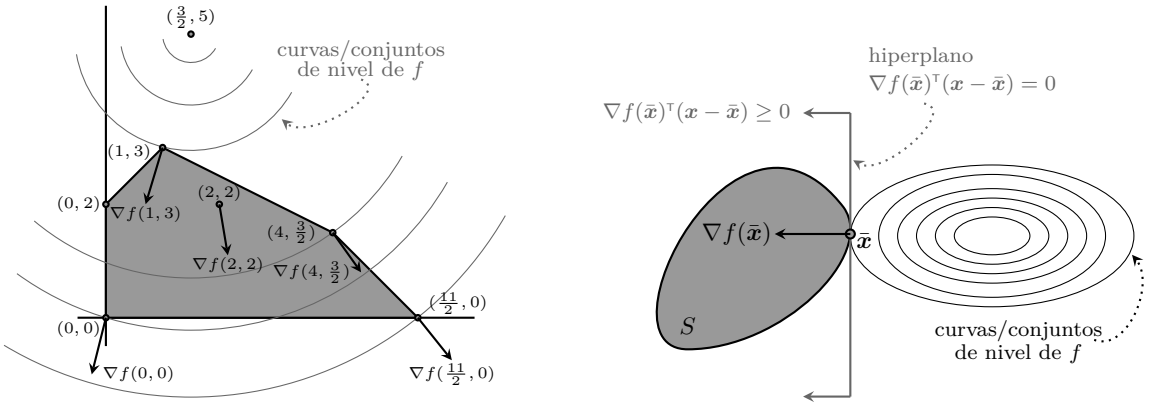
Toda la discusión la llevaremos a cabo sobre el siguiente problema de minimización, al que nos referiremos como Problema P:

$$\begin{aligned} \text{minimizar} \quad & f(\boldsymbol{x}) = (x_1 - \tfrac{3}{2})^2 + (x_2 - 5)^2 \\ \text{sujeto a} \quad & -x_1 + x_2 \leq 2 \\ & x_1 + 2x_2 \leq 7 \\ & 2x_1 + 2x_2 \leq 11 \\ & x_1 \geq 0 \\ & x_2 \geq 0. \end{aligned}$$

El Problema P consiste en encontrar el punto del conjunto S que minimiza la distancia al punto $(\frac{3}{2}, 5)$, donde S es el conjunto definido por las restricciones. Claramente, el conjunto S es un conjunto convexo, pues es un poliedro (intersección de semiespacios). Podemos preguntarnos además si la función f es convexa. Como claramente es una función dos veces diferenciable podemos recurrir a su matriz hessiana. En este caso, como tenemos una función cuadrática tenemos que la hessiana es constante: para todo $\boldsymbol{x} \in \mathbb{R}^2$, la hessiana de la función f en $\boldsymbol{x}$ viene dada por

$$\boldsymbol{H}(\boldsymbol{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

Como esta matriz tiene todos sus autovalores positivos, es definida positiva y, por tanto, la función f es estrictamente convexa. Entonces, como f es convexa y S es convexo, P es un problema de programación convexa y podemos aplicar todos los resultados de la sección anterior. En la Figura 2.4(a) representamos gráficamente el problema P.



(a) Ilustrando el Teorema 2.3 y sus corolarios.

(b) Geometría del Teorema 2.3.

Figura 2.4: Ilustrando los resultados de optimalidad global de funciones convexas.

Antes de avanzar en la discusión del problema P, hacemos una pequeña discusión acerca del gradiente y de los productos de la forma $\nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}})$. Recordemos que el gradiente de una función diferenciable f en un punto $\bar{\boldsymbol{x}}$ viene dado por

$$\nabla f(\bar{\boldsymbol{x}}) = \Big(\frac{\partial f(\bar{\boldsymbol{x}})}{\partial x_1}, \frac{\partial f(\bar{\boldsymbol{x}})}{\partial x_2}, \ldots, \frac{\partial f(\bar{\boldsymbol{x}})}{\partial x_n}\Big).$$

Cada componente del gradiente nos dice cuánto crecería/decrecería la función (localmente) si nos movemos en la dirección marcada por el correspondiente vector de la base canónica. Tomemos ahora un punto del interior de S, por ejemplo el punto $(2, 2)$, tenemos que $f(2, 2) = 9.25$ y $\nabla f(2, 2) = (1, -6)^\intercal$. Supongamos ahora que queremos ver si $(2, 2)$ es un óptimo global. Como estamos en un problema de programación convexa, optimalidad global equivale a optimalidad local, con lo que podemos ver qué pasa si nos movemos a puntos que están cerca de $(2, 2)$. A continuación discutimos distintos movimientos que se podrían llevar a cabo desde el punto $(2, 2)$:

- Podemos movernos en la dirección $(1,0)^\intercal$. Por ejemplo, pasando al punto $(2.5,2) = (2,2) + 0.5 \cdot (1,0) \in S$. En este caso, nos encontramos con que $f(2.5,2) = 10 > 9.25 = f(2,2)$. Es decir, el movimiento que hemos realizado ha empeorado la función objetivo. Esto no debería sorprendernos, pues $\nabla f(2,2) = (1,-6)^\intercal$ ya nos está diciendo que en el punto $(2,2)$ la función está creciendo en la dirección $(1,0)^\intercal$ a velocidad 1.

- Podemos movernos en la dirección $(0,1)^\intercal$. Por ejemplo, pasando al punto $(2,2.5) = (2,2) + 0.5 \cdot (0,1) \in S$. Ahora tenemos $f(2,2.5) = 6.5 < 9.25 = f(2,2)$, con lo que hemos mejorado. Nuevamente, esto ya lo podíamos anticipar dado que el gradiente nos dice que en la dirección $(0,1)^\intercal$ la función decrece a velocidad 6. De hecho, esto también explica que al pasar de $(2,2)$ a $(2,2.5)$ hayamos mejorado más de lo que empeoramos cuando pasamos al punto $(2.5,2)$.

- Es importante destacar que las velocidades 1 y -6 son locales, y lo que realmente nos dicen es que, si nos alejamos muy poco en las direcciones $(1,0)$ y $(0,1)$, las tasas de variación estarán cerca de 1 y -6, respectivamente. Pero si el alejamiento es grande estos números pueden perder su significado. Por ejemplo, el gradiente nos dice que si nos movemos desde $(2,2)$ en la dirección $(-1,0)^\intercal$, el valor de f debería decrecer. Supongamos entonces que tomamos el punto $(0,2) = (2,2) + 2 \cdot (-1,0) \in S$. En este caso, tenemos que $f(0,2) = 11.25 > 9.25 = f(2,2)$ con lo que, en contra de lo que sugería el gradiente, hemos empeorado.

- Veamos ahora qué pasa si nos movemos en la dirección $-\nabla f(2,2) = (-1,6)^\intercal$, por ejemplo al punto $(1.95,2.3) = (2,2) + 0.05 \cdot (-1,6) \in S$. En este caso tenemos que $f(1.95,2.3) = 7.4925 < 9.25 = f(2,2)$. Como el gradiente nos marca una dirección de crecimiento de la función, moverse en dirección contraria suele ser una buena opción cuando queremos minimizar. De hecho, $-\nabla f(\boldsymbol{x})$ es la dirección de máximo decrecimiento/descenso de la función f en el punto $\boldsymbol{x}$ (localmente).

- Volvamos ahora al primer movimiento, del punto $(2,2)$ al punto $(2.5,2)$. Tenemos que

$$\nabla f(2,2)^\intercal \Big((2.5,2) - (2,2) \Big) = \nabla f(2,2)^\intercal (0.5,0) = 0.5 \geq 0.$$

  Es decir, si tomamos $\bar{\boldsymbol{x}} = (2,2)$ y $\boldsymbol{x} = (2.5,2)$, el producto de la forma $\nabla f(\bar{\boldsymbol{x}})^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}})$ es no negativo, con lo que $\boldsymbol{d} = (0.5,0)$ no es una dirección de descenso (Proposición 2.8). Análogamente, por la convexidad de f tenemos $f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}) > f(\bar{\boldsymbol{x}})$ y ya vemos que movernos de $\bar{\boldsymbol{x}}$ a $\boldsymbol{x}$ no iba a mejorar la función objetivo. Aquí es importante destacar una característica de las funciones convexas. Como comentamos en el punto anterior, en general el gradiente sólo nos da condiciones locales. Sin embargo, cuando trabajamos con funciones convexas hay información local que que se puede hacer global. En este caso, el gradiente nos decía que, localmente, moverse en la dirección $(1,0)^\intercal$ iba a aumentar la función objetivo. La convexidad nos dice que ningún movimiento en esa dirección, sea pequeño o no, mejorará la función objetivo. Dado $\lambda > 0$, y $\boldsymbol{x} = \bar{\boldsymbol{x}} + \lambda(1,0)$, tenemos que $f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}) = f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\intercal \lambda \cdot (1,0) = f(\bar{\boldsymbol{x}}) + \lambda > f(\bar{\boldsymbol{x}})$.

El análisis que acabamos de llevar a cabo nos permite sacar un par de conclusiones iniciales de cara al diseño de algoritmos de optimización basados en la búsqueda de direcciones de

descenso de la función: i) incluso encontrando una buena dirección de descenso, es importante tener un buen método para decidir cuánto nos movemos en dicha dirección y ii) las direcciones de descenso hay que buscarlas entre aquellas direcciones $\boldsymbol{d} \in \mathbb{R}^n$ tales que $\nabla f(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} \leq 0$ (con menor estricto si tenemos asegurada la convexidad). Dicho de otra forma, direcciones que formen un ángulo de al menos 90° con el gradiente. En el caso del punto $(2, 2)$, como pertenece al interior del conjunto S siempre podremos movernos en alguna dirección que cumpla esa propiedad, salvo que tengamos $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$, que es precisamente la condición necesaria y suficiente de optimalidad en un punto del interior de la región factible.

La discusión anterior nos da un procedimiento bastante natural para la búsqueda de puntos óptimos. Supongamos que estamos en un punto $\bar{\boldsymbol{x}} \in S$ y encontramos otro punto $\boldsymbol{x} \in S$ tal que $\nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) < 0$. Entonces, esto nos está indicando que moviéndonos en la *dirección de descenso* $\boldsymbol{d} = (\boldsymbol{x} - \bar{\boldsymbol{x}})$ podremos mejorar la función objetivo. En la práctica, es habitual hacer lo siguiente: una vez tenemos identificada una dirección de descenso $\boldsymbol{d}$, se decide cuánto moverse en esa dirección resolviendo lo que se conoce como *búsqueda lineal* (*line search* en inglés). Una búsqueda lineal consiste en resolver un subproblema de minimización unidimensional, en este caso dado por

$$\begin{array}{rl} \text{minimizar} & f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) \\ \text{sujeto a} & \bar{\boldsymbol{x}} + \lambda \boldsymbol{d} \in S \\ & \lambda \geq 0. \end{array}$$

Como veremos más adelante, este tipo de algoritmos pueden ser muy efectivos, pues es fácil definirlos de tal manera que aseguren la convergencia a un óptimo local (y, bajo convexidad, global) y consisten en resolver una sucesión de problemas de optimización esencialmente unidimensionales.

Si miramos ahora a la Figura 2.4(b), tenemos un ejemplo en el que un punto $\bar{\boldsymbol{x}}$ en el que $\nabla f(\bar{\boldsymbol{x}}) \neq \mathbf{0}$ es un mínimo global. Lo que sucede en este caso, es que el conjunto S está completamente contenido en el semiespacio $\{\boldsymbol{x} : \nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \geq 0\}$. Es decir, todas las direcciones hacia las que me puedo mover desde $\bar{\boldsymbol{x}}$ sin salir del conjunto factible S son direcciones de crecimiento, con lo que no nos sirven para reducir el valor de la función objetivo. Eso es justamente lo que me dice el Teorema 2.9: $D^D \cap D^F = \emptyset$.

Si volvemos al Problema P y miramos el punto $(1, 3)$, tenemos que $\nabla f(1, 3) = (-1, -4)^\intercal$. En la Figura 2.4(a) se puede ver que este vector forma un ángulo menor de 90° con cualquier vector $(x_1 - 1, x_2 - 3)$ con $(x_1, x_2) \in S$. Por tanto, la discusión anterior nos asegura que, efectivamente, $(1, 3)$ es un óptimo global del problema. Como ilustración, podemos ver lo que pasa tomando $\boldsymbol{x} = (0, 2)$. En este caso, $(x_1 - 1, x_2 - 3) = (-1, -1)$ y al multiplicarlo por el $\nabla f(1, 3)$ obtenemos $(-1) \cdot (-1) + (-1) \cdot (-4) = 5 > 0$.

Asimismo, miremos ahora qué pasa con el punto $(0, 0)$, que también pertenece a la frontera de S. En este caso, la situación es la opuesta a la que teníamos para el punto $(1, 3)$: es fácil verificar gráficamente que, para todo $\boldsymbol{x} \in S$, $\nabla f(0, 0)^\intercal \cdot (\boldsymbol{x} - (0, 0)) < 0$. Por tanto, cualquier dirección de la forma $\boldsymbol{x} - (0, 0)$, con $\boldsymbol{x} \in S$, sería una dirección de descenso. En general, no sólo es importante elegir de forma adecuada cuánto nos movemos en una cierta dirección de descenso, sino que también es importante saber qué dirección de descenso elegir cuando tenemos muchas a nuestra disposición. Como ya comentamos, en el caso diferenciable, la dirección opuesta al gradiente es una buena candidata (en el caso no diferenciable habría que ver cómo elegir entre los distintos subgradientes).

Si nos fijamos nuevamente en el punto $(0,0)$ y en las curvas de nivel, está claro que estamos ante un máximo local. Del mismo modo, también el punto $(\frac{11}{2},0)$ es un máximo local y $f(0,0) = 27.25 < 41 = f(\frac{11}{2},0)$. Es decir, tenemos un óptimo local que no es un óptimo global; esto es porque al ser el problema de maximizar, para tener que optimalidad local implica optimalidad global necesitaríamos que la función f fuese cóncava en vez de convexa.

Una vez hecha esta discusión acerca de lo que hay detrás del Teorema 2.3, podemos pensar un poco en el papel de las distintas hipótesis del mismo. Para ello, nos centraremos en el caso diferenciable, es decir, en el Corolario 2.5.

***S* no convexo.** En este caso, de la primera parte de la demostración del Teorema 2.3 se deduce que la condición de que, para todo $\boldsymbol{x} \in S$, $\nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x}-\bar{\boldsymbol{x}}) \geq 0$, sigue siendo suficiente para asegurar la optimalidad global de $\bar{\boldsymbol{x}}$. Todo lo que hace falta es apoyarse en la convexidad de f. Para verlo, supongamos que existe $\hat{\boldsymbol{x}} \in S$ tal que $f(\bar{\boldsymbol{x}}) > f(\hat{\boldsymbol{x}})$. Entonces, por la convexidad de f,

$$f(\bar{\boldsymbol{x}}) > f(\hat{\boldsymbol{x}}) \geq f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\intercal(\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}).$$

Pero esto implicaría que $\nabla f(\bar{\boldsymbol{x}})^\intercal(\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}) < 0$. Fijémonos que esto nos permite interpretar la condición de que, para todo $\boldsymbol{x} \in S$, $\nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \geq 0$, como que el hiperplano $\{\boldsymbol{x} : \nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) = 0\}$ separa S de las soluciones que mejoran con respecto a $\bar{\boldsymbol{x}}$.

***f* no convexa.** En este caso, la condición $\nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x}-\bar{\boldsymbol{x}}) \geq 0$ ni siquiera asegura que estemos ante un mínimo local:

- La función $f(x) = x^3$ cumple la condición en $x = 0$ pues $f'(0) = 0$ y 0 no es un óptimo local.
- La función $f(x) = x^2(\text{sen}(x)^2 + \frac{x^2}{100} + 1)$, usada al principio de la Sección 1.3 (ver Figura 1.8(b)) tiene infinitos puntos en los que se anula la derivada y, por tanto, infinitos puntos donde se cumple la condición. Sin embargo, infinitos de ellos son máximos locales, infinitos mínimos locales y sólo uno es un mínimo global.

***f* diferenciable no convexa y *S* convexo.** Pensemos ahora en la otra dirección del resultado del Corolario 2.5. Supongamos que $\bar{\boldsymbol{x}} \in S$ es un mínimo global. Entonces, necesariamente tendremos que tener que, para todo $\boldsymbol{x} \in S$, $\nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \geq 0$. En otro caso, si existe $\hat{\boldsymbol{x}}$ tal que $\nabla f(\bar{\boldsymbol{x}})^\intercal(\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}) < 0$, tendríamos que, moviéndonos en la dirección $\boldsymbol{d} = \hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}$ hacia puntos de la forma $\bar{\boldsymbol{x}} + \lambda\boldsymbol{d}$, si λ es lo suficientemente pequeño permaneceríamos dentro del conjunto S (pues es convexo) y el valor de la función objetivo se reduciría (definición de gradiente).

***f* diferenciable no convexa y *S* no convexo.** La discusión anterior nos permite llegar a un concepto más general. Si tenemos simplemente diferenciabilidad y $\bar{\boldsymbol{x}}$ es un mínimo local y hay alguna dirección $\boldsymbol{d}$ en la que $\bar{\boldsymbol{x}} + \lambda\boldsymbol{d} \in S$ si λ es suficientemente pequeño, entonces $\nabla f(\bar{\boldsymbol{x}})^\intercal\boldsymbol{d} \geq 0$. Dicho en palabras, si estamos en un óptimo local y la función objetivo se reduce en una dirección, entonces es porque esa dirección nos echa fuera del conjunto.

2.4 Generalizaciones del concepto de función convexa

Como hemos visto en este tema y en el anterior, las funciones convexas tienen propiedades que las hacen bastante especiales, como la garantía de existencia de derivadas direccionales y de subgradientes, o la garantía de que todo mínimo local es un mínimo global.

Para terminar este tema presentamos una serie de definiciones que generalizan la idea de convexidad y cada una de ellas mantiene algunas de las principales propiedades de las funciones convexas con lo que, según el tipo de funciones que tengamos, puede ser importante verificar si se encuentran en alguna de estas clases. Nuevamente, a partir de las variaciones de la idea de convexidad que presentamos en esta sección se pueden definir variaciones de la concavidad sin más que cambiar el signo de la función f. En la Figura 2.5 presentamos un esquema con las diferentes clases de convexidad y las relaciones lógicas entre ellas.

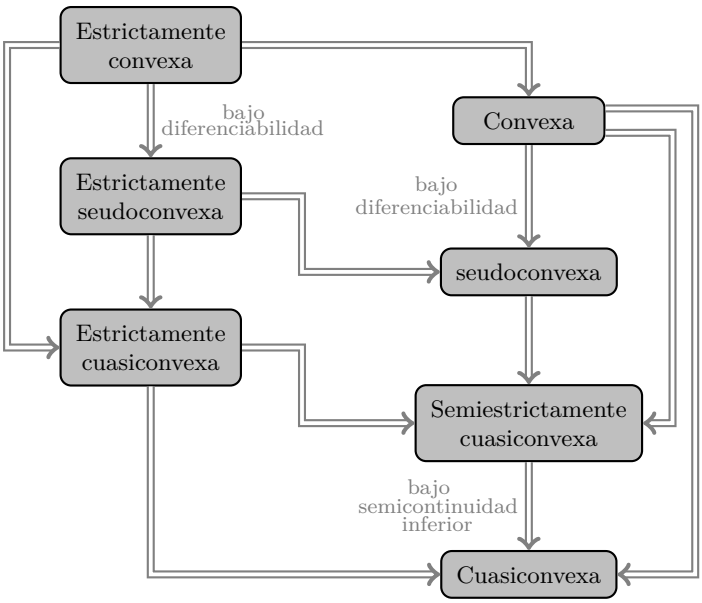


Figura 2.5: Relaciones entre los distintos tipos de convexidad.

Definición 2.3. Sea $S \subseteq \mathbb{R}^n$ conjunto no vacío y convexo y sea $f : S \to \mathbb{R}$. Entonces,

(i) La función f es *cuasiconvexa* si, para todo $\boldsymbol{x} \in S$ e $\boldsymbol{y} \in S$ y para todo $\lambda \in (0,1)$, se tiene

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \text{máx}\{f(\boldsymbol{x}), f(\boldsymbol{y})\}.$$

La función f es *semiestrictamente cuasiconvexa* si, siempre que $f(\boldsymbol{x}) \neq f(\boldsymbol{y})$, la desigualdad anterior es estricta. La función f es *estrictamente cuasiconvexa* si, siempre que $\boldsymbol{x} \neq \boldsymbol{y}$, la desigualdad anterior es estricta.

(ii) Supongamos que f es diferenciable en S. Entonces, f es *seudoconvexa* si, para todo $\boldsymbol{x} \in S$ e $\boldsymbol{y} \in S$, $f(\boldsymbol{y}) < f(\boldsymbol{x})$ implica que $\nabla f(\boldsymbol{x})^\intercal(\boldsymbol{y} - \boldsymbol{x}) < 0$.

La función f es *estrictamente seudoconvexa* si, siempre que $\boldsymbol{x} \neq \boldsymbol{y}$, $f(\boldsymbol{y}) \leq f(\boldsymbol{x})$ implica que $\nabla f(\boldsymbol{x})^\intercal(\boldsymbol{y} - \boldsymbol{x}) < 0$.

A continuación presentamos una recolección de las propiedades más importantes de las funciones convexas y vemos cuáles se mantienen cuando estudiamos las distintas generalizaciones (ver Figura 2.6).

- **Epigrafo convexo.** Funciones convexas y estrictamente convexas.

  De hecho, vimos que función convexa si y sólo si epigrafo convexo (Teorema 1.14).

- **Conjuntos de nivel inferiores convexos.** Funciones convexas y estrictamente convexas, seudoconvexas y estrictamente seudoconvexas, cuasiconvexas y estrictamente cuasiconvexas.

  En este caso, se puede probar que una función es cuasiconvexa si y sólo si los conjuntos de nivel inferiores son convexos.

- **Óptimo local implica óptimo global.** Funciones convexas y estrictamente convexas, seudoconvexas y estrictamente seudoconvexas, estrictamente cuasiconvexas y semiestrictamente cuasiconvexas.

- **Unicidad óptimo global.** Estrictamente convexas, estrictamente seudoconvexas y estrictamente cuasiconvexas.

- **$\nabla f(\bar{x}) = 0 \Rightarrow$ óptimo global.** Funciones convexas y estrictamente convexas, seudoconvexas y estrictamente seudoconvexas.

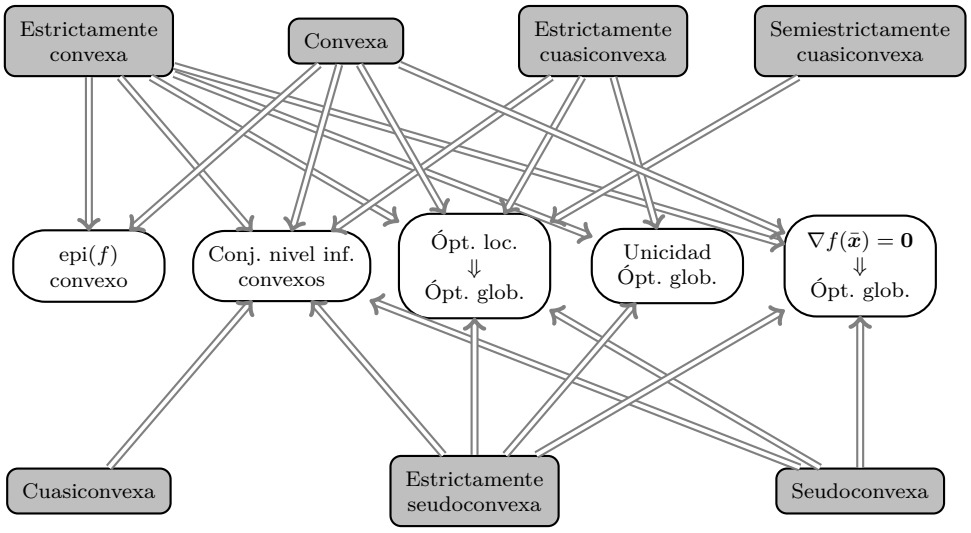


Figura 2.6: Relaciones entre los distintos tipos de convexidad.

2.5 Ejercicios adicionales

•**Ejercicio 2.5.** El comité ejecutivo de una gran constructora tiene que elegir entre $p \in \mathbb{N}$ proyectos que se le han presentado para la construcción de un puente. Cada uno de estos

proyectos viene caracterizado por un vector $\boldsymbol{x}^j \in \mathbb{R}^n$, donde cada componente del vector $\boldsymbol{x}^j$ representa distintas variables relativas a la estructura a construir: longitud, altura, grosor, número de pilares,...

Para evaluar los distintos proyectos la empresa dispone de una serie de funciones f_k, con $k \in \{1, 2, \ldots, 100\}$. Estas funciones son convexas y $f_k(\boldsymbol{x})$ representa una medida del riesgo asociado a la opción $\boldsymbol{x}$. Supongamos que la empresa está interesada en elegir la solución que minimize el máximo riesgo de acuerdo a estas funciones.

- Modeliza el problema de optimización al que se enfrenta este comité ejecutivo, escribiendo su formulación como un problema de programación matemática.
- ¿Estamos ante un problema de optimización convexa? ◁

••Ejercicio 2.6. El departamento de riesgos de un banco tiene que elegir entre distintas carteras de inversión. Cada cartera de inversión viene dada por un vector $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \geq \mathbf{0}$, donde cada componente x_i representa el dinero invertido en comprar acciones de la empresa i. Además, $\sum_{i=1}^n x_i = b$, donde b es el presupuesto que la ejecutiva del banco ha decidido invertir. Tras un exhaustivo análisis, desde el departamento han reducido a un centenar de posibilidades los escenarios que se podrán producir en el futuro, cada uno de ellos con una probabilidad estimada de ocurrencia p_k. Además, para cada escenario existe una función cóncava f_k tal que $f_k(\boldsymbol{x})$ denota el beneficio obtenido por la empresa si finalmente se produce el escenario k.

Modeliza el problema de optimización al que se enfrentaría el departamento de riesgos en cada uno de los siguientes casos, escribiendo su formulación como un problema de programación matemática. Razona además para cada uno de ellos si estamos o no ante un problema de optimización convexa:

- El departamento de riesgos quiere maximizar el beneficio esperado con su inversión.
- Estamos ante un departamento conservador que lo único que quiere es maximizar el beneficio que se produciría en el peor caso.
- El departamento de riesgos busca un compromiso entre beneficio y riesgo, con lo que su objetivo será de la forma maximizar $\alpha E(\boldsymbol{x}) - (1-\alpha) R(\boldsymbol{x})$. Donde $E(\boldsymbol{x})$ denota el beneficio esperado y $R(\boldsymbol{x})$ es una medida de riesgo dada, en este caso, por la semivarianza asociada a la solución $\boldsymbol{x}$, y se define como $SV(\boldsymbol{x}) = \sum_{k=1}^{100} \left(\text{mín}\{f_k(\boldsymbol{x}) - E(\boldsymbol{x}), 0\} \right)^2$. ◁

••Ejercicio 2.7. Considera el siguiente problema de optimización:

$$\begin{array}{rl} \text{minimizar} & f(\boldsymbol{x}) = (x_1 - 4)^2 + (x_2 - 6)^2 \\ \text{sujeto a} & x_2 \geq x_1^2 \\ & x_2 \leq 4. \end{array}$$

Escribe una condición necesaria y suficiente de optimalidad global. Verifica si se cumple o no para el punto $(2, 4)$. ◁

Tema 3
Lenguajes de Modelización: AMPL

Contenidos

3.1 Modelización de problemas de optimización matemática

A la hora de transcribir a un ordenador un problema general de optimización matemática, nos encontramos con una situación de mayor complejidad que en el caso de estar ante problemas de programación lineal y entera, cuya descripción pasa esencialmente por especificar la matriz de restricciones, el vector de lados derechos, el vector de costes y, en caso de ser necesario, la especificación de las variables que han de tomar valores enteros. Esto hace que sea relativamente sencillo definir un estándar para la formulación de problemas de programación lineal y entera, gracias al cual es relativamente sencillo resolver un mismo problema con distintas herramientas de optimización lineal y comparar el rendimiento de las mismas.

En el caso de los optimizadores para problemas de optimización no lineal, habitualmente llamados *solvers* por su nombre en inglés, es habitual que cada una trabaje internamente la descripción del problema de optimización que se adapte mejor al lenguaje de programación en el que se halla desarrollado la herramienta y al tipo de algoritmos implementados en la misma.

Esto plantea la necesidad de herramientas que permitan una modelización ágil de problemas generales de optimización no lineal y que después se encarguen de traducir/adaptar dichos modelos a las necesidades de cada herramienta de optimización. Esto facilitará la tarea de probar distintos optimizadores y algoritmos sobre un mismo problema de optimización.

Las herramientas que acabamos de describir se conocen como lenguajes de modelización algebraicos. Dos lenguajes clásicos y de uso ampliamente extendido son GAMS (*General Algebraic Modeling System*, Boisvert y otros (1985)) y AMPL (*A Mathematical Programming Language*, Fourer y otros (1990)) que, aún siendo de pago, permiten un uso completo de los mismos a través de licencias académicas. Además, en los últimos años han salido multitud de alternativas gratuitas, desarrolladas como librerías de distintos lenguajes de programación como por ejemplo Pyomo para Python (*Python Optimization Modelling* (Hart y otros, 2012)) y JuMP para Julia (*Julia for Mathematical Optimization* (Dunning y otros, 2017)).

3.2 Lenguaje AMPL

Durante esta asignatura usaremos el lenguaje AMPL con una licencia académica, bajo la cual se compartirá un enlace de descarga de una versión sin restricciones del entorno de trabajo de AMPL.

En estas notas no vamos a incluir explicaciones relativas al manejo del lenguaje AMPL. En su lugar presentaremos algunas indicaciones para su uso e instalación así como algunas referencias, que se facilitarán durante el curso mediante un enlace de descarga.

Es conveniente destacar que además de servir como lenguaje de modelización algebraico, AMPL también puede usarse como lenguaje de programación, pues admite la elaboración de scripts con bucles, sentencias condicionales,... Sin embargo, aun siendo útil a nivel académico, las funcionalidades de programación de AMPL son bastante limitadas, siendo mucho más eficiente la programación directa de scripts en otros lenguajes que después se enganchen con AMPL a través de las correspondientes APIs (*Application Program Interface*).

3.2.1 Opciones de "Instalación" y "Ejecución"

Opción 1. Instalar una versión de AMPL con su interfaz (AMPL IDE, *Integrated Development Environment*) y resolviendo los problemas desde el propio programa. Una versión completa de AMPL con los optimizadores asociados, con una licencia académica para el presente curso, puede descargarse de:[1]

- **Windows 64bit:** http://bit.ly/AMPLWin64PM20-21
- **Linux 64bit:** http://bit.ly/AMPLLinux64PM20-21
- **Mac-OS 64bit:** http://bit.ly/AMPLMac64PM20-21. Interfaz AMPL IDE para Mac-OS en fichero aparte: http://bit.ly/AMPLMacInterf20-21.

En la página https://ampl.com/try-ampl/ampl-for-courses/ampl-course-install/ puede verse información adicional relativa a la "instalación" de la herramienta.

Opción 2. Sin instalar nada. Es posible trabajar directamente con AMPL editando el código desde cualquier editor de texto y llamando a las herramientas de optimización a través de la plataforma online de NEOS (https://neos-server.org/neos/).

La principal ventaja de la primera opción es que las ejecuciones se hacen localmente en el propio ordenador, con lo que los resultados se escriben directamente en ficheros. Por otro lado, lanzar problemas sobre NEOS tiene la ventaja de no requerir instalación y, en el caso de querer resolver problemas que consuman bastante tiempo, no tendremos nuestro ordenador saturado con dicha tarea pues la ejecución del optimizador se hace remotamente. En cualquier caso, no es esperable que en el curso de esta asignatura se vayan a resolver problemas computacionalmente muy exigentes.

3.2.2 Herramientas de optimización disponibles

En la web de NEOS (http://www.neos-server.org/neos/solvers/) hay una gran cantidad de optimizadores (*solvers*) disponibles para ejecución online a través de AMPL. Estos optimizadores incluyen, entre otros, Gurobi, CPLEX y FICO-Xpress para problemas de programación lineal y entera y BARON, Couenne , CONOPT, Ipopt, Knitro, LOQO, MINOS y SNOPT para problemas no lineales (cabe destacar que BARON y Couenne son los únicos de los anteriores que pueden garantizar optimalidad global de las soluciones obtenidas).

En caso de optar por trabajar localmente con AMPL, la mayoría de los optimizadores que acabamos de mencionar también estarán disponibles. Además, se pueden incluir fácilmente optimizadores adicionales sin más que descargar la versión correspondiente al sistema operativo del ordenador y copiarla en la carpeta de la instalación de AMPL:

Optimizadores de código abierto. Destacan el optimizador local Ipopt y el optimizador global Couenne: https://ampl.com/products/solvers/open-source/.

[1] Alternativamente, se puede descargar directamente una versión de demostración de http://ampl.com/try-ampl/download-a-demo-version/, que estará limitada a problemas con un número relativamente reducido variables.

Otros optimizadores gratuitos. Existe un amplio catálogo de optimizadores, algunos diseñados para problemas más generales y otros más específicos, que han sido "enganchados" con AMPL y pueden descargarse típicamente desde las páginas web de los desarrolladores: https://ampl.com/products/solvers/all-solvers-for-ampl/.

3.2.3 Referencias para aprender AMPL

En la web se pueden encontrar multitud de referencias para aprender AMPL. Durante esta asignatura se facilitarán, entre otros, los siguientes tutoriales cuyos contenidos se adaptan bien al grado de profundidad que requerirá esta asignatura:

- El tutorial "AMPL: A Mathematical Programming Language" (Fourer y otros, 1990) escrito por los padres de AMPL es un buen comienzo.
- Aunque quizá como primera toma de contacto sea más accesible "Introduction to AMPL: A Tutorial", escrito por Kaminsky y Rajan.
- También se facilitarán par de referencias en castellano, una escrita en el año 2014 por Jorge Hans Alayo Gamarra y otra en el año 2000 por Pedro Luis Luque.

Además de los tutoriales que acabamos de mencionar, el alumno interesado también podrá acceder libremente a los siguientes recursos:

- El libro de AMPL (http://ampl.com/resources/the-ampl-book/chapter-downloads/), que está disponible gratis en la web de AMPL y tiene multitud de contenidos y ejemplos.
- Información acerca de las funciones disponibles en AMPL en el manual de referencia (http://ampl.com/BOOK/CHAPTERS/24-refman.pdf) incluido como apéndice del libro de AMPL.

Tema 4

Optimización sin Restricciones. Algoritmos

Contenidos

4.1 Introducción

Aunque la mayoría de los problemas que surgen en la práctica son problemas con restricciones, hay bastantes motivos por los que es importante estar familiarizado con las técnicas de optimización sin restricciones:

(i) Los problemas sin restricciones también son habituales en la realidad.

(ii) Muchos algoritmos de optimización con restricciones resuelven el problema de partida transformándolo en una sucesión de problemas no restringidos, ya sea mediante el uso de métodos de penalización y métodos de barrera o mediante multiplicadores de Lagrange y conceptos de dualidad (como veremos en el Tema 7).

(iii) La mayoría de los métodos se basan en encontrar una dirección de descenso apropiada y después elegir el desplazamiento óptimo en dicha dirección. La búsqueda de dicho desplazamiento se reduce habitualmente a minimizar una función de una variable sin restricciones (o con restricciones sencillas como cotas inferiores y superiores para las variables).

(iv) Una buena parte de las técnicas e ideas de la optimización sin restricciones se pueden extender y adaptar de modo natural para dar lugar a métodos de resolución para problemas con restricciones.

4.2 Algoritmos

En el resto de estas notas presentaremos una buena cantidad de algoritmos y, aunque a estas alturas suponemos ya una familiarización con el concepto de algoritmo, presentamos aquí unas pinceladas con ciertas observaciones que pueden ser de relevancia.

4.2.1 El concepto de algoritmo

Recordemos que nuestro objetivo principal consiste en encontrar una solución óptima de un problema de optimización P de la forma:

$$\begin{array}{rl} \text{minimizar} & f(\boldsymbol{x}) \\ \text{sujeto a} & \boldsymbol{x} \in S. \end{array}$$

El procedimiento de resolución consistirá en el empleo de algoritmos. En nuestro caso, se trata de procesos iterativos que, siguiendo unas instrucciones prefijadas, generan una sucesión de puntos con la intención de que converjan a un punto con las propiedades deseadas.

Siendo un poco más formales, un algoritmo se puede representar mediante la llamada *correspondencia algorítmica*, $\boldsymbol{A}$, que, a partir de un punto inicial $\boldsymbol{x}^1$, genera una sucesión $\boldsymbol{x}^1$, $\boldsymbol{x}^2$, $\boldsymbol{x}^3$,... donde $\boldsymbol{x}^{t+1} \in \boldsymbol{A}(\boldsymbol{x}^t)$. Cada una de estas aplicaciones de la correspondencia $\boldsymbol{A}$ constituye una iteración del algoritmo.

Idealmente, lo que desearíamos es que tener algoritmos que construyesen sucesiones de puntos convergentes a un óptimo global del problema. Sin embargo, es habitual tener que

conformarse con objetivos menos ambiciosos, especialmente cuando estamos fuera del dominio de problemas de optimización convexa (y perdemos la propiedad de que todo óptimo local es un óptimo global). Debido a esto, es habitual que los algoritmos se diseñen de tal manera que terminen cuando se alcanzan puntos de un cierto conjunto deseable Ω. Por ejemplo, podríamos tener lo siguiente:

- $\Omega = \{\bar{\boldsymbol{x}} : \bar{\boldsymbol{x}}$ es un óptimo local de P$\}$.
- $\Omega = \{\bar{\boldsymbol{x}} : f(\bar{\boldsymbol{x}}) \leq b\}$, donde b es un valor aceptable de la función objetivo.
- $\Omega = \{\bar{\boldsymbol{x}} : f(\bar{\boldsymbol{x}}) - \nu^* \leq \varepsilon\}$, donde ν^* es el valor de la función objetivo en el óptimo (que se supondría conocido en este caso) y ε es una precisión prefijada.
- $\Omega = \{\bar{\boldsymbol{x}} : \bar{\boldsymbol{x}}$ cumple las condiciones de Karush-Kuhn-Tucker de P$\}$. Estas condiciones, muy relacionadas con el concepto de óptimo local, las veremos en el tema siguiente.

Una parte muy importante del desarrollo de algoritmos es precisamente asegurar, mediante demostraciones matemáticas, que la sucesión generada por un algoritmo siempre converge a un punto del conjunto deseable en cuestión. En caso de que esto no sea posible, habrá que buscar condiciones suficientes a imponer sobre el problema P o sobre la solución inicial $\boldsymbol{x}^1$ que aseguren dicha convergencia. Llevar a cabo este tipo de análisis para los distintos algoritmos que presentaremos en este y otros temas requeriría introducir algunos conceptos y resultados auxiliares. Sin embargo, en lo que a algoritmos se refiere, nosotros nos centraremos en las definiciones e intuiciones de los mismos y, aunque presentaremos los resultados relativos a su convergencia, no los demostraremos.

Otra cuestión importante en el desarrollo de algoritmos iterativos es la dependencia de la solución final del iterante inicial. Supongamos, por ejemplo, que tenemos un algoritmo que tiene asegurada la convergencia al conjunto Ω, dado por los óptimos locales del problema P. Es muy posible que dependiendo del iterante inicial podamos terminar en óptimos locales con valores muy distintos de la función objetivo. Aunque esta cuestión no la vamos a discutir en estos apuntes, es de gran importancia en el desarrollo de algoritmos que encuentren soluciones con buenas propiedades globales.[1]

Supongamos ahora que tenemos dos algoritmos para resolver el mismo problema y que ambos tienen asegurada la convergencia a puntos de un conjunto deseable Ω. En este caso, a la hora de decidir entre ambos suele ser habitual compararlos en base a su velocidad de convergencia teórica. Es esta cuestión a la que está dedicado el siguiente apartado.

Antes de pasar a hablar de velocidades de convergencia es importante destacar otro criterio práctico para valorar la calidad de un algoritmo, que consiste en estudiar su eficacia para minimizar funciones cuadráticas. La razón es que, aunque la aproximación lineal de una función cerca de un mínimo suele ser bastante pobre (pues todas las componentes del gradiente se están haciendo cero), la aproximación cuadrática suele ser bastante adecuada. Por tanto, un algoritmo que no tiene un buen comportamiento para optimizar funciones cuadráticas es probable que

[1] Un enfoque muy habitual en la práctica, dada la sencillez de implementación, es el llamado *multiarranque*, que consiste en lanzar el mismo algoritmo partiendo de iterantes iniciales distintos y quedándose finalmente con la mejor de las soluciones obtenidas en las distintas pasadas.

no tenga un buen comportamiento para funciones no lineales en general, al menos cuando nos acerquemos al óptimo (pues cerca del óptimo la aproximación cuadrática será razonablemente buena).

4.2.2 Velocidad de convergencia

A continuación desarrollamos brevemente el concepto de *velocidad de convergencia asintótica*, aunque otros criterios basados en *velocidades de convergencia medias* también son posibles (aunque no tan estudiados).

Sea $\{r^t\}_{t\in\mathbb{N}} \subset \mathbb{R}$ una sucesión con límite $\bar{r}$ y tal que, para todo $t \in \mathbb{N}$, $r^t \neq \bar{r}$. El *orden de convergencia* es el supremo de los números no negativos p verificando que existe $\rho \in \mathbb{R}$ tal que

$$\limsup_{t\to\infty} \frac{|r^{t+1} - \bar{r}|}{|r^t - \bar{r}|^p} = \rho.$$[2]

A efectos de lo que vamos a comentar aquí se puede pensar simplemente en términos de límite en vez de límite superior, si eso nos resulta más sencillo.[3] En particular, los órdenes de convergencia más habituales en la práctica son los siguientes:

Convergencia lineal. Se da cuando $p = 1$ y $\rho \in (0, 1)$. En este caso tenemos que, asintóticamente, $|r^{t+1} - \bar{r}| = \rho|r^t - \bar{r}|$ y uno incluso podría verse tentado a decir que

$$|r^{t+1} - \bar{r}| \approx \rho \overset{t}{\cdots} \rho\, |r^1 - \bar{r}|.$$

Es por esto que a veces se usa el nombre de *convergencia geométrica* para referirse a la convergencia lineal, aunque este nombre debería reservarse para situaciones en las que la sucesión sea realmente una sucesión geométrica (y no sólo de modo asintótico).

Equivalentemente, la sucesión tiene convergencia lineal si existen $\rho \in (0, 1)$ y $k \in \mathbb{N}$ tales que, para todo $t > k$, $|r^{t+1} - \bar{r}| \leq \rho|r^t - \bar{r}|$.

Convergencia superlineal. Se da cuando $p > 1$ o $p = 1$ y $\rho = 0$. Equivalentemente, la sucesión tiene convergencia superlineal si existe una sucesión $\{\rho^t\}_{t\in\mathbb{N}}$ con $\lim_{t\to\infty} \rho^t = 0$ tal que, para todo $t \in \mathbb{N}$, $|r^{t+1} - \bar{r}| \leq \rho^t|r^t - \bar{r}|$.

Convergencia cuadrática. Es un caso particular de la convergencia superlineal en el que $p = 2$.

En la práctica, los algoritmos con convergencia lineal pueden llegar a ser bastante lentos (pues ρ puede estar muy cerca de 1) y aquellos con convergencia cuadrática casi siempre son bastante rápidos.

[2] En general el límite superior se define como $\limsup_{t\to\infty} \boldsymbol{x}^t = \lim_{t\to\infty}(\sup_{k>t} \boldsymbol{x}^k)$. En el caso de sucesiones de números reales, el límite superior es el menor número real b tal que, para todo $\varepsilon > 0$, $b + \varepsilon$ es mayor que casi todos los términos de la sucesión (donde casi todos quiere decir "todos salvo una cantidad finita").

[3] Formalmente se debe trabajar con límites superiores, ya que no toda sucesión tiene límite pero toda sucesión tiene límite superior. Por supuesto, en caso de que una sucesión sea convergente su límite es también su límite superior.

Las definiciones de velocidad de convergencia que acabamos de presentar se pueden generalizar inmediatamente para el caso en el que tenemos sucesiones de vectores $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}} \subset \mathbb{R}^n$ con límite $\bar{\boldsymbol{x}}$ y tales que, para todo $t \in \mathbb{N}$, $\boldsymbol{x}^t \neq \bar{\boldsymbol{x}}$. Para ello bastaría reemplazar las distancias de la forma $|r^t - \bar{r}|$ por la distancia euclídea $\|\boldsymbol{x}^t - \bar{\boldsymbol{x}}\|$.[4] Esto permite hablar de velocidad de convergencia a una solución $\bar{\boldsymbol{x}}$. El caso unidimensional es útil incluso en problemas definidos en $\mathbb{R}^n$, con $n > 1$, para hablar de velocidad de convergencia a un determinado valor de la función objetivo.

Para terminar, es importante destacar que los conceptos de velocidad y orden de convergencia que acabamos de comentar simplemente nos dan información acerca del número de iteraciones que necesitará el método. Otra consideración muy importante a la hora de establecer comparaciones entre algoritmos es la del tiempo requerido para realizar cada iteración.

4.2.3 El criterio de parada

Una vez que tenemos un algoritmo cuya convergencia ha sido demostrada teóricamente, a la hora de implementarlo es importante definir algún *criterio* de parada, que especifique cuando termina el algoritmo. Por supuesto, lo natural sería terminar el algoritmo en cuanto encontrásemos un punto del conjunto deseable Ω. Sin embargo, en la mayoría de los casos la convergencia a puntos de Ω sólo se alcanza en el límite y es necesario definir reglas prácticas de terminación de los algoritmos. Idealmente nos gustaría poner un criterio según el cual se terminase el algoritmo cuando estuviésemos lo suficientemente cerca del punto límite pero, desafortunadamente, no conocemos ese punto límite, pues es precisamente lo que pretendemos encontrar con el algoritmo. Es por esto que la mayoría de los criterios de parada se "apoyan" en que toda sucesión convergente es de Cauchy y, por tanto, sabemos que para todo $\varepsilon > 0$ habrá un término de la sucesión tal que la distancia entre dos términos cualesquiera posteriores es menor que $\varepsilon > 0$.

A continuación presentamos una serie de criterios de parada habituales (y que pueden ser usados conjuntamente), donde $\varepsilon > 0$ y $k \in \mathbb{N}$ son dos valores prefijados:

- $\|\boldsymbol{x}^{t+k} - \boldsymbol{x}^t\| \leq \varepsilon$. El algoritmo termina si en k iteraciones nos hemos movido menos de distancia ε en la sucesión de puntos generada.
- $\frac{\|\boldsymbol{x}^{t+k} - \boldsymbol{x}^t\|}{\|\boldsymbol{x}^t\|} \leq \varepsilon$. Similar al criterio anterior, pero usando la distancia relativa en vez de la distancia absoluta. A la hora de implementar criterios de parada relativos hay que tener cuidado pues si $\boldsymbol{x}^t \to \mathbf{0}$, el denominador tiende a cero y el test de parada podría no cumplirse nunca. Es por eso que lo normal es hacer algo del tipo $\frac{\|\boldsymbol{x}^{t+k} - \boldsymbol{x}^t\|}{\|\boldsymbol{x}^t\| + \delta} \leq \varepsilon$, donde $\delta > 0$ es un valor pequeño, pero que asegure que el denominador nunca se acerca demasiado a cero (tomar $\delta = \varepsilon$ es una opción habitual).
- $|f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+k})| < \varepsilon$. En este caso se mira cuánto ha mejorado la función objetivo en las últimas k iteraciones.
- $\frac{|f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+k})|}{|f(\boldsymbol{x}^t)|} < \varepsilon$. Similar al criterio anterior, pero usando la mejora relativa. Nuevamente, en la práctica suele optarse por algo de la forma $\frac{|f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+k})|}{|f(\boldsymbol{x}^t)| + \delta} < \varepsilon$, con $\delta > 0$ un valor pequeño.

[4] Aunque en la práctica la distancia euclídea es la más habitual, a veces es necesario trabajar con otros conceptos de distancia. Por ejemplo cuando tenemos sucesiones en espacios de dimensión infinita.

- $\|\nabla f(\boldsymbol{x}^t)\| < \varepsilon$. Este criterio podría ser de utilidad especialmente en problemas de optimización sin restricciones, donde sabemos que $\nabla f(\boldsymbol{x}) = \mathbf{0}$ es condición necesaria para tener optimalidad local.

4.3 Condiciones de optimalidad en problemas sin restricciones

Del mismo modo que en la Sección 2.1 estudiamos las condiciones de optimalidad en problemas de optimización convexa, a continuación presentamos los resultados correspondientes a funciones generales en el caso de problemas sin restricciones.

El siguiente resultado nos dice que la condición que, bajo diferenciabilidad, aseguraba optimalidad global en el caso convexo, ahora es únicamente una condición necesaria de optimalidad local.

Proposición 4.1. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función diferenciable en $\bar{\boldsymbol{x}}$. Si $\bar{\boldsymbol{x}}$ es un mínimo local, entonces $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$.*

Demostración. Supongamos que $\bar{\boldsymbol{x}}$ es un mínimo local y que $\nabla f(\bar{\boldsymbol{x}}) \neq \mathbf{0}$. Entonces, tomando $\boldsymbol{d} = -\nabla f(\bar{\boldsymbol{x}})$ tenemos que $\nabla f(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} = -\|\nabla f(\bar{\boldsymbol{x}})\|^2 < 0$ y la Proposición 2.8 nos asegura que $\boldsymbol{d}$ es una dirección de descenso, lo que contradice que $\bar{\boldsymbol{x}}$ sea un mínimo local. □

Nuevamente, la función $f(\boldsymbol{x}) = \boldsymbol{x}^3$ y el punto $\bar{\boldsymbol{x}} = 0$ sirven como ejemplo de que la condición del resultado anterior no es una condición suficiente. La condición $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$ usa información relativa a las derivadas parciales de primer orden de f, y por eso se conoce como *condición de primer orden*. Podemos conseguir una condición necesaria algo más fina introduciendo información relativa a las derivadas segundas.

Proposición 4.2. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función dos veces diferenciable en $\bar{\boldsymbol{x}}$. Si $\bar{\boldsymbol{x}}$ es un mínimo local, entonces $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$ y $\boldsymbol{H}(\bar{\boldsymbol{x}})$ es semidefinida positiva.*

Demostración. Supongamos que $\bar{\boldsymbol{x}}$ es un mínimo local. Consideremos una dirección $\boldsymbol{d} \in \mathbb{R}^n$. Por las condiciones sobre la diferenciabilidad de f tenemos que, para todo $\lambda > 0$.

$$f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) = f(\bar{\boldsymbol{x}}) + \lambda \nabla f(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} + \frac{1}{2}\lambda^2 \boldsymbol{d}^\intercal \boldsymbol{H}(\bar{\boldsymbol{x}})\boldsymbol{d} + \lambda^2 \varphi(\lambda \boldsymbol{d})\|\boldsymbol{d}\|^2,$$

con $\lim_{\lambda \to 0} \varphi(\lambda \boldsymbol{d}) = 0$. Por la Proposición 4.1, $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$, con lo que la ecuación anterior se puede expresar como

$$\frac{f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda^2} = \frac{1}{2}\boldsymbol{d}^\intercal \boldsymbol{H}(\bar{\boldsymbol{x}})\boldsymbol{d} + \varphi(\lambda \boldsymbol{d})\|\boldsymbol{d}\|^2.$$

Como $\bar{\boldsymbol{x}}$ es un mínimo local, $f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}}) \geq 0$ si λ es suficientemente pequeño, en cuyo caso también $\frac{1}{2}\boldsymbol{d}^\intercal \boldsymbol{H}(\bar{\boldsymbol{x}})\boldsymbol{d} + \varphi(\lambda \boldsymbol{d})\|\boldsymbol{d}\|^2 \geq 0$. Por tanto, si hacemos tender λ a cero tendremos $\boldsymbol{d}^\intercal \boldsymbol{H}(\bar{\boldsymbol{x}})\boldsymbol{d} \geq 0$. Como $\boldsymbol{d}$ fue elegido de modo arbitrario, $\boldsymbol{H}(\bar{\boldsymbol{x}})$ es semidefinida positiva. □

Para terminar nos apoyaremos en la matriz hessiana para presentar una condición suficiente de optimalidad local.

Proposición 4.3. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función dos veces diferenciable en $\bar{\boldsymbol{x}}$. Si $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$ y $\boldsymbol{H}(\bar{\boldsymbol{x}})$ es definida positiva, entonces $\bar{\boldsymbol{x}}$ es un mínimo local estricto.*

Demostración. Como f es dos veces diferenciable tenemos que, para todo $\bar{\boldsymbol{x}} \in \mathbb{R}^n$,

$$f(\boldsymbol{x}) = f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) + \frac{1}{2}(\boldsymbol{x} - \bar{\boldsymbol{x}})^\intercal \boldsymbol{H}(\bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}}) + \varphi(\boldsymbol{x} - \bar{\boldsymbol{x}})\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2,$$

con $\lim_{\boldsymbol{x}\to\bar{\boldsymbol{x}}} \varphi(\boldsymbol{x}-\bar{\boldsymbol{x}}) = 0$. Supongamos que $\bar{\boldsymbol{x}}$ no es un mínimo local estricto. Entonces existe una sucesión de puntos $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ cuyo límite es $\bar{\boldsymbol{x}}$ y tal que, para todo $t \in \mathbb{N}$, $\boldsymbol{x}^t \neq \bar{\boldsymbol{x}}$ y $f(\boldsymbol{x}^t) \leq f(\bar{\boldsymbol{x}})$. Teniendo en cuenta que $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$, la ecuación anterior implica que, para todo $t \in \mathbb{N}$,

$$0 \geq f(\boldsymbol{x}^t) - f(\bar{\boldsymbol{x}}) = \frac{1}{2}(\boldsymbol{x}^t - \bar{\boldsymbol{x}})^\intercal \boldsymbol{H}(\bar{\boldsymbol{x}})(\boldsymbol{x}^t - \bar{\boldsymbol{x}}) + \varphi(\boldsymbol{x}^t - \bar{\boldsymbol{x}})\|\boldsymbol{x}^t - \bar{\boldsymbol{x}}\|^2.$$

Si ahora dividimos por $\|\boldsymbol{x}^t - \bar{\boldsymbol{x}}\|^2$ y definimos $\boldsymbol{d}^t = (\boldsymbol{x}^t - \bar{\boldsymbol{x}})/\|\boldsymbol{x}^t - \bar{\boldsymbol{x}}\|^2$, tenemos que

$$0 \geq \frac{1}{2}(\boldsymbol{d}^t)^\intercal \boldsymbol{H}(\bar{\boldsymbol{x}})\boldsymbol{d}^t + \varphi(\boldsymbol{d}^t)\|\boldsymbol{x}^t - \bar{\boldsymbol{x}}\|^2.$$

Si tomamos ahora la sucesión $\{\boldsymbol{d}^t\}_{t\in\mathbb{N}}$, como todos sus elementos tienen norma 1, está definida en un compacto y tendrá una subsucesión convergente con límite $\boldsymbol{d}$. Además, como $\lim_{\boldsymbol{x}\to\bar{\boldsymbol{x}}} \varphi(\boldsymbol{x}-\bar{\boldsymbol{x}}) = 0$, en el límite tendremos $0 \geq \boldsymbol{d}^\intercal \boldsymbol{H}(\bar{\boldsymbol{x}})\boldsymbol{d}$, lo que contradice que $\boldsymbol{H}(\bar{\boldsymbol{x}})$ sea definida positiva y, por tanto, $\bar{\boldsymbol{x}}$ es un mínimo local estricto. □

Los algoritmos que presentaremos en el resto de esta sección tienen como objetivo buscar puntos que cumplan las condiciones necesarias de optimalidad local que acabamos de presentar y, a ser posible, también las condiciones suficientes. Por tanto, los resultados teóricos alrededor de estos algoritmos suelen centrarse en comprobar que los puntos límite de la sucesión generada por el algoritmo verifican que en ellos se anula el gradiente y, a ser posible, también alguna condición sobre la matriz hessiana. Como ya comentamos en el apartado anterior, en la práctica estos algoritmos suelen complementarse con alguna heurística que intente reducir las probabilidades de quedarse con un mal óptimo local.

4.4 Optimización unidimensional sin usar derivadas

La optimización unidimensional es la columna vertebral de muchos algoritmos de optimización para problemas de programación no lineal. Muchos de estos algoritmos proceden de la siguiente manera: dado un punto $\boldsymbol{x}^t$, se buscan una dirección $\boldsymbol{d}^t$ y un paso adecuado λ^t, en base a los cuales se define $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \lambda^t \boldsymbol{d}^t$. Esta idea ya la esbozamos en la Sección 2.3.

Dirección. La dirección $\boldsymbol{d}^t$ suele ser una *dirección de descenso* (ver Definición 2.1).

Paso. El paso λ^t puede intentar tomarse de forma óptima, minimizando $g(\lambda) = f(\boldsymbol{x}^t + \lambda \boldsymbol{d}^t)$, con lo que tenemos un problema unidimensional. Estos problemas se conocen como problemas de *búsqueda lineal* (line search).

En esta sección y en la siguiente discutiremos algunos métodos clásicos para resolver problemas de búsqueda lineal. Una primera idea podría ser buscar un punto en el que se anule la derivada de la función g. Sin embargo, esto en la práctica puede tener múltiples inconvenientes:

- La función g puede resultar bastante compleja, ya que está definida implícitamente a partir de la función multidimensional f, lo que puede dificultar trabajar con ella analíticamente.
- Si f no es diferenciable, g tampoco lo será.
- Si f es diferenciable, tenemos $g'(\lambda) = \boldsymbol{d}^{\intercal} \nabla f(\boldsymbol{x} + \lambda \boldsymbol{d})$ y buscar puntos para los que $g'(\lambda) = 0$ equivale a resolver la ecuación $\boldsymbol{d}^{\intercal} \nabla f(\boldsymbol{x} + \lambda \boldsymbol{d}) = 0$, que normalmente será no lineal.
- Incluso si somos capaces de encontrar puntos tales que $g'(\lambda) = 0$, salvo que tengamos asegurada la convexidad de g, podremos estar ante mínimos locales, máximos locales e incluso puntos de silla.

Por los motivos que acabamos de exponer, lo habitual es utilizar técnicas numéricas para minimizar la función g, aprovechándose del carácter unidimensional de la misma.

En general, asumiremos que la minimización se realiza dentro de un intervalo cerrado. Más concretamente, trabajaremos con el problema de minimizar $g(\lambda)$ con $\lambda \in [a, b]$. Todo lo que sabemos es que λ no está fuera de este intervalo y su ubicación dentro del mismo es desconocida, con lo que este intervalo se conoce como *intervalo de incertidumbre*. El objetivo de los métodos de búsqueda consiste en ir reduciendo sucesivamente este intervalo hasta que tengamos una precisión que consideremos suficiente.

A continuación presentamos un sencillo resultado, que sirve de motivación para algunos de los métodos que definiremos posteriormente.[5]

Proposición 4.4. *Sea $g : \mathbb{R} \to \mathbb{R}$ una función estrictamente convexa sobre el intervalo $[a, b]$. Sean λ y μ dos puntos en $[a, b]$ con $\lambda < \mu$.*

(i) *Si $g(\lambda) > g(\mu)$, entonces, para todo $z \in [a, \lambda)$, $g(z) > g(\mu)$. Es decir, el mínimo global se encuentra a la derecha de λ.*

(ii) *Si $g(\lambda) < g(\mu)$, entonces, para todo $z \in (\mu, b]$, $g(z) > g(\lambda)$. Es decir, el mínimo global se encuentra a la izquierda de μ.*

Demostración. Demostraremos únicamente el primer caso, siendo similar la demostración del segundo. Supongamos que $g(\lambda) > g(\mu)$ y tomemos $z \in [a, \lambda)$. Claramente, existe $\gamma \in (0, 1)$ tal que $\lambda = \gamma z + (1 - \gamma)\mu$. Por la convexidad de g,

$$g(\lambda) < \gamma g(z) + (1 - \gamma) g(\mu) \overset{g(\lambda) > g(\mu)}{<} \gamma g(z) + (1 - \gamma) g(\lambda).$$

Restando $g(\lambda)$ en los dos lados de la desigualdad obtenemos $0 < \gamma g(z) - \gamma g(\lambda)$, con lo que $g(z) > g(\lambda)$. Como $g(\lambda) > g(\mu)$, concluimos que $g(z) > g(\mu)$. □

[5] Para la Proposición 4.4 es suficiente con asumir que la función es estrictamente cuasiconvexa. Del mismo modo, para otros resultados que mencionaremos más adelante, también se puede trabajar con seudoconvexidad en vez de convexidad. Para facilitar la exposición, presentamos todos los resultados asumiendo la condición de convexidad, que engloba a ambas (añadiendo diferenciabilidad cuando sea preciso).

Si lo que queremos es minimizar la función g, la Proposición 4.4 nos dice cómo se reduce el intervalo de incertidumbre en cada caso. Esta reducción, que podemos ver en la Figura 4.1, es en la que se apoyan algunos de los algoritmos que presentaremos en esta sección, con lo que la convexidad es una propiedad necesaria para asegurar que éstos alcanzan un óptimo global. En el caso de que la función original no sea convexa, uno puede intentar dividir el intervalo de incertidumbre en subintervalos en los que sí sea convexa, encontrar el mínimo en cada uno de ellos y al final quedarse con el menor de los mínimos obtenidos. Alternativamente, uno puede aplicar directamente dichos algoritmos sobre el intervalo completo, conformándose con la posible convergencia de los mismos a óptimos locales en vez de óptimos globales.

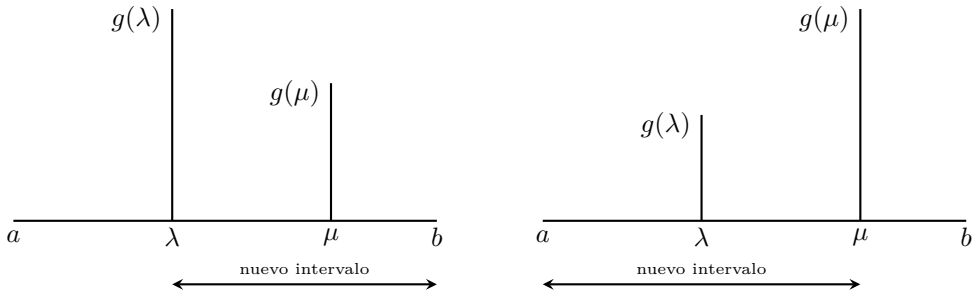


Figura 4.1: Ilustración de los dos casos en la Proposición 4.4.

4.4.1 Búsqueda uniforme

La búsqueda uniforme es el enfoque más simple, y que consiste en elegir una malla del intervalo de incertidumbre $[a^1, b^1]$ y evaluar g en todos los puntos de dentro de la malla, que serán de la forma $a^1 + k\delta$ con $k \in \{1, \ldots, n\}$ y tales que $b^1 = a^1 + (n+1)\delta$. Una vez hecho esto, si $\hat{\lambda}$ es un punto de la malla donde la función ha tomado un valor mínimo, la convexidad de f implica que el óptimo está en el subintervalo $[\hat{\lambda} - \delta, \hat{\lambda} + \delta]$ (ver Figura 4.2).

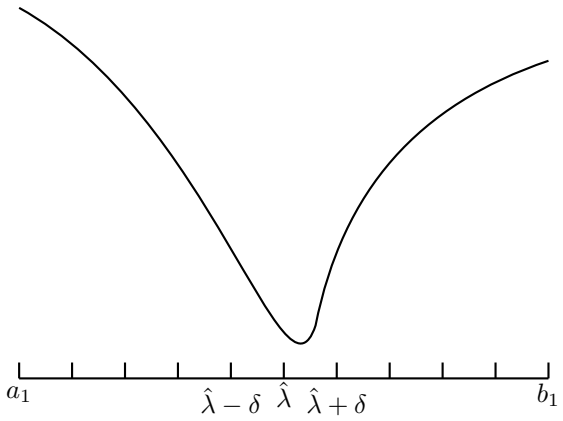


Figura 4.2: Ilustración del método de búsqueda uniforme.

MÉTODO DE BÚSQUEDA UNIFORME

INICIALIZACIÓN

Elegir $l > 0$, la longitud deseada en el intervalo final. Definir $n \in \mathbb{N}$ como el número natural más pequeño tal que $\frac{2(b^1-a^1)}{n+1} < l$. Definir $\delta = \frac{b^1-a^1}{n+1}$.

PASO 1

Calcular $\hat{k} = \text{argmin}_{k\in\{1,\dots,n\}} g(a^1 + k\delta)$. Definir $\hat{\lambda} = a^1 + \hat{k}\delta$. FIN. Devolvemos el intervalo $[\hat{\lambda} - \delta, \hat{\lambda} + \delta]$.

Figura 4.3: Esquema del método de búsqueda uniforme.

Nótese que, bajo convexidad, el algoritmo, representado en la Figura 4.3, no necesita evaluar g en los extremos del intervalo $[a^1, b^1]$. Por tanto, después de n evaluaciones de la función objetivo, la longitud del nuevo intervalo de incertidumbre pasa a ser 2δ. Un método habitual para implementar búsqueda uniforme en la práctica, en vez de fijar directamente un tamaño n que nos asegure la precisión deseada, consiste en empezar con un valor relativamente pequeño de n (poca precisión) e ir aumentando la precisión a medida que vamos quedándonos son subintervalos más pequeños. En cualquier caso, los siguientes métodos serán más eficientes pues, en vez de generar por adelantado todos los puntos en los que se va a evaluar la función, van utilizando información de las iteraciones precedentes a la hora de generar los nuevos puntos.

4.4.2 Método de dicotomía

La Proposición 4.4 nos dice cómo se va a reducir el intervalo de incertidumbre inicial $[a^1, b^1]$ tras dos evaluaciones de la función objetivo en puntos λ^1 y μ^1. En un caso la longitud del nuevo intervalo será $\mu^1 - a^1$ y en otro $b^1 - \lambda^1$. Como a priori no sabemos en cuál de estos dos casos vamos a estar, la idea del método de dicotomía consiste en minimizar el valor máximo que pueden tomar estas dos longitudes. Claramente, la forma de conseguir esto es tomar $\lambda^1 = \mu^1 = \frac{a^1+b^1}{2}$, reduciendo a la mitad la longitud del intervalo con cada iteración. Sin embargo, como dos evaluaciones son necesarias para aplicar la Proposición 4.4, el método de dicotomía coloca λ^1 y μ^1 simétricamente a distancia $\varepsilon > 0$ del punto medio, $\frac{a^1+b^1}{2}$ (ver Figura 4.4).

El método de dicotomía, representado en la Figura 4.5, asegura que en cada iteración t se obtiene un nuevo intervalo de incertidumbre cuya longitud es $\varepsilon + \frac{b^t-a^t}{2}$. Apoyándose en esto se puede probar fácilmente que la longitud del intervalo $[a^{t+1}, b^{t+1}]$ es exactamente $\frac{1}{2^t}(b^1 - a^1) + 2\varepsilon(1 - \frac{1}{2^t})$. Apoyándonos en esta forma podemos calcular el número de iteraciones necesarias para conseguir una determinada longitud final l.

4.4.3 Método de la sección áurea

En cada iteración del método de dicotomía se realizan dos evaluaciones de la función objetivo y se reduce la longitud del intervalo de incertidumbre prácticamente a la mitad. A continuación presentamos el método de la sección áurea (Wilde, 1964), cuya idea es hacer únicamente una evaluación por iteración, sacrificando un poco la reducción de la longitud del intervalo. La idea consiste en elegir λ^t y μ^t asegurando lo siguiente:

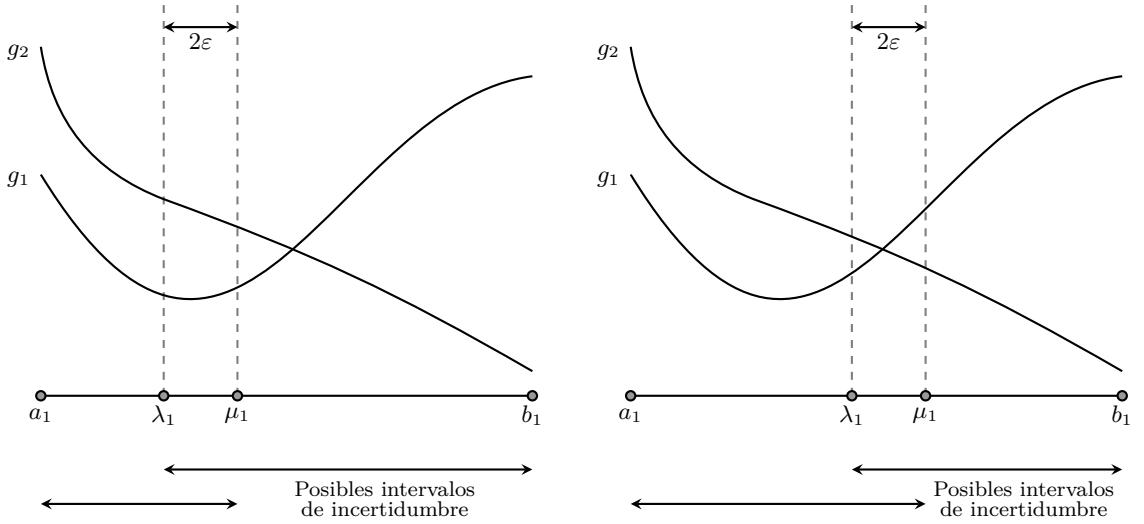


Figura 4.4: Ilustración del método de dicotomía.

MÉTODO DE DICOTOMÍA

INICIALIZACIÓN

Elegir $\varepsilon > 0$. Elegir $l > 0$, la longitud deseada en el intervalo final. Denotemos al intervalo inicial por $[a^1, b^1]$ y definamos $t = 1$.

PASO 1

- Si $[b^t - a^t] < l$, FIN. Devolvemos el intervalo $[a^t, b^t]$.
- En otro caso, definir $\lambda^t = \frac{a^t+b^t}{2} - \varepsilon$ y $\mu^t = \frac{a^t+b^t}{2} + \varepsilon$. Ir al PASO 2.

PASO 2

- Si $g(\lambda^t) < g(\mu^t)$, definir $a^{t+1} = a^t$ y $b^{t+1} = \mu^t$.
- En otro caso, definir $a^{t+1} = \lambda^t$ y $b^{t+1} = b^t$.

Reemplazar t por $t + 1$ e ir al PASO 1.

Figura 4.5: Esquema del método de dicotomía.

- La longitud del nuevo intervalo de incertidumbre es independiente del resultado de la iteración actual (aunque no tiene por qué ser dicha reducción prácticamente igual a 0.5 como en el método de dicotomía). Por tanto, $b^t - \lambda^t = \mu^t - a^t$. Si escribimos λ^t como combinación convexa de a^t y b^t tenemos $\lambda^t = \alpha a^t + (1 - \alpha)b^t$, con $\alpha \in (0, 1)$. Entonces,

$$\mu^t - a^t = b^t - \lambda^t = b^t - (\alpha a^t + (1 - \alpha)b^t) = \alpha(b^t - a^t).$$

  Con lo que $\lambda^t = a^t + (1 - \alpha)(b^t - a^t)$ y $\mu^t = a^t + \alpha(b^t - a^t)$ y

$$b^{t+1} - a^{t+1} = \alpha(b^t - a^t).$$

- En la iteración $t + 1$, o bien $\lambda^{t+1} = \mu^t$ o $\mu^{t+1} = \lambda^t$. Esta es la clave del método de la sección áurea, pues asegura que en cada nueva iteración sólo será necesaria una nueva

evaluación de la función objetivo. Para esto hay que distinguir dos casos, que ilustramos en la Figura 4.6.

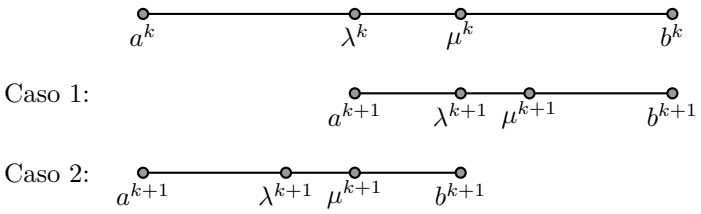


Figura 4.6: Ilustración del método de la sección áurea.

- Si $g(\lambda^t) > g(\mu^t)$, entonces $a^{t+1} = \lambda^t$ y $b^{t+1} = b^t$. En este caso, para que $\lambda^{t+1} = \mu^t$ aplicando la fórmula para el cálculo de λ^t obtenida anteriormente (y reemplazando t por $t+1$) tendremos

$$\mu^t = \lambda^{t+1} = a^{t+1} + (1-\alpha)(b^{t+1} - a^{t+1}) = \lambda^t + (1-\alpha)(b^t - \lambda^t).$$

  Si en esta ecuación aplicamos las sustituciones $\lambda^t = a^t + (1-\alpha)(b^t - a^t)$ y $\mu^t = a^t + \alpha(b^t - a^t)$ obtenemos

$$a^t + \alpha(b^t - a^t) = a^t + (1-\alpha)(b^t - a^t) + (1-\alpha)(b^t - (a^t + (1-\alpha)(b^t - a^t))).$$

  Simplificando esta expresión llegamos a la ecuación $\alpha^2 + \alpha - 1 = 0$.
- Si $g(\lambda^t) \leq g(\mu^t)$, un análisis análogo nos lleva de nuevo a la ecuación $\alpha^2 + \alpha - 1 = 0$.

Por tanto, como las raíces de la ecuación $\alpha^2 + \alpha - 1 = 0$ son $\frac{-1+\sqrt{5}}{2} \approx 0.618$ y $\frac{-1-\sqrt{5}}{2} \approx -1.618$ y $\alpha \in (0,1)$, nos quedaremos con la primera solución con lo que $\alpha = 0.618$. Nótese que el número de oro es $\frac{1+\sqrt{5}}{2} = 1.618 = \alpha + 1 = \frac{1}{\alpha}$ y de ahí que este método se conozca con el nombre de método de la sección áurea. La descripción algorítimica del método está representada en la Figura 4.7.

Exceptuando la primera iteración, en la que dos evaluaciones funcionales son necesarias, el método de la sección áurea realiza una única evaluación de g en cada iteración y consigue reducir la longitud del intervalo de incertidumbre a 0.618 veces su longitud original. En este sentido, tras n evaluaciones funcionales tendríamos que la longitud con el método de dicotomía sería $l = (\frac{1}{2})^{\frac{n}{2}}(b^1 - a^1)$ y con el método de la sección áurea $l = (0.618)^{n-1}(b^1 - a^1)$.

Para terminar, es interesante mencionar otro método de búsqueda lineal sin derivadas, que se conoce como método de búsqueda de Fibonacci, atribuido a Kiefer (1953). El procedimiento es prácticamente análogo al del método de la sección áurea, pero α varía iteración a iteración y sus valores son fijados a priori dependiendo del número total de iteraciones deseadas. En caso de que este dicho número sea n, tendremos $\alpha^t = \frac{F^{n-t}}{F^{n-t+1}}$, donde F^k denota el k-ésimo término de la sucesión de Fibonacci.[6] En el método de Fibonacci ha de especificarse de antemano el número de iteraciones y la reducción del intervalo de incertidumbre cambia iteración a iteración, pero

[6] La sucesión de Fibonacci se define como $F^0 = F^1 = 1$ y, para todo $k \geq 1$, $F^{k+1} = F^k + F^{k-1}$.

MÉTODO DE LA SECCIÓN ÁUREA

INICIALIZACIÓN

Elegir $l > 0$, la longitud deseada en el intervalo final. Denotemos al intervalo inicial por $[a^1, b^1]$.

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

Definamos $\lambda^1 = a^1 + (1-\alpha)(b^1 - a^1)$, $\mu^1 = a^1 + \alpha(b^1 - a^1)$ y $t = 1$.

PASO 1

- Si $[b^t - a^t] < l$, FIN. Devolvemos el intervalo $[a^t, b^t]$.
- En otro caso, ir al PASO 2.

PASO 2

- Si $g(\lambda^t) > g(\mu^t)$, definir $a^{t+1} = \lambda^t$ y $b^{t+1} = b^t$. Además, definir $\lambda^{t+1} = \mu^t$ y $\mu^{t+1} = a^{t+1} + \alpha(b^{t+1} - a^{t+1})$. Evaluar $g(\mu^{t+1})$.
- Si $g(\lambda^t) \leq g(\mu^t)$, definir $a^{t+1} = a^t$ y $b^{t+1} = \mu^t$. Además, definir $\mu^{t+1} = \lambda^t$ y $\lambda^{t+1} = a^{t+1} + (1-\alpha)(b^{t+1} - a^{t+1})$. Evaluar $g(\lambda^{t+1})$.

Reemplazar t por $t+1$ e ir al PASO 1.

Figura 4.7: Esquema del método de la sección áurea.

se mantiene la propiedad de que sólo una evaluación de la función objetivo es necesaria en cada iteración. Además, se sabe que $\lim_{k\to\infty} \frac{F^k}{F^{k+1}} = \alpha = \frac{-1+\sqrt{5}}{2}$ con lo que para valores grandes de n el comportamiento de este método es muy similar al del método de la sección áurea.

4.4.4 Comparación de los métodos de búsqueda lineal sin derivadas

La Tabla 4.1 presenta, para cada uno de los métodos, la expresión para la longitud del intervalo de incertidumbre resultante como función del número de evaluaciones de la función objetivo. Estos valores son bastante sencillos de obtener a partir de las descripciones de los métodos. Además, presentamos también el valor para el método de bisección, que veremos en la Sección 4.5.1.

Longitud Intervalo final	Valor teórico	$b^1 - a^1 = 1$ $n = 10$	$b^1 - a^1 = 1$ $n = 20$
Método			
Búsqueda uniforme	$l = \frac{2(b^1-a^1)}{n+1}$	0.1818	0.095238
Dicotomía	$l = 0.5^{\frac{n}{2}}(b^1 - a^1)$	0.0313	0.000977
Sección áurea	$l = 0.618^{n-1}(b^1 - a^1)$	0.0132	0.000107
Fibonacci	$l = \frac{b^1-a^1}{F^n}$	0.0112	0.000091
Bisección	$l = 0.5^n(b^1 - a^1)$	0.0009	0.000001

Tabla 4.1: Comparativa

La Tabla 4.1 deja muy clara la ineficiencia del método de búsqueda uniforme y también la mejora que los métodos de la sección áurea y Fibonacci representan con respecto al método de

dicotomía.

Es importante añadir que, desde el punto de vista teórico, se puede demostrar que el método de Fibonacci es el método más eficiente de entre todos los que no usan derivadas en lo que respecta al número de evaluaciones funcionales que son necesarias para conseguir una determinada reducción del intervalo de incertidumbre.

En la práctica, de entre los métodos descritos en esta sección, el método de la sección áurea suele ser el elegido.

4.5 Optimización unidimensional usando derivadas

4.5.1 Método de bisección

Nuevamente, suponemos que queremos minimizar una función real g sobre un intervalo cerrado y acotado. Supongamos además que g es convexa y diferenciable.[7]

La idea del método de bisección es reemplazar los dos valores de la función g que la Proposición 4.4 requiere para poder reducir el intervalo de incertidumbre por una única observación de su derivada. Supongamos que en la iteración t evaluamos g' en un punto λ^t del intervalo de incertidumbre $[a^t, b^t]$. Entonces, pueden suceder tres cosas:

$g'(\lambda^t) = 0$. En este caso, la convexidad de g asegura que λ^t es el mínimo buscado.

$g'(\lambda^t) > 0$. Por la convexidad de g, para todo $\lambda > \lambda^t$, $g(\lambda) \geq g(\lambda^t) + g'(\lambda^t)(\lambda - \lambda^t) > g(\lambda^t)$. Por tanto, el mínimo de g estará a la izquierda de λ^t.

$g'(\lambda^t) < 0$. Por la convexidad de g, para todo $\lambda < \lambda^t$, $g(\lambda) \geq g(\lambda^t) + g'(\lambda^t)(\lambda - \lambda^t) > g(\lambda^t)$. Por tanto, el mínimo de g estará a la derecha de λ^t.

Razonando igual que cuando definimos el método de dicotomía, si queremos minimizar el valor máximo que puede tomar la longitud del nuevo intervalo de incertidumbre, lo mejor es que λ^t sea el punto medio del intervalo de incertidumbre actual. Esto es exactamente lo que hace el método de bisección, representado en la Figura 4.8,. De hecho, este método es análogo al método de dicotomía, donde en cada iteración se evalúa la derivada en vez de los dos puntos centrales.[8]

4.5.2 Método de Newton

El método de Newton se basa en ideas desarrolladas por Isaac Newton en la segunda mitad del siglo XVII para encontrar raíces de funciones, es decir, resolver ecuaciones de la forma $f(\boldsymbol{x}) = 0$, para una referencia más reciente se puede consultar Fletcher (1987). En el caso de la optimización lo que busca es resolver $\nabla f(\boldsymbol{x}) = 0$, o $g'(\lambda) = 0$ en el caso unidimensional que nos atañe por el momento. Requiere que la función g sea dos veces diferenciable y se

[7] Para el correcto funcionamiento del método de bisección bastaría con pedir que g sea seudoconvexa.

[8] En este sentido, el método de dicotomía puede interpretarse como una evaluación de la derivada por diferencias finitas.

Método de bisección

Inicialización

Elegir $l > 0$, la longitud deseada en el intervalo final. Denotemos al intervalo inicial por $[a^1, b^1]$ y definamos $t = 1$.

Paso 1

- Si $[b^t - a^t] < l$, Fin. Devolvemos el intervalo $[a^t, b^t]$.
- En otro caso, definir $\lambda^t = \frac{a^t + b^t}{2}$. Ir al Paso 2.

Paso 2

- Si $g'(\lambda^t) = 0$, Fin. Devolvemos λ^t, que es el mínimo buscado.
- Si $g'(\lambda^t) > 0$, definir $a^{t+1} = a^t$ y $b^{t+1} = \lambda^t$.
- Si $g'(\lambda^t) < 0$, definir $a^{t+1} = \lambda^t$ y $b^{t+1} = b^t$.

Reemplazar t por $t + 1$ e ir al Paso 1.

Figura 4.8: Esquema del método de bisección.

basa en trabajar con la aproximación cuadrática de dicha función en el punto actual λ^t. Más concretamente, esta aproximación cuadrática viene dada por:

$$q(\lambda) = g(\lambda^t) + g'(\lambda^t)(\lambda - \lambda^t) + \frac{1}{2} g''(\lambda^t)(\lambda - \lambda^t)^2.$$

Este método se basa en tomar como siguiente iterante, λ^{t+1}, el punto donde $q'(\lambda) = 0$ y que, en el caso de que $q(\lambda)$ sea una función convexa, coincidirá con su mínimo. Dado que $q'(\lambda) = g'(\lambda^t) + g''(\lambda^t)(\lambda - \lambda^t)$, el método de Newton definirá λ^{t+1} como

$$\lambda^{t+1} = \lambda^t - \frac{g'(\lambda^t)}{g''(\lambda^t)}.$$

El esquema del método de Newton está representado en la Figura 4.9. Es importante hacer

Método de Newton

Inicialización

Elegir $\varepsilon > 0$, precisión para el criterio de parada. Denotemos al intervalo inicial por $[a^1, b^1]$. Definamos $t = 1$ y $\lambda^1 = \frac{a^1 + b^1}{2}$.

Paso 1

- Si $|g'(\lambda^t)| < \varepsilon$, Fin. Devolvemos λ^t.
- En otro caso, definir $\lambda^{t+1} = \lambda^t - \frac{g'(\lambda^t)}{g''(\lambda^t)}$. Ir al Paso 2.

Paso 2

- Si $|\lambda^{t+1} - \lambda^t| < \varepsilon$, Fin. Devolvemos λ^t.
- En otro caso, reemplazar t por $t + 1$ e ir al Paso 1.

Figura 4.9: Esquema del método de Newton.

un par de limitaciones importantes de este método:

- El método de Newton sólo está definido si no pasa por ningún λ^t para el cual $g''(\lambda^t) = 0$.
- Además, nada nos asegura que el algoritmo avance en direcciones en las que mejora la función objetivo. Por ejemplo, si las aproximaciones cuadráticas no son convexas, entonces el algoritmo podría estar encontrando máximos locales de dichas aproximaciones, con lo que el comportamiento estaría lejos de ser el deseable.

A pesar de estos dos problemas, el método de Newton es uno de los métodos más populares en optimización no lineal. Los principales motivos son los siguientes:

- Se puede demostrar que, si el método parte de un punto suficientemente próximo a un punto en el que el gradiente de la función se anula, entonces el algoritmo converge a ese punto (que si no tenemos convexidad no tiene por qué ser un mínimo). Este resultado es cierto aunque la función a minimizar no sea convexa.
- El método se puede generalizar de modo bastante sencillo a dimensiones más altas (como veremos en la Sección 4.8).
- Los métodos de cuasi-Newton, que trabajan con aproximaciones de la derivada segunda, permiten asegurar que dicha aproximación nunca se anula sin comprometer el resto de propiedades del algoritmo.
- El método de Newton y sus variantes suelen tener buenas velocidades de convergencia.

4.6 Optimización unidimensional: métodos inexactos

Todos los métodos que hemos descrito hasta ahora son lo que se denominan *métodos exactos*, pues permiten tener un control exacto de la distancia máxima a la solución óptima. Sin embargo, se basan en supuestos de convexidad que muchas veces no se cumplen en la práctica. Además, a la hora de trabajar con implementaciones hechas en ordenador rara vez se puede hablar de métodos 100 % exactos.

Para terminar la exposición de técnicas de búsqueda de línea queremos mencionar brevemente un par de métodos inexactos de gran popularidad en la práctica:

Búsqueda lineal por ajuste cuadrático. Se trata de un método que no usa derivadas. En cada iteración realiza un ajuste cuadrático basado en las evaluaciones en tres puntos del intervalo de incertidumbre. Iteración a iteración se actualiza exactamente uno de los tres puntos y se realiza un nuevo ajuste.

Backtracking. Este método es una de las múltiples adaptaciones de la conocida como regla de Armijo-Goldstein (Armijo, 1966), que son muy empleadas en la práctica. Una de sus mayores virtudes es que requiere un número relativamente reducido de evaluaciones de la función objetivo. El esquema general es el siguiente. Se eligen parámetros β y γ con $0 < \beta < \gamma < 1$. Después, se elige λ como el primer γ^k en la sucesión de potencias de γ: $\gamma, \gamma^2, \gamma^3, \ldots$ tal que:

$$\frac{g(\gamma^k) - g(0)}{\gamma^k} < \beta g'(0).$$

Nótese que $g'(0) < 0$ por estar trabajando con una dirección de descenso. La idea es muy natural. No será necesario incrementar k indefinidamente pues

$$\lim_{k\to\infty} \frac{g(\gamma^k) - g(0)}{\gamma^k} = g'(0).$$

Si pensamos en $\beta = 0.9$, lo que hace este método es partir de γ cercano a 1 y quedarse con el paso más grande posible (de los obtenidos mediante potencias de γ) que obtiene un decrecimiento de la función g comparable al marcado por la derivada $g'(0)$. Más concretamente, en vez de exigir un decrecimiento relativo de al menos $g'(0)$ (que se estaría garantizado en el límite cuando $k \to \infty$), nos conformaremos con un decrecimiento relativo mejor que $0.9 \cdot g'(0)$.

4.7 Optimización multidimensional sin usar derivadas

Pasamos ahora a estudiar problemas de optimización multidimensional sin restricciones de la forma minimizar$_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$, empezando en esta sección por dos métodos que no requieren el uso de derivadas. Como ya comentamos, muchos de los métodos que veremos en el resto de las secciones de este tema se apoyan en la realización de búsquedas lineales, y la clave está en escoger de forma adecuada las direcciones en las que realizar dichas búsquedas.

4.7.1 Método de descenso por coordenadas

Se trata de un método clásico usado desde los orígenes de la optimización matemática y es difícil atribuirle una autoría concreta. Se basa en usar como direcciones de búsqueda, de modo alternativo, las direcciones de los ejes coordenados. El método de descenso por coordenadas está representado en la Figura 4.10, donde $\boldsymbol{e}^1, \ldots, \boldsymbol{e}^n$ denotan los vectores de la base canónica. En dicha figura hemos tomado un criterio de parada relativo de la forma $\frac{\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|}{\|\boldsymbol{x}^t\|} \leq \varepsilon$, pero se podrían haber elegido otros criterios en su lugar.

Método de descenso por coordenadas

Inicialización
Elegir $\varepsilon > 0$. Elegir un punto inicial $\boldsymbol{x}^1$. Definir $\boldsymbol{y}^1 = \boldsymbol{x}^1$ y $t = j = 1$.

Paso 1
Calcular λ^j, solución del problema minimizar$_{\lambda\in\mathbb{R}} f(\boldsymbol{y}^j + \lambda \boldsymbol{e}^j)$. Definir $\boldsymbol{y}^{j+1} = \boldsymbol{y}^j + \lambda^j \boldsymbol{e}^j$.

- Si $j < n$ reemplazar j por $j + 1$ y repetir el Paso 1.
- Si $j = n$, ir al Paso 2.

Paso 2
Definir $\boldsymbol{x}^{t+1} = \boldsymbol{y}^{n+1}$.

- Si $\frac{\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|}{\|\boldsymbol{x}^t\|} \leq \varepsilon$, Fin. Devolvemos $\boldsymbol{x}^{t+1}$.
- En otro caso, definir $\boldsymbol{y}^1 = \boldsymbol{x}^{t+1}$ y $j = 1$. Reemplazar t por $t + 1$ e ir al Paso 1.

Figura 4.10: Esquema del método de descenso por coordenadas.

Proposición 4.5. *Supongamos que se cumplen los siguientes supuestos:*

- *La función f es diferenciable.*
- *El mínimo de cualquier búsqueda lineal asociada a la función f es único.*
- *La sucesión de puntos generada por el algoritmo está contenida en un conjunto compacto.*

Entonces, todo punto de acumulación $\bar{\boldsymbol{x}}$ de la sucesión generada por el método de descenso por coordenadas cumple $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$.

Como ya comentamos al principio de este tema, vamos a centrarnos en las intuiciones detrás los algoritmos y en sus propiedades, tratando de abstraernos de las demostraciones en la medida de lo posible.

Nótese que, a pesar de ser un método que no usa derivadas, la convergencia del método de descenso por coordenadas necesita diferenciabilidad. En la Figura 4.11 podemos ver el tipo de comportamiento que puede aparecer en funciones no diferenciables, y que ocasiona que el algoritmo se atasque en su búsqueda al no producirse mejora en ninguna de las direcciones $\boldsymbol{e}^j$. Se trata de la función

$$f(\boldsymbol{x}) = \begin{cases} x_1^2 + (x_2 - 6)^2 & \text{si } x_2 \leq x_1 \\ (x_1 - 6)^2 + x_2^2 & \text{si } x_2 > x_1, \end{cases}$$

que para puntos debajo de la diagonal $D = \{\boldsymbol{x} : x_1 = x_2\}$ coincide con la distancia al punto $(0, 6)$ y en puntos encima de la diagonal con la distancia al punto $(6, 0)$. Esta función es diferenciable en $\mathbb{R}^2 \backslash D$ y, como vemos en la Figura 4.11, si empezamos el algoritmo en el punto $(2, 1)$ y nos movemos en la dirección $(1, 0)$, el paso óptimo será -1, llegando al punto $(1, 1) \in D$. Pero el algoritmo de descenso por coordenadas no consigue mejorar cuando llega a puntos de la diagonal (esto se deduce fácilmente de la forma de las curvas de nivel en la figura, donde movimientos en horizontal o vertical no permiten mejorar).

Otro problema habitual en el algoritmo de descenso por coordenadas es el conocido como zigzag, que se produce cuando las curvas de nivel de la función a minimizar definen "valles estrechos", como representamos en la Figura 4.12.

En este caso tenemos que el algoritmo, a la hora de minimizar la función $f(\boldsymbol{x}) = (3 - x_1)^2 + 7(x_2 - x_1^2)^2$, tiene un avance muy lento hacia el único óptimo global, el punto $(3, 9)$. En la figura vemos las primeras 50 iteraciones del algoritmo, que necesitaría más de 800 iteraciones para verificar el criterio de parada con $\varepsilon = 10^{-4}$.

En el siguiente apartado presentamos una modificación del método de descenso por coordenadas, que introduce un paso de aceleración para mitigar el impacto de este tipo de zigzag.

4.7.2 Método de Hooke y Jeeves

El método de Hooke y Jeeves (1961) consiste en introducir una pequeña mejora en el método de descenso por coordenadas, que tiene como principal objetivo acelerar la convergencia del mismo. Más concretamente, supongamos que realizamos un ciclo completo de búsqueda en las direcciones $\boldsymbol{e}^j$ en el método de descenso por coordenadas, pasando del punto $\boldsymbol{x}^t$ al punto $\boldsymbol{x}^{t+1}$. Dado que la dirección $\boldsymbol{x}^{t+1} - \boldsymbol{x}^t$ es la dirección en la que el algoritmo se ha movido dentro de este ciclo, parece natural hacer una nueva búsqueda en esta dirección antes de iniciar un nuevo ciclo

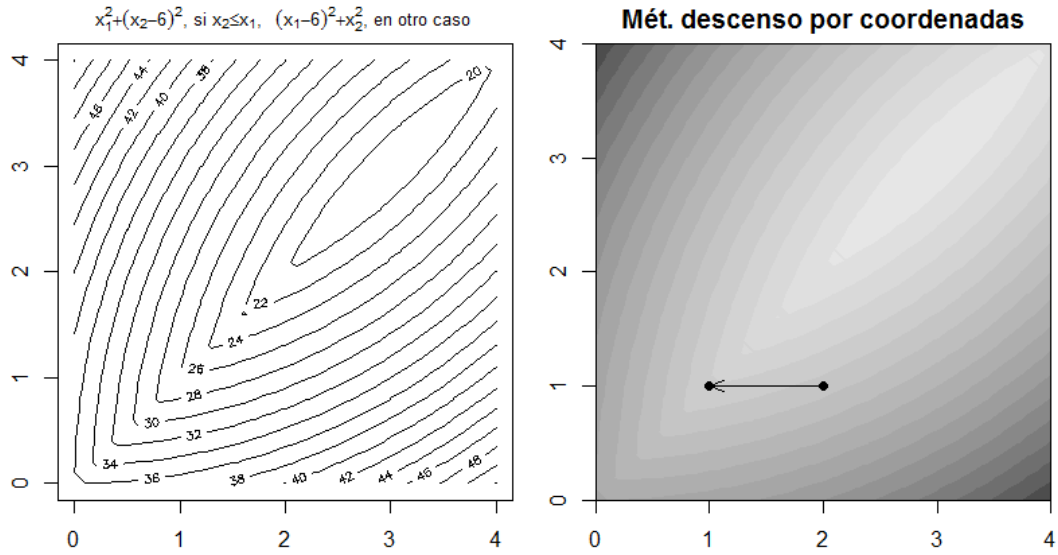


Figura 4.11: Ejemplo en el que el método de descenso por coordenadas no converge a un punto estacionario.

en las direcciones $\boldsymbol{e}^j$. Esto es exactamente lo que hace el algoritmo de Hooke y Jeeves. En este algoritmo, la búsqueda en la dirección $\boldsymbol{x}^{t+1} - \boldsymbol{x}^t$ se conoce como paso de aceleración o búsqueda por patrón, en la que se hace una búsqueda siguiendo el patrón obtenido con las búsquedas en las direcciones $\boldsymbol{e}^j$. El esquema de este algoritmo está representado en la Figura 4.13.

El siguiente resultado establece que las condiciones necesarias para la convergencia del método de Hooke y Jeeves son las mismas que las condiciones para el método de descenso por coordenadas.

Proposición 4.6. *Supongamos que se cumplen los siguientes supuestos:*

- *La función f es diferenciable.*
- *El mínimo de cualquier búsqueda lineal asociada a la función f es único.*
- *La sucesión de puntos generada por el algoritmo está contenida en un conjunto compacto.*

Entonces, todo punto de acumulación de la sucesión generada por el método de Hooke y Jeeves cumple $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$.

En la Figura 4.12 podemos ver como, efectivamente, el método de Hooke y Jeeves corrige significativamente el efecto del zigzag asociado al método de descenso por coordenadas. Aunque la convergencia sigue siendo bastante lenta en el ejemplo en cuestión, el número de iteraciones para conseguir la precisión de 10^{-4} es cuatro veces menor, quedándose en 200.

Es interesante terminar mencionando brevemente el método de Rosenbrock (1960) que, en cierto modo, se puede ver como un refinamiento del método de Hooke y Jeeves. La idea es realizar ciclos de búsquedas lineales siguiendo las direcciones marcadas por una base de vectores

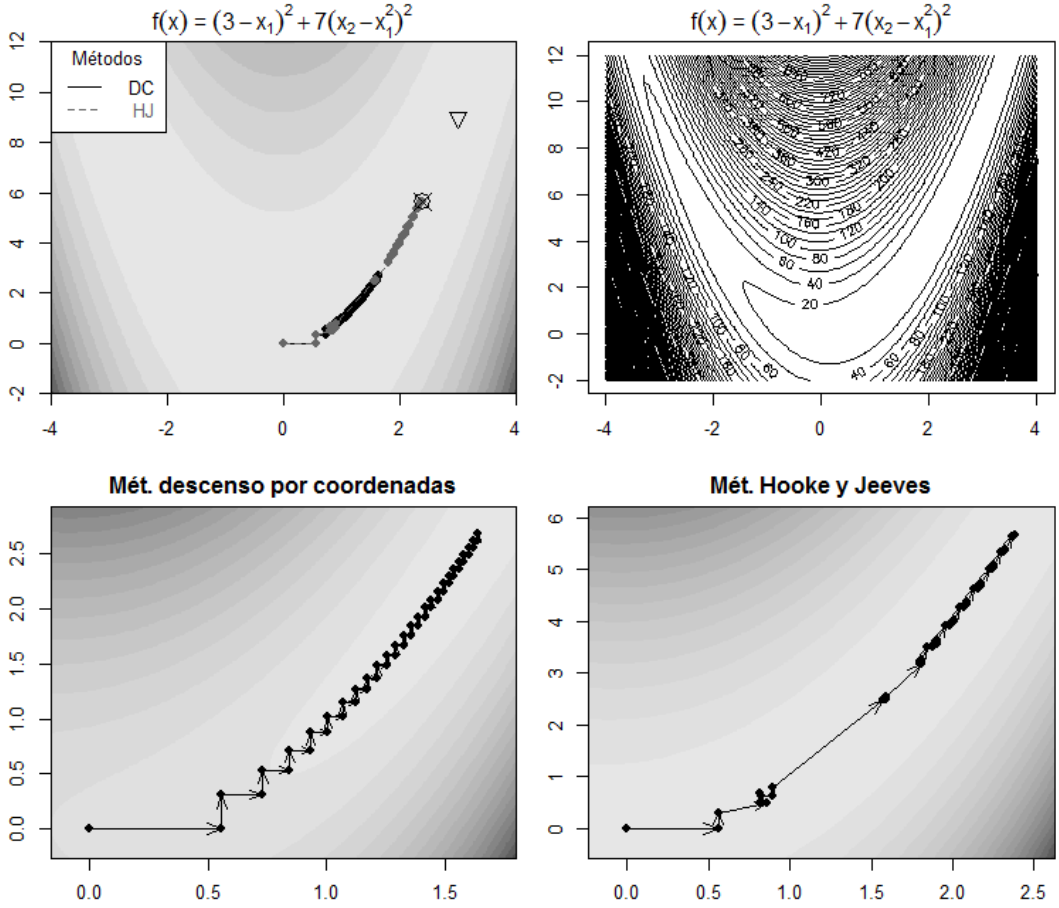


Figura 4.12: Ejemplo en el que el método de descenso por coordenadas tiene una convergencia lenta por culpa del fenómeno de zigzag.

ortogonales, pero sin atarse a las direcciones de los ejes coordenados. Esquemáticamente, este método consistiría en lo siguiente:

- Se realiza un primer ciclo de búsqueda en las direcciones $\boldsymbol{e}^j$ (como harían el método de descenso por coordenadas y Hooke y Jeeves).
- Se realiza un paso de aceleración en la dirección $\boldsymbol{d}^1 = \boldsymbol{x}^{t+1} - \boldsymbol{x}^t$ (como haría el método de Hooke y Jeeves).
- Ahora, en vez de volver a las direcciones $\boldsymbol{e}^j$, se definen vectores $\boldsymbol{d}^2 \dots \boldsymbol{d}^n$ de tal manera que $\{\boldsymbol{d}^1, \boldsymbol{d}^2, \dots, \boldsymbol{d}^n\}$ sea una base ortogonal.[9]
- En iteraciones posteriores se repite el mismo procedimiento. Al final de cada ciclo de n búsquedas, se obtiene una nueva base ortogonal a partir de la dirección que marca el patrón identificado en dicho ciclo.

[9] Esto se suele hacer mediante el método de ortogonalización de Gram-Schmidt.

Método de Hooke y Jeeves

Inicialización
Elegir $\varepsilon > 0$. Elegir un punto inicial $\boldsymbol{x}^1$. Definir $\boldsymbol{y}^1 = \boldsymbol{x}^1$ y $t = j = 1$.

Paso 1
Calcular λ^j, solución del problema minimizar$_{\lambda\in\mathbb{R}}\, f(\boldsymbol{y}^j + \lambda \boldsymbol{e}^j)$. Definir $\boldsymbol{y}^{j+1} = \boldsymbol{y}^j + \lambda^j \boldsymbol{e}^j$.

- Si $j < n$ reemplazar j por $j + 1$ y repetir el Paso 1.
- Si $j = n$, ir al Paso 2.

Paso 2
Definir $\boldsymbol{x}^{t+1} = \boldsymbol{y}^{n+1}$.

- Si $\frac{\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|}{\|\boldsymbol{x}^t\|} \leq \varepsilon$, Fin. Devolvemos $\boldsymbol{x}^{t+1}$.
- En otro caso, ir al Paso 3.

Paso 3 (aceleración)
Definir $\boldsymbol{d} = \boldsymbol{x}^{t+1} - \boldsymbol{x}^t$ y calcular $\bar{\lambda}$, solución del problema minimizar$_{\lambda\in\mathbb{R}}\, f(\boldsymbol{x}^{t+1} + \lambda \boldsymbol{d})$. Definir $\boldsymbol{y}^1 = \boldsymbol{x}^{t+1} + \bar{\lambda} d$ y $j = 1$. Reemplazar t por $t + 1$ e ir al Paso 1.

Figura 4.13: Esquema del método de Hooke y Jeeves.

En su origen, tanto el método de Hooke y Jeeves como el método de Rosenbrock fueron definidos como métodos discretos que, en cada iteración, en vez de realizar búsquedas de línea, simplemente realizaban evaluaciones de la función objetivo en los puntos obtenidos al moverse con paso δ y $-\delta$ en la dirección actual y aceptando el nuevo punto en caso de que se produjese una mejora en la función objetivo. En ambos algoritmos era clave que el paso δ se fuese ajustando dinámicamente durante las distintas iteraciones.[10]

4.8 Optimización multidimensional usando derivadas

En esta sección veremos como la búsqueda de direcciones de descenso apropiadas puede ser hecha de modo mucho más eficiente si podemos apoyarnos en información relativa a las derivadas de la función en estudio.

4.8.1 Método de máximo descenso

Se trata de un método clásico, también conocido como el método de Cauchy, pues fue él quien lo propuso en 1847, para una referencia más reciente se puede consultar Gonzaga (1990). La idea del método es la base de muchos de los métodos más importantes de optimización multidimensional sin restricciones.

Recordemos que, dada una función $f : \mathbb{R}^n \to \mathbb{R}$ y un punto $\boldsymbol{x} \in \mathbb{R}^n$, un vector $\boldsymbol{d} \in \mathbb{R}^n$ es una dirección de descenso si existe $\delta > 0$ tal que, para todo $\lambda \in (0, \delta)$, $f(\boldsymbol{x} + \lambda \boldsymbol{d}) < f(\boldsymbol{x})$. En

[10] Un paso fijo no es deseable ya que si es demasiado pequeño resultará en un algoritmo muy lento y si es demasiado grande puede detenerse en puntos que están lejos de cumplir alguna condición de optimalidad. Por supuesto, lo que es grande o pequeño depende de la función en cuestión y en general no se puede saber con antelación.

particular, si para una dirección $\boldsymbol{d}$ tenemos que $f'(\boldsymbol{x},\boldsymbol{d}) = \lim_{\lambda\to 0^+} \frac{f(\boldsymbol{x}+\lambda\boldsymbol{d})-f(\boldsymbol{x})}{\lambda} < 0$, entonces $\boldsymbol{d}$ es una dirección de descenso. El método de máximo descenso busca la dirección d que minimice este último límite, es decir, la dirección en la que, localmente, más rápido esté descendiendo la función f en el punto x. El siguiente resultado muestra que esa dirección es precisamente la dirección opuesta al gradiente y por eso este método se conoce también como el método del gradiente.

Proposición 4.7. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función diferenciable en un punto $\boldsymbol{x} \in \mathbb{R}^n$ tal que $\nabla f(\boldsymbol{x}) \neq \boldsymbol{0}$. Entonces, la solución óptima del problema de* minimizar $f'(\boldsymbol{x},\boldsymbol{d})$ *sujeto a $\|\boldsymbol{d}\| \leq 1$ viene dada por $\bar{\boldsymbol{d}} = \frac{-\nabla f(x)}{\|\nabla f(x)\|}$. En otras palabras, la dirección de máximo descenso de f en $\boldsymbol{x}$ viene dada por $-\nabla f(x)$.*

Demostración. Dado que f es diferenciable en $\boldsymbol{x}$ tenemos que

$$f'(\boldsymbol{x},\boldsymbol{d}) = \lim_{\lambda\to 0^+} \frac{f(\boldsymbol{x}+\lambda\boldsymbol{d}) - f(\boldsymbol{x})}{\lambda} = \nabla f(\boldsymbol{x})^\intercal \boldsymbol{d},$$

con lo que queremos encontrar $\bar{\boldsymbol{d}}$, el mínimo de $\nabla f(\boldsymbol{x})^\intercal \boldsymbol{d}$ sujeto a $\|\boldsymbol{d}\| \leq 1$. Ahora bien, si denotamos por θ_d al ángulo que forman los vectores $\nabla f(\boldsymbol{x})$ y $\boldsymbol{d}$, tenemos que

$$\nabla f(\boldsymbol{x})^\intercal \boldsymbol{d} = \|\nabla f(\boldsymbol{x})\|\|\boldsymbol{d}\|\cos(\theta_d) \geq -\|\nabla f(\boldsymbol{x})\|\|\boldsymbol{d}\| \geq -\|\nabla f(\boldsymbol{x})\|.$$

Además, las desigualdades anteriores sólo son igualdades cuando $\cos(\theta_d) = -1$ y $\|d\| = 1$, lo que implica que $\bar{\boldsymbol{d}}$ va en la dirección opuesta a $\nabla f(\boldsymbol{x})$ y tiene norma 1 o, lo que es lo mismo, $\bar{\boldsymbol{d}} = \frac{-\nabla f(x)}{\|\nabla f(x)\|}$. □

En la Figura 4.14 representamos el método de máximo descenso, que consiste en la realización sucesiva de búsquedas lineales en las direcciones que vayan marcando los gradientes. Dado que el objetivo del algoritmo es encontrar un punto crítico ($\nabla f(\boldsymbol{x}) = \boldsymbol{0}$) hemos tomado como criterio de parada que la norma del gradiente sea lo suficientemente pequeña, aunque este criterio no es esencial para el algoritmo y se podrían tomar otros.[11]

Método de máximo descenso

Inicialización
Elegir $\varepsilon > 0$. Elegir un punto inicial $\boldsymbol{x}^1$. Definir $t = 1$.

Paso 1
- Si $\|\nabla f(\boldsymbol{x}^t)\| < \varepsilon$, Fin. Devolvemos $\boldsymbol{x}^t$.
- En otro caso, definir $\boldsymbol{d}^t = -\nabla f(\boldsymbol{x}^t)$. Calcular λ^t, solución de minimizar$_{\lambda\geq 0}\, f(\boldsymbol{x}^t + \lambda\boldsymbol{d}^t)$. Definir $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \lambda^t\boldsymbol{d}^t$. Reemplazar t por $t+1$. Repetir el Paso 1.

Figura 4.14: Esquema del método de máximo descenso.

Proposición 4.8. *Supongamos que se cumplen los siguientes supuestos:*

[11] De hecho, para facilitar las comparaciones entre algoritmos, en las implementaciones usadas para generar los ejemplos de estas notas se ha tomado la distancia relativa como criterio de parada (posiblemente complementada con algún otro criterio, dependiendo del algoritmo).

- *La función f es continuamente diferenciable.*[12]
- *La sucesión de puntos generada por el algoritmo está contenida en un conjunto compacto.*

Entonces, todo punto de acumulación $\bar{\boldsymbol{x}}$ de la sucesión generada por el método de máximo descenso cumple $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$.

Uno de los principales motivos de la popularidad del método de máximo descenso es su sencillez de implementación. Sin embargo, este método tiene importantes limitaciones y es por eso que los métodos que presentamos en las siguientes secciones suelen ser preferibles (métodos que, de uno u otro modo, buscan direcciones apropiadas a partir del gradiente).

El método de gradiente también es muy susceptible de sufrir el problema del zigzag que ya comentamos en el caso del método de descenso por coordenadas. Intuitivamente, lo que sucede es que en cada iteración t "agotamos" lo que nos podemos mover desde $\boldsymbol{x}^t$ en la dirección del gradiente y, por lo tanto, como desde el punto $\boldsymbol{x}^{t+1}$ no podemos mejorar moviéndonos en la dirección de $\nabla f(\boldsymbol{x}^t)$ (ni con pasos positivos ni con pasos negativos), tendremos que $\nabla f(\boldsymbol{x}^{t+1})$ será ortogonal a $\nabla f(\boldsymbol{x}^t)$. Esta ortogonalidad de iterantes consecutivos del método es la que llevará al zigzag. De hecho, si volvemos a la Figura 4.12, donde se ilustró el zigzag del método de descenso por coordenadas para la función $f(\boldsymbol{x}) = (3 - x_1)^2 + 7(x_2 - x_1^2)^2$, tenemos que $\nabla f(0, 0) = c(-6, 0)$, con lo que en la primera iteración del método de máximo descenso se hará una búsqueda en la dirección $\boldsymbol{e}^1$. Como acabamos de argumentar, la búsqueda en la segunda iteración será entonces en la dirección $\boldsymbol{e}^2$, volviendo a $\boldsymbol{e}^1$ en la tercera y así sucesivamente. Por tanto, en este ejemplo el método de máximo descenso coincide exactamente con el método de descenso por coordenadas, sufriendo entonces el mismo problema de zigzag y convergencia lenta.

El hecho de que el método de gradiente coincida en este ejemplo con el método de descenso por coordenadas no tiene mayor trascendencia y rara vez se producirá en problemas en $\mathbb{R}^n$ con $n > 2$, pues el gradiente de la primera iteración no condicionará de manera tan crítica el comportamiento del algoritmo en las iteraciones siguientes (pues el espacio de vectores ortogonales al gradiente tendrá dimensión mayor que 1). En cualquier caso, la ortogonalidad de las direcciones usadas en iteraciones consecutivas sí que puede seguir dando lugar al fenómeno de zigzag.

A continuación presentamos un resultado relativo para la velocidad de convergencia del método de máximo descenso, para el cual hemos de introducir previamente un concepto importante en el diseño de algoritmos. Supongamos que $\boldsymbol{A}_{n\times n}$ es una matriz diagonalizable y que $\lambda_{\text{máx}}$ y $\lambda_{\text{mín}}$ denotan, respectivamente, los autovalores de $\boldsymbol{A}$ con mayor y menor valor absoluto. Entonces, el número de condicionamiento de la matriz $\boldsymbol{A}$, $\kappa(\boldsymbol{A})$, se define como

$$\kappa(\boldsymbol{A}) = \frac{|\lambda_{\text{máx}}|}{|\lambda_{\text{mín}}|}.$$

En general, cuanto más grande sea el número de condicionamiento de una matriz, más problemas experimentan los algoritmos que, directa o indirectamente, dependen de dicha matriz. Una de las razones es que un valor grande de $\kappa(\boldsymbol{A})$ puede dar lugar a problemas de estabilidad numérica, ya que un amplio rango de valores entre $|\lambda_{\text{máx}}|$ y $|\lambda_{\text{mín}}|$ suele tener asociados

[12] Una función $f : \mathbb{R}^n \to \mathbb{R}$ es continuamente diferenciable si ∇f es una función continua.

problemas de precisión. En el caso del método de gradiente, el condicionamiento también está muy relacionado en el fenómeno de zigzag del método, y esto es lo que captura el siguiente resultado.

Proposición 4.9. *Supongamos que $f : \mathbb{R}^n \to \mathbb{R}$ es una función dos veces diferenciable y que $\bar{\boldsymbol{x}}$ es un mínimo local al que converge una sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ generada por el método de máximo descenso. Entonces, si la matriz hessiana $\boldsymbol{H}(\bar{\boldsymbol{x}})$ es definida positiva, la sucesión $\{f(\boldsymbol{x}^t)\}_{t\in\mathbb{N}}$ converge a $f(\bar{\boldsymbol{x}})$ a una velocidad lineal, es decir, asintóticamente tenemos que*

$$|f(\boldsymbol{x}^{t+1}) - f(\bar{\boldsymbol{x}})| = \rho|f(\boldsymbol{x}^t) - f(\bar{\boldsymbol{x}})|,$$

donde, además, la tasa ρ está acotada por $\frac{(\kappa(\boldsymbol{H}(\bar{\boldsymbol{x}}))-1)^2}{(\kappa(\boldsymbol{H}(\bar{\boldsymbol{x}}))+1)^2}$.

La Proposición 4.9 es un resultado negativo para el método de máximo descenso. No sólo la velocidad de convergencia del método será como mucho lineal, sino que además la tasa asociada depende del condicionamiento de la matriz hessiana en el punto límite $\bar{\boldsymbol{x}}$. En particular, como $\frac{(\kappa(\boldsymbol{H}(\bar{\boldsymbol{x}}))-1)^2}{(\kappa(\boldsymbol{H}(\bar{\boldsymbol{x}}))+1)^2}$ es creciente en $\kappa(\boldsymbol{H}(\bar{\boldsymbol{x}}))$ y converge a 1 cuando $\kappa(\boldsymbol{H}(\bar{\boldsymbol{x}})) \to \infty$, la convergencia puede llegar a ser muy lenta para valores grandes de $\kappa(\boldsymbol{H}(\bar{\boldsymbol{x}})$.

Este resultado negativo también puede ser explicado de modo intuitivo. En cierta manera, la calidad de un paso del método de máximo descenso en una iteración t depende de cómo de buena sea la aproximación de Taylor de primer orden de f en $\boldsymbol{x}^t$. La propia definición del gradiente nos asegura que

$$f(\boldsymbol{x}^t + \lambda\boldsymbol{d}) = f(\boldsymbol{x}^t) + \lambda\,\nabla f(\boldsymbol{x}^t)^\intercal\boldsymbol{d} + \lambda\varphi(\lambda\boldsymbol{d}),$$

donde $\lim_{\lambda\to 0}\varphi(\lambda\boldsymbol{d}) = 0$ y $\boldsymbol{d}$ es una dirección de búsqueda con $\|\boldsymbol{d}\| = 1$. Si $\boldsymbol{x}^t$ está cerca de un punto $\bar{\boldsymbol{x}}$ con $\nabla f(\bar{\boldsymbol{x}}) = \boldsymbol{0}$ y f es continuamente diferenciable, entonces $\|\nabla f(\boldsymbol{x}^t)\|$ será un número pequeño. Por tanto, como $\|\boldsymbol{d}\| = 1$, $\lambda\,\nabla f(\boldsymbol{x}^t)^\intercal\boldsymbol{d}$ será también pequeño. El método de máximo descenso se basa en la linealización en torno a $\boldsymbol{x}^t$ para encontrar la dirección de búsqueda, descartando posible influencia del término $\lambda\varphi(\lambda\boldsymbol{d})$. Sin embargo, cuando $\lambda\,\nabla f(\boldsymbol{x}^t)^\intercal\boldsymbol{d}$ es pequeño, la influencia de $\lambda\varphi(\lambda\boldsymbol{d})$ deja de ser despreciable y las direcciones basadas en la linealización en torno a $\boldsymbol{x}^t$ dejan de ser efectivas.

La forma más habitual de modificar el método de máximo descenso para mejorar su comportamiento consiste en *desviar el gradiente*, siendo esta la idea principal de los métodos que discutiremos en el resto de esta sección. En vez de movernos directamente en la dirección $-\nabla f(\boldsymbol{x})$, usaremos una dirección de la forma $-\boldsymbol{A}\,\nabla f(\boldsymbol{x})$ o $-\nabla f(\boldsymbol{x}) + \boldsymbol{v}$ donde la matriz $\boldsymbol{A}$ y el vector $\boldsymbol{v}$ han de ser elegidos de modo apropiado. De hecho, es habitual que la matriz $\boldsymbol{A}$ use directa o indirectamente información de la hessiana de f. Esto puede ser muy útil pues, la aproximación de segundo orden de una función en torno a un punto crítico sí que es efectiva, lo que es clave para las tasas de convergencia cuadráticas que obtienen habitualmente estos métodos.

4.8.2 Método de Newton

En esta sección presentamos la generalización al caso multidimensional del método de Newton introducido en la Sección 4.5.2 como método para realizar búsquedas de línea. Recordemos que

el método de máximo descenso elige la dirección de búsqueda basándose en una aproximación de orden uno de la función, es decir, la linealización de la función en torno al punto en cuestión, y que se basa en el uso del gradiente. La idea del método de Newton consiste en ir un paso más allá y tomar dicha dirección basándose en la aproximación de orden dos de la función. De hecho, en su versión más clásica, el método de Newton ni siquiera usa búsquedas de línea, realizando un paso unitario en la dirección obtenida al desviar el gradiente multiplicándolo por la inversa de la matriz hessiana:

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \boldsymbol{H}(\boldsymbol{x}^t)^{-1}\,\nabla f(\boldsymbol{x}^t).$$

Al igual que en el caso unidimensional, esta expresión surge de tomar como nuevo iterante el punto donde se anula el gradiente de la aproximación cuadrática, $q(\boldsymbol{x})$, de la función en torno al iterante actual. Formalmente,

$$q(\boldsymbol{x}) = f(\boldsymbol{x}^t) + \nabla f(\boldsymbol{x}^t)^\intercal(\boldsymbol{x} - \boldsymbol{x}^t) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^t)^\intercal \boldsymbol{H}(\boldsymbol{x}^t)(\boldsymbol{x} - \boldsymbol{x}^t).$$

La expresión $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \boldsymbol{H}(\boldsymbol{x}^t)^{-1}\,\nabla f(\boldsymbol{x}^t)$ se obtiene sin más que resolver la ecuación $\nabla q(\boldsymbol{x}) = \boldsymbol{0}$ (y el cálculo es esencialmente el mismo ya presentado para el caso unidimensional). En la Figura 4.15 representamos el esquema completo del método de Newton básico (en la práctica se suele implementar acompañado de una búsqueda de línea).

Método de Newton

Inicialización

Elegir $\varepsilon > 0$. Elegir un punto inicial $\boldsymbol{x}^1$. Definir $t = 1$.

Paso 1

- Si $\|\nabla f(\boldsymbol{x}^t)\| < \varepsilon$, Fin. Devolvemos $\boldsymbol{x}^t$.
- En otro caso, definir $\boldsymbol{d}^t = -\boldsymbol{H}(\boldsymbol{x}^t)^{-1}\,\nabla f(\boldsymbol{x}^t)$ y $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \boldsymbol{d}^t$. Reemplazar t por $t+1$. Repetir el Paso 1.

Figura 4.15: Esquema del método de Newton básico.

Aunque el método de Newton es de gran interés desde el punto de vista conceptual, su implementación directa en la práctica tiene bastantes problemas, lo que hace que lo habitual sea usar distintas modificaciones del mismo. Los principales problemas del método de Newton son los siguientes:

P-I No tiene por qué estar bien definido, pues requiere que la matriz hessiana sea invertible en todos los puntos que el método va construyendo.

P-II Incluso si el método está bien definido, la dirección $\boldsymbol{d}^t$ construida por el método no tiene por qué ser una dirección de descenso. Por ejemplo, si la hessiana fuese la opuesta de la matriz identidad, entonces tendríamos $\boldsymbol{d}^t = \nabla f(\boldsymbol{x}^t)$, que es justamente la dirección de máximo crecimiento de la función.

P-III Además, incluso si la dirección $\boldsymbol{d}^t$ es una dirección de descenso, puede ser que con el paso unitario elegido por el método ya no lo sea.

P-IV Desde el punto de vista computacional este método también presenta dos inconvenientes:

- Requiere calcular explícitamente la matriz hessiana en cada iteración, lo cual puede ser muy costoso computacionalmente.
- Además del propio cálculo de la matriz hessiana, el método requiere hacer manipulaciones con la misma: cálculo de la inversa y multiplicación matriz por vector. Cuando la dimensión del problema crece, el requerir manipulaciones matriciales en vez de simplemente vectoriales (como las del método de máximo descenso) puede resultar en mucho tiempo invertido en cada iteración del algoritmo.

Desde el punto de vista de la convergencia teórica del algoritmo y de la velocidad de dicha convergencia (pensada como el orden del número de iteraciones necesario para acercarse suficientemente al óptimo), P-IV no es relevante, aunque sí lo es por supuesto a la hora de decidirse por uno u otro algoritmo; eso sí, en la práctica, un buen equilibrio entre el número necesario de iteraciones y tiempo requerido por iteración es fundamental para el comportamiento de un algoritmo. P-III tiene fácil arreglo, pues bastaría con modificar el método de Newton para que haga una búsqueda de línea en cada iteración en la dirección $\boldsymbol{d}$. Si nos fijamos ahora en los puntos P-I y P-II, y pensamos que nuestro algoritmo debe encontrar un óptimo local, tenemos que cerca del mismo ambos problemas desaparecen:

- Por un lado la optimalidad local del punto asegura que la matriz es definida positiva y por tanto invertible.
- Si $\boldsymbol{H}(\boldsymbol{x}^t)$ es definida positiva también lo es $\boldsymbol{H}(\boldsymbol{x}^t)^{-1}$, con lo que $\boldsymbol{d} = -\boldsymbol{H}(\boldsymbol{x}^t)^{-1}\nabla f(\boldsymbol{x}^t)$ es una dirección de descenso pues

$$\nabla f(\boldsymbol{x}^t)^\intercal \boldsymbol{d} = -\nabla f(\boldsymbol{x}^t)^\intercal \boldsymbol{H}(\boldsymbol{x}^t)^{-1}\nabla f(\boldsymbol{x}^t) < 0.$$

Son precisamente estas dos observaciones las que permiten obtener el siguiente resultado de convergencia para el método de Newton.

Proposición 4.10. *Supongamos que $f : \mathbb{R}^n \to \mathbb{R}$ es una función dos veces continuamente diferenciable. Sea $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ tal que $\nabla f(\bar{\boldsymbol{x}}) = \boldsymbol{0}$ y $H(\bar{\boldsymbol{x}})$ es invertible. Si partimos de un punto $\boldsymbol{x}^1$ suficientemente próximo a $\bar{\boldsymbol{x}}$, entonces el método de Newton converge a $\bar{\boldsymbol{x}}$, y el orden de convergencia es al menos cuadrático.*

Más concretamente, es necesario que existan constantes α_1 y α_2 tales que:

(i) $\|\boldsymbol{H}(\bar{\boldsymbol{x}})^{-1}\| \leq \alpha_1$.[13]

(ii) $\|\nabla f(\bar{\boldsymbol{x}}) - \nabla f(\boldsymbol{x}) - \boldsymbol{H}(\boldsymbol{x})(\bar{\boldsymbol{x}} - \boldsymbol{x})\| \leq \alpha_2\|\bar{\boldsymbol{x}} - \boldsymbol{x}\|^2$ *para todo $\boldsymbol{x}$ tal que* $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq \|\boldsymbol{x}^1 - \bar{\boldsymbol{x}}\|$.

Es importante destacar que los supuestos de la Proposición 4.10 se cumplen, en particular, si $\bar{\boldsymbol{x}}$ es un mínimo o un máximo local (estricto). Por tanto, salvo que estemos trabajando con funciones convexas, en cuyo caso la condición $\nabla f(\bar{\boldsymbol{x}}) = \boldsymbol{0}$ asegura que estamos ante un mínimo global, hay que tener cuidado al aplicar el método de Newton. Es habitual diseñar métodos

[13] El resultado es cierto, por ejemplo, para la norma matricial definida como $\|\boldsymbol{A}\| = \max_{\|\boldsymbol{x}\|=1}\|\boldsymbol{A}\boldsymbol{x}\|$.

basados en el método de Newton que lejos del óptimo tengan un comportamiento parecido al del método de máximo descenso y cerca del mismo tengan el orden de convergencia del método de Newton. En esta línea discutiremos el método de Levenberg-Marquardt, aunque antes presentamos una reinterpretación del método de Newton.

Reinterpretación del método de Newton como un método de máximo descenso

Consideremos ahora un ejemplo como el discutido en la Sección 2.3, que consistía en minimizar la distancia al punto $\bar{\boldsymbol{x}} = (\frac{3}{2}, 5)$ sujeto a una serie de restricciones. Más concretamente, la función a minimizar era $f(\boldsymbol{x}) = (x_1 - \frac{3}{2})^2 + (x_2 - 5)^2$. En la Figura 2.4(a) representamos las curvas de nivel de esta función y también los gradientes en distintos puntos. En este caso, para todo punto $\boldsymbol{x} \in \mathbb{R}^2$, $\nabla f(\boldsymbol{x}) = 2(x_1 - \bar{x}_1, x_2 - \bar{x}_2)$. Es decir, el gradiente apunta en la dirección del vector $\boldsymbol{x} - \bar{\boldsymbol{x}}$ y, por tanto, la recta que pasa por $\boldsymbol{x}$ en la dirección del gradiente pasa también por el punto $\bar{\boldsymbol{x}}$. Esto hace que el método de máximo descenso converja en una iteración para este tipo de problemas, como ilustramos en la Figura 4.16. Por supuesto, el método de Newton, que converge en una iteración para cualquier problema cuadrático también convergerá en una iteración. En este ejemplo, la matriz hessiana se obtiene al multiplicar por dos la matriz identidad, $\boldsymbol{H}(\boldsymbol{x}) = 2\boldsymbol{I}$, y su inversa es $\boldsymbol{H}(\boldsymbol{x})^{-1} = \frac{1}{2}\boldsymbol{I}$. Por tanto, la dirección $\boldsymbol{d} = -\boldsymbol{H}(\boldsymbol{x}^t)^{-1}\nabla f(\boldsymbol{x}^t) = -\frac{1}{2}\nabla f(\boldsymbol{x}^t)$, como era de esperar, pues, si tanto el método de máximo descenso como el método de Newton convergen en una iteración, deberán moverse en la misma dirección.

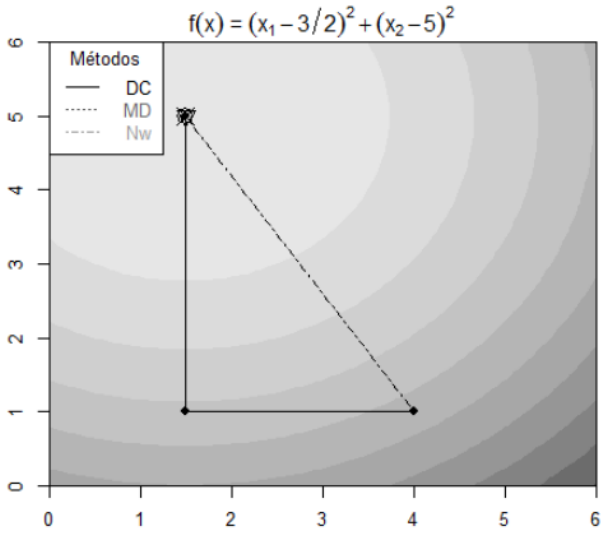


Figura 4.16: Métodos de descenso por coordenadas, máximo descenso y Newton sobre un problema de minimización de distancia euclídea.

Para problemas cuadráticos más generales, el método de máximo descenso ya no tiene por qué converger en una iteración (ni mucho menos) y, sin embargo, el método de Newton mantiene esta propiedad. La razón es que el método de Newton se puede interpretar como la aplicación del método de descenso a una versión "reescalada" del problema original que

tiene la misma propiedad del ejemplo que acabamos de discutir. Para ver esto formalmente, supongamos que $\boldsymbol{H}(x^t)$ es definida positiva y su inversa admite una *factorización de Cholesky* dada por $\boldsymbol{H}(x^t)^{-1} = \boldsymbol{L}\boldsymbol{L}^\intercal$.[14] Consideremos ahora el cambio de variable asociado al reescalado $\boldsymbol{x} = \boldsymbol{L}\boldsymbol{y}$, que transforma la función $f(\boldsymbol{x})$ en $f^*(\boldsymbol{y}) = f(\boldsymbol{L}\boldsymbol{y})$ y el iterante $\boldsymbol{x}^t$ se convierte en $\boldsymbol{y}^t = \boldsymbol{L}^{-1}x^t$. Aplicando la regla de la cadena tenemos que el gradiente en $\boldsymbol{y}^t$ es $\nabla f^*(\boldsymbol{y}^t) = \boldsymbol{L}^\intercal \nabla f(\boldsymbol{L}\boldsymbol{y}^t) = \boldsymbol{L}^\intercal \nabla f(\boldsymbol{x}^t)$. Entonces, un paso unitario en la dirección opuesta al gradiente nos llevaría al punto $\boldsymbol{y}^{t+1} = \boldsymbol{y}^t - \boldsymbol{L}^\intercal \nabla f(\boldsymbol{x}^t)$ que, en el espacio de variables original sería

$$\boldsymbol{x}^{t+1} = \boldsymbol{L}\boldsymbol{y}^{t+1} = \boldsymbol{L}\boldsymbol{y}^t - \boldsymbol{L}\boldsymbol{L}^\intercal \nabla f(\boldsymbol{x}^t) = \boldsymbol{x}^t - \boldsymbol{H}(\boldsymbol{x}^t)^{-1} \nabla f(\boldsymbol{x}^t),$$

que es justamente un paso del método de Newton. Resumiendo, esta reinterpretación nos permite ver el método de Newton como un cambio de coordenadas que en particular transforma cualquier problema cuadrático en un problema de minimización de la distancia euclídea como el de la Figura 4.16.

Método de Levenberg-Marquardt (o método de Gauss-Newton)

Este método consiste en una variación natural del método de Newton, que permite solventar de manera sencilla los problemas P-I, P-II y P-III. Además, como comentaremos más adelante, el método de Levenberg-Marquardt surgió en el contexto de los problemas de ajustes de mínimos cuadrados, en los cuales también solventa bastante exitosamente el problema P-IV. Históricamente, el método fue introducido independientemente por Levenberg (1944) y Marquardt (1963), aunque fue Gauss quien, más de 100 años antes, observó que el método de Newton podía adaptarse para resolver problemas de ajustes de mínimos cuadrados sin requerir el uso de la matriz hessiana.

La idea del método consiste en reemplazar el uso de $\boldsymbol{H}(\boldsymbol{x}^t)^{-1}$ por el de una aproximación $\boldsymbol{A}$, simétrica y definida positiva. Esto asegura que siempre tendremos direcciones de descenso y permite solucionar los problemas P-I y P-II. Además, complementado con una búsqueda de línea soluciona también el problema P-III. Por supuesto, si tomamos como matriz $\boldsymbol{A}$ la matriz identidad, pasamos al método de máximo descenso y perdemos todas las bondades del método de Newton, con lo que es importante que $\boldsymbol{A}$ sea efectivamente una aproximación de $\boldsymbol{H}(\boldsymbol{x}^t)^{-1}$. El método de Levenberg-Marquardt se basa en trabajar con la matriz $(\varepsilon^t \boldsymbol{I} + \boldsymbol{H}(\boldsymbol{x}^t))^{-1}$, donde ε^t debería ser el menor escalar que asegure que todos los autovalores de $\varepsilon^t \boldsymbol{I} + \boldsymbol{H}(\boldsymbol{x}^t)$ son mayores que un cierto valor prefijado $\delta > 0$. Por tanto, el paso fundamental del método viene dado por:

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \lambda^t \boldsymbol{d}^t,$$

donde λ^t es el paso óptimo en la dirección $\boldsymbol{d}^t = (\varepsilon^t \boldsymbol{I} + \boldsymbol{H}(\boldsymbol{x}^t))^{-1} \nabla f(\boldsymbol{x}^t)$. Por supuesto, todavía queda abierto el tema de determinar, en cada iteración, el valor ε^t. Una de las claves del método de Levenberg-Marquardt, es que este valor se puede calcular aprovechándose de la factorización de la matriz $\varepsilon^t \boldsymbol{I} + \boldsymbol{H}(\boldsymbol{x}^t)$, usada a su vez para resolver el sistema $(\varepsilon^t \boldsymbol{I} + \boldsymbol{H}(\boldsymbol{x}^t))(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t) = \nabla f(\boldsymbol{x}^t)$. Aunque no vamos a entrar aquí en los detalles, hay formas de llevar esto a cabo sin que el cálculo de los ε^t suponga una carga importante para el algoritmo.

[14] Toda matriz simétrica y definida positiva $\boldsymbol{A}$ puede ser descompuesta como $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\intercal$, donde $\boldsymbol{L}$ es una matriz triangular inferior cuyos términos diagonales son todos mayores que cero. Esta descomposición se conoce con el nombre de factorización de Cholesky.

Supongamos ahora que aplicamos este algoritmo a un problema de programación no lineal y que nos encontramos en un punto $\boldsymbol{x}^t$ lejos de un mínimo local. En este caso, es posible que la matriz $\boldsymbol{H}(\boldsymbol{x}^t)$ esté lejos de ser definida positiva, con lo que ε^t deberá tomar un valor grande, $(\varepsilon^t \boldsymbol{I} + \boldsymbol{H}(\boldsymbol{x}^t))$ se parecerá a un múltiplo de la matriz identidad y la iteración actual será muy similar a una iteración del método de máximo descenso (especialmente si lo complementamos con una búsqueda de línea). Por otro lado, si estamos cerca de un mínimo local, entonces $\boldsymbol{H}(\boldsymbol{x}^t)$ será definida positiva, podremos tomar $\varepsilon^t = 0$ y la iteración actual se corresponderá con una iteración del método de Newton. En este sentido, el método de Levenberg-Marquardt tiene un comportamiento muy similar al del método de máximo descenso cuando está lejos de un mínimo y al método de Newton cuando está cerca, con lo que hereda de este último los resultados relativos a su orden de convergencia.

El gran problema de este método, y que hace que en la práctica sean mucho más populares los métodos cuasi-Newton que discutiremos en la siguiente sección, es que requiere el cálculo explícito de la matriz hessiana en cada iteración. Sin embargo, hay una clase de problemas muy importante en la práctica para la cual este método se puede aplicar sin necesidad de trabajar con las derivadas segundas de las funciones. Se trata de los problemas de mínimos cuadrados, en cuya formulación general se trata de minimizar problemas con función objetivo

$$f(\boldsymbol{x}) = \frac{1}{2}\sum_{l=1}^{m} h^l(\boldsymbol{x})^2 = \frac{1}{2}\|\boldsymbol{h}(\boldsymbol{x})\|^2,$$

donde las funciones $h^l : \mathbb{R}^n \to \mathbb{R}$ son dos veces diferenciables y la constante $\frac{1}{2}$ está simplemente por comodidad para los cálculos. Este tipo de problemas aparecen habitualmente en identificación de modelos (muy habitual en estadística y en el estudio de modelos físicos y químicos), donde $\boldsymbol{x}$ representa un vector de parámetros desconocidos en un modelo de la forma $\boldsymbol{y} = \varphi(\boldsymbol{x}, \boldsymbol{z})$, con $\varphi : \mathbb{R}^{n+s} \to \mathbb{R}$, donde $\boldsymbol{z}$ representa las variables de entrada del modelo ($\boldsymbol{z}$ variable explicativa e $\boldsymbol{y}$ variable explicada). Entonces, dada una colección de observaciones $(\boldsymbol{z}^l, \boldsymbol{y}^l)$, con $l \in \{1, \ldots, m\}$, tenemos que las funciones $h^l(\boldsymbol{x})$ representarán las diferencias entre los valores observados, $\boldsymbol{y}^l$, y los valores predichos por el modelo:

$$h^l(\boldsymbol{x}) = \boldsymbol{y}^l - \varphi(\boldsymbol{x}, \boldsymbol{z}^l).^{15}$$

Por tanto, la minimización de $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{h}(\boldsymbol{x})\|^2$ es equivalente a la minimización de los errores cuadráticos del ajuste. En el caso de que la función φ sea lineal tendríamos un problema clásico de regresión lineal, donde la solución del problema de mínimos cuadrados puede ser obtenida de forma explícita de múltiples maneras. Sin embargo, cuando el ajuste es no lineal, la resolución explícita del problema de mínimos cuadrados ya no es posible y es necesario recurrir a métodos generales de optimización sin restricciones.

Veamos ahora por qué el método de Levenberg-Marquardt es especialmente apropiado para este tipo de problemas. Para empezar, es fácil obtener las siguientes expresiones para el gradiente y la hessiana de la función f (para evitar confusión entre la hessiana de f y la hessiana

[15] La extensión al caso de mínimos cuadrados ponderados es inmediata, sin más que poner un peso w^l multiplicando cada término $\boldsymbol{y}^l - \varphi(\boldsymbol{x}, \boldsymbol{z}^l)$.

de las funciones h^l, denotamos la hessiana por ∇^2):

$$\nabla f(\boldsymbol{x}) = \sum_{l=1}^{m} h^l(\boldsymbol{x}) \nabla h^l(\boldsymbol{x}),$$
$$\nabla^2 f(\boldsymbol{x}) = \sum_{l=1}^{m} \nabla h^l(\boldsymbol{x})^\intercal \nabla h^l(\boldsymbol{x}) + \sum_{l=1}^{m} h^l(\boldsymbol{x}) \nabla^2 h^l(\boldsymbol{x}).$$

A la hora de aplicar el método de Levenberg-Marquardt a problemas de mínimos cuadrados, tendríamos que el problema computacional P-IV aparece en el cálculo del segundo sumando de la hessiana de f, que es donde aparece la hessiana de los ajustes, $\nabla^2 h^l(\boldsymbol{x})$. La aplicación del método pasa por desechar este segundo sumando, tomando $\boldsymbol{A}(\boldsymbol{x}) = \sum_{l=1}^{m} \nabla h^l(\boldsymbol{x})^\intercal \nabla h^l(\boldsymbol{x})$ como aproximación de la hessiana y trabajando entonces en el método con $\varepsilon^t \boldsymbol{I} + \boldsymbol{A}(\boldsymbol{x}^t)$. Esta aproximación descartando el sumando $\sum_{l=1}^{m} h^l(\boldsymbol{x}) \nabla^2 h^l(\boldsymbol{x})$ tiene muy buen comportamiento en la práctica y está justificada por los siguientes motivos:

- Por un lado, para problemas de mínimos cuadrados bien construidos es de esperar que los ajustes sean razonablemente buenos, con lo que los valores $h^l(\boldsymbol{x})$ deberían tomar valores próximos a cero cerca del ajuste óptimo.
- El término $\sum_{l=1}^{m} h^l(\boldsymbol{x}) \nabla^2 h^l(\boldsymbol{x})$ puede tener efectos no deseados en el ajuste en el caso de que tengamos algún "outlier", l, para el cual $h^l(\boldsymbol{x})$ no va de la mano con el ajuste del resto de observaciones. En este caso, el término $\sum_{l=1}^{m} h^l(\boldsymbol{x}) \nabla^2 h^l(\boldsymbol{x})$ podría provocar que los outliers acabaran teniendo un peso excesivo en el resultado final del ajuste.
- Además, en el caso de que las funciones h^l sean sólo ligeramente no lineales, los valores de las hessianas $\nabla^2 h^l(\boldsymbol{x})$ también serán pequeños.

4.8.3 Direcciones conjugadas: Gradiente conjugado y cuasi-Newton

En la sección anterior, en la discusión del método de Newton, comentamos que el método de Newton se puede ver como un método de gradiente aplicado sobre una transformación del problema original mediante un cambio de variable que se apoyaba en la matriz hessiana. En particular, es el uso de la hessiana el que permite que la convergencia en una iteración del método de gradiente para la minimización de distancias euclídeas, como el ilustrado en la Figura 4.16, se obtenga con el método de Newton para cualquier función cuadrática.

A pesar de eso, el mayor problema del método de Newton (y heredado por Levenberg-Marquardt, aunque bastante atenuado en el caso de problemas de mínimos cuadrados) es la necesidad de calcular y procesar la matriz hessiana en todas las iteraciones del método. En esta sección presentamos una importante familia de métodos que mantiene las buenas propiedades del método de Newton sin tener que manipular explícitamente la matriz hessiana.

Tanto el método de Newton como los métodos que discutiremos en esta sección para minimizar funciones no lineales se basan en generalizar métodos que tienen un buen comportamiento en problemas cuadráticos. Una de las razones para seguir este enfoque es que cerca de un mínimo local la función cuadrática dada por la aproximación de Taylor de segundo orden de la función objetivo es una buena aproximación de la misma. Por tanto, un método rara vez va a

tener un buen comportamiento cerca del óptimo de un problema no cuadrático si no tiene un buen comportamiento para problemas cuadráticos.

Direcciones conjugadas y problemas cuadráticos

Consideremos nuevamente el ejemplo de la Figura 4.16, con la función $f(\boldsymbol{x}) = (x_1 - \frac{3}{2})^2 + (x_2 - 5)^2$. Si miramos el comportamiento del método de descenso por coordenadas vemos que alcanza la convergencia en dos iteraciones independientemente del iterante inicial. En la primera iteración nos moveremos en la dirección $(1, 0)$ hasta situarnos en la vertical del punto $(\frac{3}{2}, 5)$, es decir, la primera coordenada será igual a $\frac{3}{2}$. En la segunda iteración nos moveremos en la dirección $(0, 1)$ hasta situarnos en el punto $(\frac{3}{2}, 5)$, con función objetivo igual a cero. ¿Qué tiene de especial este problema para que el método de descenso por coordenadas lo resuelva en sólo dos iteraciones? La clave está en que la función $f(\boldsymbol{x})$ es separable, pues puede ser escrita como dos sumandos independientes para cada una de las dos coordenadas x_1 y x_2.

Veamos ahora cómo formalizar este concepto para funciones cuadráticas generales. Dada una función $f : \mathbb{R}^n \to \mathbb{R}$, decimos que f es una función cuadrática si se puede expresar como[16]

$$f(\boldsymbol{x}) = b + \boldsymbol{c}^\intercal \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^\intercal \boldsymbol{Q}\boldsymbol{x} = b + \sum_{i=1}^{n} c_i x_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} Q_{ij} x_i x_j.$$

Además, este problema es separable si $\boldsymbol{Q}$ es una matriz diagonal, en cuyo caso tenemos

$$f(\boldsymbol{x}) = b + \sum_{i=1}^{n} (c_i x_i + \frac{1}{2} Q_{ii} (x_i)^2).$$

En el caso del ejemplo de la Figura 4.16 tenemos que es separable pues

$$f(\boldsymbol{x}) = (x_1 - \frac{3}{2})^2 + (x_2 - 5)^2 = \frac{109}{4} + (-3x_1 + (x_1)^2) + (-10x_2 + (x_2)^2).$$

Es precisamente esta separabilidad la que hace que el mínimo de la función n-dimensional se pueda calcular mediante la minimización sucesiva (e independiente) en cada una de las n direcciones de los ejes coordenados. Del mismo modo que el método de Newton permitía replicar el comportamiento del método del gradiente para problemas separables en problemas cuadráticos generales, se puede decir que los métodos de direcciones conjugadas buscan hacer lo propio con respecto al método de descenso por coordenadas. Más concretamente, son métodos que al aplicarse sobre problemas cuadráticos consiguen identificar n direcciones de tal manera que el óptimo del problema se consigue mediante la minimización sucesiva (e independiente) en estas direcciones. Es aquí donde entra el concepto de direcciones conjugadas, que definimos a continuación.

Definición 4.1. Sea $\boldsymbol{Q}_{n\times n}$ una matriz simétrica y definida positiva. Las direcciones $\boldsymbol{d}^1, \ldots, \boldsymbol{d}^n$ se denominan $\boldsymbol{Q}$*-conjugadas* si son todas distintas de cero y son $\boldsymbol{Q}$*-ortogonales*. Es decir, para todo $i \neq j$,

$$(\boldsymbol{d}^i)^\intercal \boldsymbol{Q}\boldsymbol{d}^j = 0.$$

[16] Nuevamente, el uso del coeficiente $\frac{1}{2}$ es para obtener expresiones más sencillas al calcular el gradiente y la matriz hessiana.

Si recordamos el método de máximo descenso, allí teníamos que dos direcciones $\boldsymbol{d}^1$ y $\boldsymbol{d}^2$ usadas en iteraciones consecutivas del método tenían la propiedad de que eran ortogonales, $(\boldsymbol{d}^1)^\intercal \boldsymbol{d}^2 = 0$, y que esta propiedad era en parte culpable de los fenómenos de zigzag que puede presentar el método en la práctica. El concepto de direcciones conjugadas en vez de direcciones ortogonales es la clave de los algoritmos que veremos a continuación. El siguiente resultado muestra que las direcciones conjugadas se pueden usar para definir un nuevo sistema de coordenadas.

Proposición 4.11. *Si las direcciones $\boldsymbol{d}^1, \ldots, \boldsymbol{d}^n$ son $\boldsymbol{Q}$-conjugadas, entonces son linealmente independientes.*

Demostración. Supongamos que existen coeficientes $\lambda_1, \ldots, \lambda_n$ tales que $\sum_{i=1}^n \lambda_i \boldsymbol{d}^i = \boldsymbol{0}$. Fijemos ahora $j \in \{1, \ldots, n\}$ y multipliquemos escalarmente ambos lados de esta igualdad por $\boldsymbol{Q}\boldsymbol{d}^j$. Entonces, usando que las direcciones son $\boldsymbol{Q}$-conjugadas obtenemos

$$(\sum_{i=1}^n \lambda_i \boldsymbol{d}^i)^\intercal \boldsymbol{Q}\boldsymbol{d}^j = \sum_{i=1}^n \lambda_i (\boldsymbol{d}^i)^\intercal \boldsymbol{Q}\boldsymbol{d}^j = \lambda_j (\boldsymbol{d}^j)^\intercal \boldsymbol{Q}\boldsymbol{d}^j = \boldsymbol{0}.$$

Como $\boldsymbol{Q}$ es definida positiva y $\boldsymbol{d}^j \neq \boldsymbol{0}$, tenemos que $\lambda_j = 0$. Como esto es cierto para todo $j \in \{1, \ldots, n\}$, obtenemos el resultado deseado. □

Consideremos ahora una función cuadrática de la forma $f(\boldsymbol{x}) = \boldsymbol{c}^\intercal \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^\intercal \boldsymbol{Q}\boldsymbol{x}$ y tomemos n direcciones $\boldsymbol{Q}$-conjugadas $\boldsymbol{d}^1, \ldots, \boldsymbol{d}^n$.[17] Nótese que la hessiana $\boldsymbol{H}(\boldsymbol{x})$ coincide, para todo $\boldsymbol{x}$, con la matriz $\boldsymbol{Q}$, con lo que el concepto de $\boldsymbol{Q}$-conjugación en problemas cuadráticos quiere decir direcciones conjugadas con respecto a la matriz hessiana. Supongamos ahora que queremos minimizar $f(\boldsymbol{x})$ y partimos de un punto inicial $\boldsymbol{x}^1$. Entonces, dado otro punto cualquiera $\boldsymbol{x} \in \mathbb{R}^n$, como las direcciones $\boldsymbol{d}^1, \ldots, \boldsymbol{d}^n$ forman una base tendremos que $\boldsymbol{x} - \boldsymbol{x}^1$ puede ser representado como combinación lineal de dichas direcciones. Es decir, existen $y_1, \ldots, y_n$ tales que

$$\boldsymbol{x} = \boldsymbol{x}^1 + \sum_{i=1}^n y_i \boldsymbol{d}^i.$$

Apoyándonos en este cambio de sistema de coordenadas podemos escribir $f(\boldsymbol{x})$ como función del vector $\boldsymbol{y}$ y obtenemos una nueva función F dada por

$$F(\boldsymbol{y}) = \boldsymbol{c}^\intercal (\boldsymbol{x}^1 + \sum_{i=1}^n y_i \boldsymbol{d}^i) + \frac{1}{2}(\boldsymbol{x}^1 + \sum_{i=1}^n y_i \boldsymbol{d}^i)^\intercal \boldsymbol{Q}(\boldsymbol{x}^1 + \sum_{i=1}^n y_i \boldsymbol{d}^i).$$

Ahora, usando que las direcciones $\boldsymbol{d}^1, \ldots, \boldsymbol{d}^n$ son $\boldsymbol{Q}$-conjugadas y que la matriz $\boldsymbol{Q}$ es simétrica, esta expresión se reduce a

$$F(\boldsymbol{y}) = \boldsymbol{c}^\intercal \boldsymbol{x}^1 + \sum_{i=1}^n y_i \boldsymbol{c}^\intercal \boldsymbol{d}^i + \frac{1}{2}(\boldsymbol{x}^1)^\intercal \boldsymbol{Q}\boldsymbol{x}^1 + \sum_{i=1}^n y_i (\boldsymbol{x}^1)^\intercal \boldsymbol{Q}\boldsymbol{d}^i + \frac{1}{2}\sum_{i=1}^n (y_i)^2 (\boldsymbol{d}^i)^\intercal \boldsymbol{Q}\boldsymbol{d}^i.$$

[17] A partir de aquí siempre que trabajemos con una función cuadrática asumiremos que $b = 0$, pues simplifica la notación y eliminar este término constante no afecta a la solución óptima.

Claramente, esta nueva función es separable y podría ser resuelta minimizando sucesivamente con respecto a las direcciones $\boldsymbol{e}^1, \ldots, \boldsymbol{e}^n$ (en el espacio de variables $\boldsymbol{y}$). Entonces, teniendo en cuenta que los términos constantes no afectan al proceso de optimización, la minimización en cada dirección sería equivalente a encontrar el valor de y_i que minimice

$$\boldsymbol{c}^\intercal \boldsymbol{x}^1 + y_i \boldsymbol{c}^\intercal \boldsymbol{d}^i + \frac{1}{2}(\boldsymbol{x}^1)^\intercal \boldsymbol{Q}\boldsymbol{x}^1 + y_i(\boldsymbol{x}^1)^\intercal \boldsymbol{Q}\boldsymbol{d}^i + \frac{1}{2}(y_i)^2(\boldsymbol{d}^i)^\intercal \boldsymbol{Q}\boldsymbol{d}^i =$$

o, equivalentemente,

$$\boldsymbol{c}^\intercal(\boldsymbol{x}^1 + y_i\boldsymbol{d}^i) + \frac{1}{2}(\boldsymbol{x}^1 + y_i\boldsymbol{d}^i)^\intercal \boldsymbol{Q}(\boldsymbol{x}^1 + y_i\boldsymbol{d}^i).$$

Esto se corresponde con la expresión de $f(\boldsymbol{x}^1 + y_i\boldsymbol{d}^i)$ con lo que minimizar la función F en la dirección $\boldsymbol{e}^i$ es equivalente a minimizar f desde $\boldsymbol{x}^1$ en la dirección $\boldsymbol{d}^i$.

Lo que acabamos de hacer es esbozar la propiedad fundamental de las direcciones conjugadas: permiten resolver problemas cuadráticos en espacios de dimensión n en n iteraciones. Antes de continuar presentamos un ejemplo en el que se pone en práctica el concepto de direcciones conjugadas.

Ejemplo 4.1. Considérese el problema

$$\text{minimizar } f(\boldsymbol{x}) = -12x_2 + 8x_1^2 + 8x_2^2 - 8x_1x_2 = -12x_2 + \frac{1}{2}\boldsymbol{x}^\intercal \begin{pmatrix} 16 & -8 \\ -8 & 16 \end{pmatrix} \boldsymbol{x}.$$

Veamos cómo generar direcciones conjugadas para este problema (la matriz $\boldsymbol{Q}$ es definida positiva y sus autovalores son 24 y 8). Supongamos que empezamos con el vector $(\boldsymbol{d}^1)^\intercal = (1, 0)$. Entonces, $\boldsymbol{d}^2$ ha de cumplir que $0 = (\boldsymbol{d}^1)^\intercal \boldsymbol{Q}\boldsymbol{d}^2 = 16d_1^2 - 8d_2^2$. Por tanto, tomando $(\boldsymbol{d}^2)^\intercal = (1, 2)$ tenemos un par de direcciones $\boldsymbol{Q}$-conjugadas.

•**Ejercicio 4.1.** Partiendo de la función f del Ejemplo 4.1 y de las direcciones $\boldsymbol{d}^1$ y $\boldsymbol{d}^2$ definidas en él, realiza las siguientes tareas:

- Elige un punto inicial cualquiera en $\mathbb{R}^2$ y minimiza sucesivamente la función f en las direcciones $\boldsymbol{d}^1$ y $\boldsymbol{d}^2$.
- Comprueba que el punto obtenido tras estos dos pasos es el óptimo global del problema.
- Repite el proceso eligiendo otro punto inicial distinto.
- Representa gráficamente el resultado obtenido.

Nótese que la minimización del problema original y de los problemas unidimensionales se puede hacer de modo sencillo aprovechándose de la diferenciabilidad y convexidad del problema; $\boldsymbol{H}(\boldsymbol{x}) = \boldsymbol{Q}$ es definida positiva. ◁

El hecho de que al trabajar con las direcciones conjugadas podamos pensar en el problema original como un problema separable, sugiere que un método iterativo que se va moviendo sucesivamente siguiendo estas direcciones tendrá la propiedad de que, en cada paso, el gradiente de la función en el punto actual será ortogonal a las direcciones conjugadas ya utilizadas (pues por la separabilidad en esas direcciones ya hemos mejorado todo lo que podríamos mejorar). Esto es precisamente lo que dice el siguiente resultado.

Proposición 4.12. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función cuadrática dada por $f(\boldsymbol{x}) = \boldsymbol{c}^\intercal \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^\intercal \boldsymbol{Q}\boldsymbol{x}$ con $\boldsymbol{Q}$ simétrica y definida positiva y sean $\boldsymbol{d}^1, \dots, \boldsymbol{d}^n$ direcciones conjugadas. Sea $\boldsymbol{x}^1$ un punto inicial arbitrario. Para $k \in \{1, \dots, n\}$, sea λ^k la solución óptima del problema $\min_{\lambda \in \mathbb{R}} f(\boldsymbol{x}^k + \lambda \boldsymbol{d}^k)$ y sea $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \lambda^k \boldsymbol{d}^k$. Entonces, para todo $k \in \{1, \dots, n\}$,*

$$\nabla f(\boldsymbol{x}^{k+1})^\intercal \boldsymbol{d}^i = 0, \text{ para todo } i \leq k.$$

A partir de esta proposición, dado que los vectores $\boldsymbol{d}^1, \dots, \boldsymbol{d}^n$ forman una base de $\mathbb{R}^n$, el siguiente resultado es inmediato.

Corolario 4.13. *En las hipótesis de la Proposición 4.12, el punto $\boldsymbol{x}^{n+1}$ es un mínimo global del problema de minimizar la función f.*

Es importante destacar que la convergencia al óptimo en n pasos es un resultado teórico y que rara vez se consigue de modo exacto en la práctica, pues las búsquedas de línea unidimensionales normalmente serán aproximadas, con lo que lo que tras n iteraciones tendremos una aproximación del óptimo.

••Ejercicio 4.2. Demuestra la Proposición 4.12 y el Corolario 4.13. ◁

Utilidad de las direcciones conjugadas y cálculo de las mismas

A estas alturas, es importante plantearse la siguiente pregunta: dado el problema de optimización

$$\text{minimizar } f(\boldsymbol{x}) = \boldsymbol{c}^\intercal \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^\intercal \boldsymbol{Q}\boldsymbol{x}$$

del enunciado de la Proposición 4.12, en el que se asume que $\boldsymbol{Q}$ es simétrica y definida positiva, ¿realmente necesitamos trabajar con direcciones conjugadas para resolverlo? Los supuestos sobre la matriz $\boldsymbol{Q}$ nos aseguran que tenemos un problema de optimización convexo y es fácil comprobar que $\nabla f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{c}$. Sabemos que en estos problemas $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ es una condición necesaria y suficiente de optimalidad, lo que en este caso se reduce a resolver el sistema lineal

$$\boldsymbol{Q}\boldsymbol{x} = -\boldsymbol{c}.$$

¿Por qué estamos entonces interesados en el uso de direcciones conjugadas? Las principales razones son las siguientes:

- En primer lugar, resolver un sistema de la forma $\boldsymbol{Q}\boldsymbol{x} = -\boldsymbol{c}$ puede ser muy costoso computacionalmente cuando la dimensión n es grande (invertir una matriz es algo computacionalmente caro). De hecho, algoritmos como el método de gradiente conjugado que presentamos en la siguiente sección surgieron originalmente para resolver este tipo de problemas, no para resolver problemas de optimización.
- Además, como ya comentamos, la idea de estos métodos se puede adaptar de modo casi inmediato para atacar problemas no cuadráticos, y los algoritmos resultantes deberían beneficiarse del buen comportamiento de las técnicas de direcciones conjugadas para problemas cuadráticos.

La siguiente pregunta que deberíamos plantearnos es la siguiente: ¿cómo generamos direcciones conjugadas? Si el cálculo de las direcciones conjugadas requiere el uso de la matriz $\boldsymbol{Q}$, o de la hessiana $H(\boldsymbol{x})$ cuando pasemos a problemas no cuadráticos, entonces el problema P-IV del método de Newton no desaparece: seguiríamos teniendo que hacer cálculos y manipulaciones explícitas con la hessiana en las distintas iteraciones del algoritmo.

En la siguiente sección presentamos un método que se basa en generar direcciones conjugadas sin ni siquiera conocer la matriz $\boldsymbol{Q}$ y llevando a cabo únicamente manipulaciones vectoriales.

Método de gradiente conjugado

En esta sección presentamos el método de gradiente conjugado, cuya idea fue introducida por Hestenes y Stiefel (1952) para resolver sistemas de ecuaciones lineales. Posteriormente, se empezó a aplicar en el contexto de optimización sin restricciones apoyándose en el hecho de que el óptimo de un problema cuadrático convexo coincide con la solución del sistema $\boldsymbol{Qx} = -\boldsymbol{c}$. Fueron Fletcher y Reeves (1964) quienes extendieron el método a la resolución de sistemas no lineales y de problemas de optimización sin restricciones. Aquí comentaremos la versión de Polak y Ribière (1969).

La clave del método está en la generación de direcciones conjugadas para problemas cuadráticos sin apoyarse explícitamente en la matriz $\boldsymbol{Q}$ y cuya generalización natural a problemas no cuadráticos no requerirá el uso de la matriz hessiana. Previamente, en la Figura 4.17 presentamos una versión del método de gradiente conjugado que obtiene las direcciones conjugadas basándose en la matriz $\boldsymbol{Q}$ y el método de ortogonalización de Gram-Schmidt.

MÉTODO DE GRADIENTE CONJUGADO (APOYÁNDOSE EN LA MATRIZ $\boldsymbol{Q}$)

INICIALIZACIÓN

Elegir $\varepsilon > 0$. Elegir un punto inicial $\boldsymbol{x}^1$. Definir $t = 1$.

PASO 1

- Si $\|\nabla f(\boldsymbol{x}^t)\| < \varepsilon$, FIN. Devolvemos $\boldsymbol{x}^t$.
- En otro caso, definir, mediante Gram-Schmidt, el vector $\boldsymbol{d}^t$ como la $\boldsymbol{Q}$-ortogonalización de $-\nabla f(\boldsymbol{x}^t)$ con respecto a las direcciones $\boldsymbol{d}^j$, con $j < t$:

$$\boldsymbol{d}^t = -\nabla f(\boldsymbol{x}^t) - \sum_{i=1}^{t-1} \frac{-\nabla f(\boldsymbol{x}^t)^\intercal \boldsymbol{Q}\boldsymbol{d}^i}{(\boldsymbol{d}^i)^\intercal \boldsymbol{Q}\boldsymbol{d}^i}\boldsymbol{d}^i.$$

Calcular λ^t, solución del problema $\text{minimizar}_{\lambda\in\mathbb{R}}\, f(\boldsymbol{x}^t + \lambda\boldsymbol{d}^t)$. Definir $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \lambda^t\boldsymbol{d}^t$. Reemplazar t por $t+1$. Repetir el PASO 1.

Figura 4.17: Esquema del método de gradiente conjugado apoyándose en la matriz $\boldsymbol{Q}$.

A continuación demostramos que las direcciones $\boldsymbol{d}^t$ obtenidas con el método descrito en la Figura 4.17 son $\boldsymbol{Q}$-ortogonales y, además, se pueden calcular de modo equivalente sin hacer uso de la matriz $\boldsymbol{Q}$. Una de las claves está en aprovecharse de que, para problemas cuadráticos, la matriz hessiana se puede reconstruir apoyándose en diferencias entre gradientes, más concretamente, dados dos puntos $\boldsymbol{x}^t$ y $\boldsymbol{x}^{t+1}$ tenemos que $\nabla f(\boldsymbol{x}^{t+1}) - \nabla f(\boldsymbol{x}^t) = (\boldsymbol{Q}\boldsymbol{x}^{t+1} + \boldsymbol{c}) - (\boldsymbol{Q}\boldsymbol{x}^t + \boldsymbol{c}) =$

$\boldsymbol{Q}(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t)$.

Proposición 4.14. *Considérense el método descrito en la Figura 4.17. Entonces se cumple lo siguiente:*

(i) *Las direcciones* $\boldsymbol{d}^1, \ldots, \boldsymbol{d}^n$ *son* $\boldsymbol{Q}$*-ortogonales.*

(ii) *Cada dirección distinta de cero en* $\{\boldsymbol{d}^1, \ldots, \boldsymbol{d}^n\}$ *es una dirección de descenso.*

(iii) *Las direcciones* $\boldsymbol{d}^1, \ldots, \boldsymbol{d}^n$ *se pueden calcular sin hacer uso de la matriz* $\boldsymbol{Q}$*.*

Demostración. (I) La $\boldsymbol{Q}$-ortogonalidad se obtiene por inducción. Para $t = 1$ no hay nada que demostrar, pues $\boldsymbol{d}^1$ es la única dirección disponible. Supongamos ahora que $\boldsymbol{d}^1, \ldots, \boldsymbol{d}^{t-1}$, con $t < n$, son $\boldsymbol{Q}$-ortogonales. Dado $j < t$, tenemos

$$(\boldsymbol{d}^t)^\intercal \boldsymbol{Q}\boldsymbol{d}^j = -\nabla f(\boldsymbol{x}^t)^\intercal \boldsymbol{Q}\boldsymbol{d}^j - (\sum_{i=1}^{t-1} \frac{-\nabla f(\boldsymbol{x}^t)^\intercal \boldsymbol{Q}\boldsymbol{d}^i}{(\boldsymbol{d}^i)^\intercal \boldsymbol{Q}\boldsymbol{d}^i}\boldsymbol{d}^i)^\intercal \boldsymbol{Q}\boldsymbol{d}^j$$

y, por la hipótesis de inducción, sólo el producto con $i = j$ será distinto de cero en el segundo sumando con lo que:

$$(\boldsymbol{d}^t)^\intercal \boldsymbol{Q}\boldsymbol{d}^j = -\nabla f(\boldsymbol{x}^t)^\intercal \boldsymbol{Q}\boldsymbol{d}^j - (\frac{-\nabla f(\boldsymbol{x}^t)^\intercal \boldsymbol{Q}\boldsymbol{d}^j}{(\boldsymbol{d}^j)^\intercal \boldsymbol{Q}\boldsymbol{d}^j}\boldsymbol{d}^j)^\intercal \boldsymbol{Q}\boldsymbol{d}^j = 0.$$

(II) Para ver que las direcciones $\boldsymbol{d}^j$ son direcciones de descenso basta notar que, por la Proposición 4.12, para todo $i < j$, $\nabla f(\boldsymbol{x}^j)^\intercal \boldsymbol{d}^i = 0$, con lo que

$$\nabla f(\boldsymbol{x}^j)^\intercal \boldsymbol{d}^j = -\nabla f(\boldsymbol{x}^j)^\intercal \nabla f(\boldsymbol{x}^j) - \nabla f(\boldsymbol{x}^j)^\intercal \sum_{i=1}^{j-1} \frac{-\nabla f(\boldsymbol{x}^j)^\intercal \boldsymbol{Q}\boldsymbol{d}^i}{(\boldsymbol{d}^i)^\intercal \boldsymbol{Q}\boldsymbol{d}^i}\boldsymbol{d}^i = -\|\nabla f(\boldsymbol{x}^j)\|^2,$$

con lo que si el algoritmo no paró en el iterante $\boldsymbol{x}^j$, tenemos $\nabla f(\boldsymbol{x}^j)^\intercal \boldsymbol{d}^j < 0$.

(III) Este paso lo haremos también por inducción. Para $t = 1$ tenemos $\boldsymbol{d}^1 = -\nabla f(\boldsymbol{x}^1)$, cuyo cálculo no requiere usar la matriz $\boldsymbol{Q}$. Supongamos entonces que ya tenemos direcciones $\boldsymbol{Q}$-ortogonales $\boldsymbol{d}^1, \ldots, \boldsymbol{d}^{t-1}$ y que han sido calculadas sin usar $\boldsymbol{Q}$. Veamos cómo calcular $\boldsymbol{d}^t$.

Dado, $j \leq t$, denotemos por D^j al espacio vectorial generado por las direcciones $\boldsymbol{d}^1, \ldots, \boldsymbol{d}^j$. Por la Proposición 4.12, para todo $j < t$, $\nabla f(\boldsymbol{x}^t)^\intercal \boldsymbol{d}^j = 0$, con lo que $\nabla f(\boldsymbol{x}^t)$ es perpendicular a cualquier vector de D^j. Además, como las direcciones $\boldsymbol{d}^1, \ldots, \boldsymbol{d}^j$ son linealmente independientes y cada dirección $\boldsymbol{d}^j$ se puede expresar como combinación lineal de los gradientes $\nabla f(\boldsymbol{x}^i)$, con $i \leq j$, tenemos que $\nabla f(\boldsymbol{x}^t)$ también ha de ser ortogonal a todos los gradientes anteriores. En particular tenemos que, para todo $j \in \{2, \ldots, t-1\}$,

$$\nabla f(\boldsymbol{x}^t)^\intercal (\nabla f(\boldsymbol{x}^j) - \nabla f(\boldsymbol{x}^{j-1})) = 0.$$

Usando la expresión del gradiente de una función cuadrática tenemos

$$\nabla f(\boldsymbol{x}^j) - \nabla f(\boldsymbol{x}^{j-1}) = \boldsymbol{Q}(\boldsymbol{x}^j - \boldsymbol{x}^{j-1}) = \boldsymbol{Q}(\lambda^{j-1}\boldsymbol{d}^{j-1}) = \lambda^{j-1}\boldsymbol{Q}\boldsymbol{d}^{j-1}. \tag{4.1}$$

Por (II) sabemos que $\boldsymbol{d}^{j-1}$ es una dirección de descenso en $\boldsymbol{x}^{j-1}$ y, como tenemos un problema de optimización convexa, esto implica que $\lambda^{j-1} > 0$. Combinando ahora las dos ecuaciones anteriores tenemos que, para todo $j \in \{2, \dots, t-1\}$, $\nabla f(\boldsymbol{x}^t)^\intercal \boldsymbol{Q}\boldsymbol{d}^{j-1} = 0$. Entonces, si vamos ahora a la expresión usada para definir $\boldsymbol{d}^t$ en la Figura 4.17 tenemos

$$\boldsymbol{d}^t = -\nabla f(\boldsymbol{x}^t) - \sum_{i=1}^{t-1} \frac{-\nabla f(\boldsymbol{x}^t)^\intercal \boldsymbol{Q}\boldsymbol{d}^i}{(\boldsymbol{d}^i)^\intercal \boldsymbol{Q}\boldsymbol{d}^i}\boldsymbol{d}^i = -\nabla f(\boldsymbol{x}^t) + \alpha^t \boldsymbol{d}^{t-1}, \text{ con } \alpha^t = \frac{\nabla f(\boldsymbol{x}^t)^\intercal \boldsymbol{Q}\boldsymbol{d}^{t-1}}{(\boldsymbol{d}^{t-1})^\intercal \boldsymbol{Q}\boldsymbol{d}^{t-1}}.$$

Hemos simplificado el cálculo, pero α^t todavía depende de $\boldsymbol{Q}$. Usando ahora la Ecuación (4.1) y que $\lambda^{t-1} \neq 0$ (por ser $\boldsymbol{d}^{t-1}$ una dirección de descenso), tenemos $\boldsymbol{Q}\boldsymbol{d}^{t-1} = (\nabla f(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^{t-1}))/\lambda^{t-1}$. Entonces,

$$\alpha^t = \frac{\nabla f(\boldsymbol{x}^t)^\intercal (\nabla f(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^{t-1}))}{(\boldsymbol{d}^{t-1})^\intercal (\nabla f(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^{t-1}))},$$

con lo que α^t no depende de $\boldsymbol{Q}$ y $\boldsymbol{d}^t = -\nabla f(\boldsymbol{x}^t) + \alpha^t \boldsymbol{d}^{t-1}$ se puede calcular sin conocer $\boldsymbol{Q}$. □

El cálculo del coeficiente α^t en la demostración que acabamos de ver se puede afinar todavía un poco más. Usando que $\nabla f(\boldsymbol{x}^t)$ es ortogonal a $\boldsymbol{d}^{t-1}$ y que $\nabla f(\boldsymbol{x}^{t-1})$ es ortogonal a todos los vectores usados en la definición de $\boldsymbol{d}^{t-1}$ excepto el propio $\nabla f(\boldsymbol{x}^{t-1})$, tenemos

$$(\boldsymbol{d}^{t-1})^\intercal (\nabla f(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^{t-1})) = -(\boldsymbol{d}^{t-1})^\intercal \nabla f(\boldsymbol{x}^{t-1}) = \|\nabla f(\boldsymbol{x}^{t-1})\|^2,$$

de donde obtenemos la expresión de α^t de Polak y Ribière:

$$\alpha^t = \frac{\nabla f(\boldsymbol{x}^t)^\intercal (\nabla f(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^{t-1}))}{\|\nabla f(\boldsymbol{x}^{t-1})\|^2}.$$

Podíamos ir todavía un paso más allá, y apoyarnos en que $\nabla f(\boldsymbol{x}^t)$ es ortogonal a $\nabla f(\boldsymbol{x}^{t-1})$ para obtener la fórmula de Fletcher y Reeves:

$$\alpha^t = \frac{\|\nabla f(\boldsymbol{x}^t)\|^2}{\|\nabla f(\boldsymbol{x}^{t-1})\|^2}.$$

Aunque ambas formulaciones son equivalentes matemáticamente, en la práctica la expresión de Polak y Ribière parece más robusta a los errores numéricos de las búsquedas de línea. Además, también la hace preferible para problemas no cuadráticos el hecho de que la ortogonalidad de $\nabla f(\boldsymbol{x}^t)$ y $\nabla f(\boldsymbol{x}^{t-1})$ no tiene por qué ser cierta para los mismos.

Corolario 4.15. *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función cuadrática dada por $f(\boldsymbol{x}) = \boldsymbol{c}^\intercal \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^\intercal \boldsymbol{Q}\boldsymbol{x}$ con $\boldsymbol{Q}$ simétrica y definida positiva. Entonces, el método de gradiente conjugado encuentra la solución óptima del problema de minimización asociado en, como mucho, n iteraciones.*

Demostración. Inmediato sin más que combinar el Corolario 4.13 con la Proposición 4.14. □

En la Figura 4.18 tenemos la representación del método de gradiente conjugado de Polak y Ribière que puede aplicarse directamente a problemas no cuadráticos. En la práctica, este método se suele implementar reiniciando el cálculo del gradiente conjugado cada cierto número

de iteraciones, digamos t, tomando $\boldsymbol{d}^t = -\nabla f(\boldsymbol{x}^t)$. La razón es que la propiedad de conjugación de las direcciones construidas se puede ir perdiendo debido a problemas de precisión numérica en las búsquedas de línea realizadas iteración a iteración. Se ha observado que es efectivo reiniciar cada 10 o 20 iteraciones, independientemente de la dimensión n del problema. Esto permite combinar de forma eficiente la robustez del método de máximo descenso lejos del óptimo con la eficacia del método de gradiente conjugado cerca del mismo.

MÉTODO DE GRADIENTE CONJUGADO (SIN APOYARSE EN LA MATRIZ $\boldsymbol{Q}$)

INICIALIZACIÓN

Elegir $\varepsilon > 0$. Elegir un punto inicial $\boldsymbol{x}^1$. Definir $t = 1$.

PASO 1

- Si $\|\nabla f(\boldsymbol{x}^t)\| < \varepsilon$, FIN. Devolvemos $\boldsymbol{x}^t$.
- En otro caso, definir $\boldsymbol{d}^t = -\nabla f(\boldsymbol{x}^t) + \alpha^t \boldsymbol{d}^{t-1}$. Donde $\alpha^1 = 0$ y, si $t > 1$,

$$\alpha^t = \frac{\nabla f(\boldsymbol{x}^t)^\intercal \big(\nabla f(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^{t-1})\big)}{\|\nabla f(\boldsymbol{x}^{t-1})\|^2}.$$

Calcular λ^t, solución del problema minimizar$_{\lambda \geq 0}\, f(\boldsymbol{x}^t + \lambda \boldsymbol{d}^t)$. Definir $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \lambda^t \boldsymbol{d}^t$. Reemplazar t por $t+1$. Repetir el PASO 1.

Figura 4.18: Esquema del método de gradiente conjugado sin apoyarse en la matriz $\boldsymbol{Q}$.

Proposición 4.16. *Supongamos que se cumplen los siguientes supuestos:*

- *La función f es continuamente diferenciable.*
- *La sucesión de puntos generada por el algoritmo está contenida en un conjunto compacto.*

Entonces, todo punto de acumulación $\bar{\boldsymbol{x}}$ de la sucesión generada por el método de gradiente conjugado cumple $\nabla f(\bar{\boldsymbol{x}}) = \boldsymbol{0}$.

Además, si la sucesión es convergente a un punto $\bar{\boldsymbol{x}}$ y f es dos veces continuamente diferenciable en un entorno de $\bar{\boldsymbol{x}}$ y con $H(\bar{\boldsymbol{x}})$ definida positiva, entonces el orden de convergencia es n-superlineal.[18]

El resultado relativo a la velocidad de convergencia no es muy relevante, pues en la práctica reinicia el método introduciendo un paso de máximo descenso cada pocas iteraciones.

Terminamos este apartado con una pequeña discusión de las principales fortalezas y debilidades del método de gradiente conjugado:

- Una de las grandes ventajas del método de gradiente conjugado es la sencillez de su implementación y el hecho de que es muy ligero computacionalmente: sólo hace manipulaciones vectoriales y, además, no tiene unos grandes requerimientos de memoria (sólo hay que guardar el gradiente de la iteración anterior).

[18] Formalmente, la sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ tiene convergencia n-superlineal si la sucesión $\{\boldsymbol{y}^t\}_{t\in\mathbb{N}}$, donde $\boldsymbol{y}^t = \boldsymbol{x}^{nt}$, tiene convergencia superlineal. Con algún supuesto adicional se puede probar que la convergencia es n-cuadrática.

- Una de las principales debilidades se deriva justamente del punto precedente. Como en cada iteración sólo utiliza información del gradiente de la iteración anterior para desviar el gradiente actual, la calidad de dicha desviación está limitada con respecto a lo que se puede conseguir con los métodos cuasi-Newton que discutiremos en el siguiente apartado, y que usan mucha más información en cada paso. Sin embargo, esta información adicional pasa por tener que manipular matrices, lo que tiene asociado un mayor coste computacional.
- Por último, el método es sensible al condicionamiento de la matriz hessiana en el límite, aunque mucho menos que el método de gradiente. Hay versiones del método que llevan a cabo *precondicionamientos*, dirigidos a atenuar los problemas derivados de un mal condicionamiento.

Métodos cuasi-Newton

Los métodos cuasi-Newton deben su nombre a su similitud con el método de Newton y, desde el punto de vista matemático, se pueden ver como métodos de direcciones conjugadas. Se trata de métodos que en cada iteración desvían el gradiente multiplicándolo por una matriz, y no usando únicamente la información del gradiente de la iteración anterior como hace el método de gradiente conjugado. Más concretamente, esta familia de métodos se basa en calcular $\boldsymbol{x}^{t+1}$ mediante una búsqueda lineal desde $\boldsymbol{x}^t$ en una dirección de la forma

$$\boldsymbol{d}^t = -\boldsymbol{V}^t \nabla f(\boldsymbol{x}^t),$$

donde las matrices $\boldsymbol{V}^t$ son matrices simétricas, definidas positivas y, a ser posible, buenas aproximaciones de la inversa de la matriz hessiana (y de ahí el nombre de métodos cuasi-Newton. Además, en el caso de problemas cuadráticos las direcciones construidas por el algoritmo deberán ser direcciones conjugadas. Hay múltiples formas de definir las matrices $\boldsymbol{V}^t$, y aquí presentamos una formulación general introducida por Broyden (1967) y que engloba a algunas de las versiones más habituales. Previamente necesitamos un poco de notación. Definimos, $\boldsymbol{p}^t = \boldsymbol{x}^{t+1} - \boldsymbol{x}^t$ y $\boldsymbol{q}^t = \nabla f(\boldsymbol{x}^{t+1}) - \nabla f(\boldsymbol{x}^t)$. Entonces, podemos definir una familia de métodos cuasi-Newton basados en el cálculo

$$\boldsymbol{V}^{t+1} = \boldsymbol{V}^t + \frac{\boldsymbol{p}^t(\boldsymbol{p}^t)^\intercal}{(\boldsymbol{p}^t)^\intercal \boldsymbol{q}^t} - \frac{\boldsymbol{V}^t\boldsymbol{q}^t(\boldsymbol{q}^t)^\intercal\boldsymbol{V}^t}{(\boldsymbol{q}^t)^\intercal\boldsymbol{V}^t\boldsymbol{q}^t} + \omega^t((\boldsymbol{q}^t)^\intercal\boldsymbol{V}^t\boldsymbol{q}^t)\boldsymbol{v}^t(\boldsymbol{v}^t)^\intercal,$$

donde $\boldsymbol{V}^1$ es una matriz simétrica y definida positiva,

$$\boldsymbol{v}^t = \frac{\boldsymbol{p}^t}{(\boldsymbol{p}^t)^\intercal\boldsymbol{q}^t} - \frac{\boldsymbol{V}^t\boldsymbol{q}^t}{(\boldsymbol{q}^t)^\intercal\boldsymbol{V}^t\boldsymbol{q}^t} \quad \text{y } \omega^t \in [0,1] \text{ es un parámetro a elegir.}$$

No vamos a entrar aquí en la justificación de esta fórmula, tarea que sería más sencilla estudiando primero alguno de sus casos más representativos:

$\boldsymbol{\omega^t = 0}$. El método DFP, propuesto por Davidon (1959) y desarrollado posteriormente por Fletcher y Powell (1963).

$\boldsymbol{\omega^t = 1}$. El método BFGS, propuesto independientemente por Broyden (1970), Fletcher (1970), Goldfarb (1970) y Shanno (1970).

Se puede probar teóricamente que estos métodos generan matrices $\boldsymbol{V}^t$ que son definidas positivas y por tanto las direcciones $\boldsymbol{d}^t = -\boldsymbol{V}^t \nabla f(\boldsymbol{x}^t)$ son direcciones de descenso. Además, también se puede probar que, acompañados de una búsqueda de línea exacta, todos los métodos de esta familia son matemáticamente equivalentes para funciones continuamente diferenciables. Más concretamente, en cada iteración t, todos obtienen una dirección de búsqueda $\boldsymbol{d}^t$ que es paralela al vector $\boldsymbol{v}^t$. Sin embargo, en la práctica las búsquedas de línea no son exactas y el método BFGS suele exhibir un rendimiento superior. Además, también se puede probar que, en el caso de problemas cuadráticos, son también matemáticamente equivalentes al método de gradiente conjugado cuando este se aplica con un precondicionamiento dado por la matriz $\boldsymbol{V}^1$.

Terminamos este apartado con una pequeña discusión de las principales fortalezas y debilidades de los métodos cuasi-Newton:

- Los métodos cuasi-Newton aseguran una velocidad de convergencia cuadrática.
- En la práctica se ha observado que los métodos cuasi-Newton son menos sensibles a imprecisiones en las búsquedas de línea que el método de gradiente conjugado.
- Sin embargo, cuando la dimensión es muy alta las manipulaciones matriciales de estos métodos se vuelven muy costosas. En este sentido es importante destacar que hay versiones de "memoria limitada" de los métodos cuasi-Newton que buscan un equilibrio entre la sofisticación en el cálculo de los $\boldsymbol{V}^t$ y la ligereza del método de gradiente conjugado.
- En la práctica se ha observado que los métodos cuasi-Newton pueden generar matrices intermedias con problemas de condicionamiento, lo que puede afectar a su estabilidad numérica. Para evitar estos problemas, hay versiones de estos métodos que introducen un factor de escalado en los cálculos, aunque una discusión detallada de este procedimiento va más allá del alcance de estas notas.

4.8.4 Métodos de región de confianza

Para terminar presentamos una última familia de métodos de optimización sin restricciones, cuya filosofía se usa habitualmente en problemas de optimización con restricciones. La idea es la siguiente, queremos resolver el problema $\text{minimizar}_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$. Pensemos en los métodos ya vistos hasta ahora, en particular, en el método de Newton, el de Levenberg-Marquardt o algún método cuasi-Newton. Una forma de ver estos métodos es pensar que, en cada iteración t, trabajan con una aproximación cuadrática de f en torno al iterante $\boldsymbol{x}^t$:

$$q^t(\boldsymbol{x}) = f(\boldsymbol{x}^t) + \nabla f(\boldsymbol{x}^t)(\boldsymbol{x} - \boldsymbol{x}^t) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^t)^\intercal \boldsymbol{Q}^t (\boldsymbol{x} - \boldsymbol{x}^t)^\intercal,$$

donde $\boldsymbol{Q}^t$ coincide con la hessiana $\boldsymbol{H}(\boldsymbol{x}^t)$ en el método de Newton y entonces tenemos la aproximación de Taylor de segundo orden, o puede ser otra matriz que pretenda aproximar dicha matriz hessiana. En el caso de que $\boldsymbol{Q}^t$ sea una matriz definida positiva el mínimo de q^t se alcanza en el punto $\boldsymbol{x}^t - \boldsymbol{Q}^{-1} \nabla f(\boldsymbol{x}^t)$, que se correspondería con un paso de longitud 1

en la dirección $\boldsymbol{d}^t = -\boldsymbol{Q}^{-1}\nabla f(\boldsymbol{x}^t)$. Exceptuando el método de Newton, el resto de métodos que hemos comentado realizarían una búsqueda de línea en la dirección $\boldsymbol{d}^t$ (y también las implementaciones del método de Newton en la práctica).

La idea de los métodos de región de confianza es reemplazar esta búsqueda de línea por una búsqueda en cualquier dirección, pero dentro de un cierto entorno del punto $\boldsymbol{x}^t$. Lo más natural es partir de un cierto radio r^t y restringir la búsqueda a la bola B^t, de centro $\boldsymbol{x}^t$ y radio r^t:

$$B^t = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{x}^t\| \leq r^t\}.$$

El conjunto B^t se denomina *región de confianza* y en la iteración t buscamos resolver el subproblema $\text{minimizar}_{\boldsymbol{x}\in B^t}\, q^t(\boldsymbol{x})$. El óptimo de este problema será el nuevo iterante $\boldsymbol{x}^{t+1}$ a partir del cual definiremos una nueva aproximación q^{t+1} y así sucesivamente. Los subproblemas $\text{minimizar}_{\boldsymbol{x}\in B^t}\, q^t(\boldsymbol{x})$ no son búsquedas de línea, pero son problemas cuadráticos definidos sobre un dominio acotado y si las matrices $\boldsymbol{Q}^t$ son definidas positivas su resolución no es complicada (basta moverse desde $\boldsymbol{x}^t$ en la dirección de máximo descenso hasta alcanzar la frontera de B^t).

Por otro lado, es importante definir el radio r^t de forma adecuada, de tal manera que podamos dar pasos relativamente grandes y, al mismo tiempo, no nos salgamos de la región en la que q^t es una buena aproximación de la función f. Veamos el enfoque habitual para conseguir esto. Supongamos que $\boldsymbol{y}^t$ es el óptimo del subproblema de la iteración t, lo que queremos es evaluar cómo de bueno es el valor $f(\boldsymbol{y}^t)$ en comparación con el valor de la aproximación cuadrática $q^t(\boldsymbol{y}^t)$. Como la diferencia absoluta entre $f(\boldsymbol{y}^t)$ y $q^t(\boldsymbol{y}^t)$ no es muy informativa, lo que se suele hacer es comparar la mejora predicha por el modelo cuadrático con la mejora real obtenida. Más concretamente, dado $\gamma \in (0, 1)$, se evalúa si

$$f(\boldsymbol{y}^t) \leq f(\boldsymbol{x}^t) - \gamma(q^t(\boldsymbol{x}^t) - q^t(\boldsymbol{y}^t)).$$

Nótese que $q^t(\boldsymbol{x}^t) - q^t(\boldsymbol{y}^t)$ va a ser siempre no negativo, con lo que si γ vale por ejemplo 0.5 estaríamos pidiendo que la mejora real que se produce en la función f al pasar de $\boldsymbol{x}^t$ a $\boldsymbol{y}^t$ tiene que ser al menos 0.5 veces la mejora obtenida en el modelo cuadrático. En caso afirmativo, se considera que la aproximación es buena y se realiza un *paso de descenso* definiendo $\boldsymbol{x}^{t+1} = \boldsymbol{y}^t$, además, si la desigualdad se cumple para un valor $\bar{\gamma}$ suficientemente próximo a 1, incrementamos el radio de la región de confianza definiendo $r^{t+1} = \beta_1 r^t$ con $\beta_1 \geq 1$. En caso contrario, el modelo no es lo suficientemente bueno y hay que reducir la región de confianza, definiendo $r^{t+1} = \beta_2 r^t$, con $\beta_2 \in (0, 1)$. A este paso se le llama *paso nulo*, pues pasamos a la iteración siguiente con $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t$. El esquema general de un algoritmo de región de confianza está representado en la Figura 4.19. Para profundizar en este tipo de algoritmos el lector puede acudir tanto a la referencia principal para estas notas, Bazaraa y otros (2006), como a las siguientes referencias, especializadas en métodos de región de confianza: Conn y otros (2000) y Powell (2003).

4.9 Optimización multidimensional sin diferenciabilidad

Hasta ahora hemos visto métodos de optimización multidimensional tanto con derivadas como sin ellas, pero todos los resultados de convergencia de los mismos requerían diferenciabilidad. Para terminar la exposición de métodos de optimización sin restricciones comentaremos brevemente la idea del método de subgradiente, cuya convergencia se puede garantizar incluso en ausencia de diferenciabilidad.

ESQUEMA GENERAL DE UN MÉTODO DE REGIÓN DE CONFIANZA

INICIALIZACIÓN
Elegir $\varepsilon > 0$. Elegir $r^1 > 0$, γ y $\bar{\gamma}$ con $0 \leq \gamma \leq \bar{\gamma} \leq 1$, y $\beta_1 \geq 1$ y $\beta_2 \in (0, 1/\beta_1)$. Elegir un punto inicial $\boldsymbol{x}^1$, calcular $\nabla f(\boldsymbol{x}^1)$ y $\boldsymbol{Q}^1$, y con ellos definir q^1. Definir $t = 1$.

PASO 1
Calcular $\boldsymbol{y}^t$, solución del problema minimizar$_{\boldsymbol{x} \in B^t}$ $q^t(\boldsymbol{x}^t)$.

- Si $|f(\boldsymbol{x}^t) - q^t(\boldsymbol{y}^t)| < \varepsilon$, FIN. Devolvemos $\boldsymbol{y}^t$.
- En otro caso:
  - Si $f(\boldsymbol{y}^t) \leq f(\boldsymbol{x}^t) - \gamma\big(q^t(\boldsymbol{x}^t) - q^t(\boldsymbol{y}^t)\big)$, ir al PASO 2.
  - Si $f(\boldsymbol{y}^t) > f(\boldsymbol{x}^t) - \gamma\big(q^t(\boldsymbol{x}^t) - q^t(\boldsymbol{y}^t)\big)$, ir al PASO 3.

PASO 2 (PASO DE DESCENSO)
Si $f(\boldsymbol{y}^t) \leq f(\boldsymbol{x}^t) - \bar{\gamma}\big(q^t(\boldsymbol{x}^t) - q^t(\boldsymbol{y}^t)\big)$, definir $r^{t+1} = \beta_1 r^t$.

Definir $\boldsymbol{x}^{t+1} = \boldsymbol{y}^t$. Calcular $\nabla f(\boldsymbol{x}^{t+1})$ y $\boldsymbol{Q}^{t+1}$, y con ellos definir q^{t+1}. Reemplazar t por $t+1$ e ir al PASO 1.

PASO 3 (PASO NULO)
Definir $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t$, $r^{t+1} = \beta_2 r^t$. Reemplazar t por $t+1$ e ir al PASO 1.

Figura 4.19: Esquema de un método de región de confianza.

4.9.1 Método de subgradiente

En este apartado vamos a comentar un método de optimización para problemas de optimización convexa no diferenciables basado en el uso de subgradientes; algunas referencias clásicas de este método son Polyak (1967, 1969) y Held y otros (1974). Este método se puede ver como una generalización directa del método de máximo descenso, pero donde el gradiente se reemplaza por un subgradiente. Sin embargo, a diferencia de lo que ocurría con el método de máximo descenso, cuando se trabaja con un subgradiente no tenemos asegurado que la dirección opuesta al mismo sea efectivamente una dirección de descenso. Sin embargo, si en cada iteración nos movemos en pasos suficientemente pequeños, se puede garantizar que el algoritmo converge a una solución óptima (con lo que no se realizarán búsquedas de línea).

El esquema más básico de un algoritmo de subgradiente se puede ver en la Figura 4.20. Nótese que el criterio de parada $\|\boldsymbol{s}^t\| < \varepsilon$ podría no cumplirse nunca pues, aunque encontremos una solución para la cual el vector $\mathbf{0}$ pertenece al subdiferencial, no tenemos garantizado que el método vaya a escoger justamente dicho subgradiente. Es por ello que hay que usar otros criterios de parada adicionales basados en el número de iteraciones o valores de la cota superior (esto sería habitual usando esquemas de optimización primal-dual, donde ambos problemas se van aportando sucesivamente cotas el uno al otro).

El resultado formal detrás del método de subgradiente es el siguiente.

Proposición 4.17. *Supongamos que la función f es convexa y no diferenciable. Si la sucesión $\{\lambda^t\}$ es no negativa, converge a 0 y, además, cumple que $\sum_{t=0}^{\infty} \lambda^t = \infty$, entonces:*

- *O bien el algoritmo termina en una cantidad finita de iteraciones (por el criterio sobre $\|\boldsymbol{s}^t\|$).*

MÉTODO (RUDIMENTARIO) DE SUBGRADIENTE

INICIALIZACIÓN

Elegir $\varepsilon > 0$. Elegir un punto inicial $\boldsymbol{x}^1$. Definir la cota superior inicial $UB^1 = f(\boldsymbol{x}^1)$ y $\bar{\boldsymbol{x}} = \boldsymbol{x}^1$. Definir $t = 1$.

PASO 1

Elegir un subgradiente $\boldsymbol{s}^t$ del subdiferencial de f en $\boldsymbol{x}^t$. Ir al PASO 2.

PASO 2

- Si $\|\boldsymbol{s}^t\| < \varepsilon$, FIN. Devolvemos $\boldsymbol{x}^t$.
- En otro caso, definir $\boldsymbol{d}^t = -\frac{\boldsymbol{s}^t}{\|\boldsymbol{s}^t\|}$. Determinar el paso λ^t y definir $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \lambda^t \boldsymbol{d}^t$. Si $f(\boldsymbol{x}^{t+1}) < UB^t$, definir $UB^{t+1} = f(\boldsymbol{x}^{t+1})$ y $\bar{\boldsymbol{x}} = \boldsymbol{x}^{t+1}$. Reemplazar t por $t+1$. Repetir el PASO 1.

Figura 4.20: Esquema del método de máximo descenso.

- *O el algoritmo genera una sucesión infinita tal que UB^t converge al mínimo de f (con el punto $\bar{\boldsymbol{x}}$ siendo una solución óptima).*

Intuitivamente, la condición sobre los λ^t garantiza que, cuando cometamos un "error" moviéndonos en una dirección que no es de descenso, dicho error tendrá magnitud λ^t, y podrá ser rectificado en las restantes iteraciones. Más allá de esta intuición, el resultado se apoya fuertemente en la convexidad de f y las propiedades de los subgradientes.

A la vista de la Proposición 4.17, podría parecer que la sucesión dada por $\lambda^t = \frac{1}{t}$ es una opción razonable. Sin embargo, tiene mal comportamiento en la práctica y hay que tener en cuenta otras consideraciones a la hora de elegir los λ^t.

Además de la cuestión de cómo el elegir en cada iteración un subgradiente y un paso adecuados, este algoritmo sufre también de otros problemas que hacen adecuado trabajar con modificaciones en el espíritu del método de gradiente conjugado, desviando el subgradiente para evitar problemas derivados de que ciertos ángulos se aproximen demasiado a 90°. En cualquier caso, todas estas consideraciones requerirían un análisis bastante profundo, que va más allá de estas notas.

Aunque existen otros algoritmos para el caso de funciones convexas no diferenciables, una gran ventaja del método de subgradiente es que se puede adaptar de modo sencillo a problemas de optimización con restricciones (siempre que sean problemas de programación convexa). En este caso, en cada iteración bastaría definir $\boldsymbol{x}^{t+1}$ como la proyección de $\boldsymbol{x}^t + \lambda^t \boldsymbol{d}^t$ sobre el conjunto factible.

Por citar otros métodos también basados en el uso de subgradientes tendríamos, por ejemplo, métodos de planos de corte y bundle methods.[19]

[19] Quien desee profundizar en estos algoritmos puede referirse al libro "Nonlinear Optimization" (Andrzej Ruszczyński).

4.10 Ejemplos ilustrativos

En esta sección presentamos algunos ejemplos ilustrativos de las ideas de los algoritmos que hemos discutido en este tema, así como de las implicaciones en el rendimiento de los mismos.

Ejemplo 4.2. En la Sección 4.8.1 presentamos el método de máximo descenso y la Proposición 4.8 asegura la convergencia del mismo a un punto con gradiente igual a cero en el caso de que la función f sea continuamente diferenciable. En este ejemplo pretendemos ilustrar que el supuesto de diferenciabilidad es crucial.

Considérese la función $f : \mathbb{R}^2 \to \mathbb{R}$ definida por partes como sigue:

$$f(\boldsymbol{x}) = \begin{cases} \sqrt{x_1^2 + 3x_2^2} & \text{si } |x_2| < x_1 \\ \dfrac{x_1 + 3|x_2|}{2} & \text{si } |x_2| \geq x_1. \end{cases}$$

Estamos ante una función continua y diferenciable prácticamente en todo su dominio, aunque habiendo problemas en los puntos de la forma $|x_2| = x_1$, siendo especialmente problemático el punto $(0, 0)$.

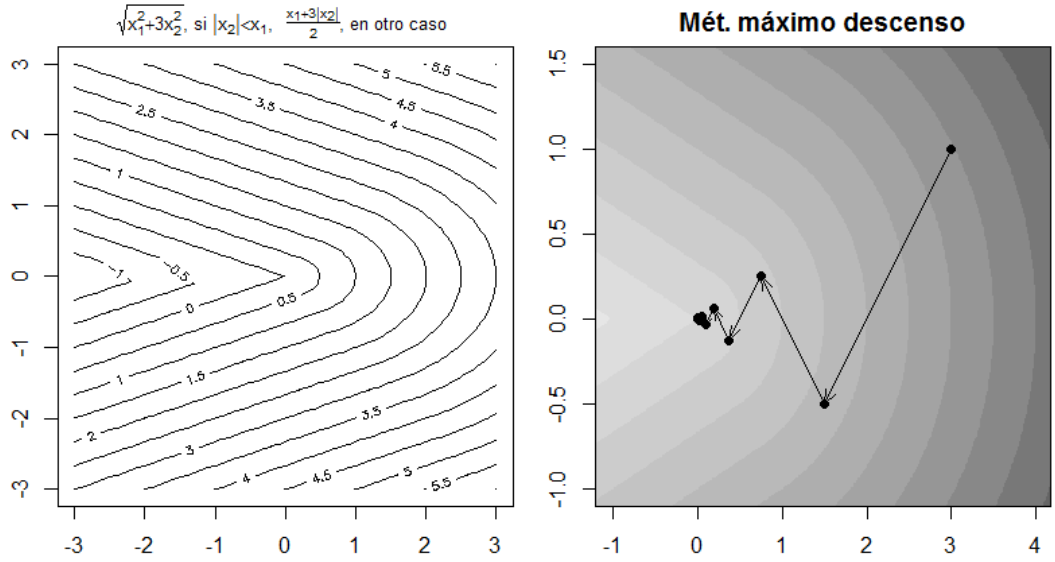


Figura 4.21: Ejemplo en el que el método de descenso por coordenadas converge a un punto que está lejos de ser un mínimo de la función.

En la Figura 4.21 vemos el comportamiento del algoritmo de máximo descenso cuando este parte del punto $\boldsymbol{x}^1 = (3, 1)$:

Primera iteración. Como $\boldsymbol{x}^1 = (3, 1)$, $|x_2^1| < x_1^1$ y el gradiente de la función en $\boldsymbol{x}^1$ sigue la dirección $(x_1^1, 3x_2^1) = (3, 3)$. Y tenemos que la dirección de máximo descenso es $(-3, -3)$. Como $x_1^1 = 3x_2^1$, no es difícil ver que $\boldsymbol{x}^2 = \boldsymbol{x}^1 - (x_1^1/2, x_1^1/2) = (x_1^1/2, -x_2^1/2) = (1.5, -0.5)$.

Segunda iteración. Ahora $\boldsymbol{x}^2 = (1.5, -0.5)$ y, nuevamente, $|x_2^2| < x_1^2$ y el gradiente sigue la dirección $(x_1^1, 3x_2^1) = (1.5, -1.5)$, que es ortogonal a la de la iteración anterior y dará lugar al punto $\boldsymbol{x}^3 = \boldsymbol{x}^2 - (x_1^2/2, -x_1^2/2) = (x_1^2/2, -x_2^2/2) = (0.75, 0.25)$.

Resto de iteraciones. En general tendremos que $\boldsymbol{x}^t = (\frac{x_1^1}{2^{t-1}}, \frac{\boldsymbol{x}_2^1}{(-2)^{t-1}}) = (\frac{3}{2^{t-1}}, \frac{1}{(-2)^{t-1}})$ y siempre se cumplirá que $|x_2^1| < x_1^1$.

Por tanto, el algoritmo siempre se moverá en una zona en la que f viene dada por $\sqrt{x_1^2 + 3x_2^2}$ y convergerá al mínimo de esta función, que es el punto $(0, 0)$ y el valor de la función en este punto es 0. Sin embargo, la función f está definida a trozos y está claro que no tiene un mínimo finito, pues en puntos de la forma $(-2t, 0)$ con $t \in \mathbb{N}$ tenemos que $f(-2t, 0) = -t$. Es importante destacar que el algoritmo funciona mal aunque todos los iterantes construidos son puntos donde la función es continuamente diferenciable. El problema radica en que precisamente el punto límite es el que no lo es. ◇

Ejemplo 4.3. Ahora presentamos un ejemplo de función continuamente diferenciable con múltiples puntos estacionarios, $\nabla f(\boldsymbol{x}) = \mathbf{0}$, y varios óptimos globales. Se trata de la función $f : \mathbb{R}^2 \to \mathbb{R}$ definida como $f(\boldsymbol{x}) = (x_1 - x_2^3)^2 + 3(x_1 - x_2)^4$. Esta función tiene cinco puntos estacionarios, con distinto comportamiento en lo que respecta a la optimalidad:

- Los puntos $(0, 0)$, $(1, 1)$ y $(-1, -1)$ son los tres únicos óptimos globales, en los que la función f toma el valor 0.
- Los puntos $(0.30874, 0.57735)$ y $(0.30874, -0.57735)$ que son puntos estacionarios donde la matriz hessiana no es semidefinida positiva y que no son óptimos locales.

En las figuras 4.22, 4.23 y 4.24 ilustramos como los distintos algoritmos pueden converger a distintos óptimos globales y como, además, dicha convergencia depende del iterante inicial. ◇

Ejemplo 4.4. Para terminar presentamos un ejemplo de función continuamente diferenciable y un comportamiento bastante periódico aunque con un único óptimo global. Se trata de la función $f : \mathbb{R}^2 \to \mathbb{R}$ definida como $f(\boldsymbol{x}) = 10 \sin(x_1)^2 \cos(x_2)^2 + (x_1^2 + x_2^2)/10$. En este caso vemos en la Figura 4.25 que todos los métodos salvo el método de Newton alcanzan exitosamente el óptimo global $(0, 0)$. El método de Newton tiene problemas para moverse por la región factible y su comportamiento no es nada deseable. La razón principal es que la implementación que hemos hecho del mismo no tiene una búsqueda de línea, con lo que no tenemos asegurado que el método mejore la función objetivo entre iteraciones. Si modificásemos la implementación incluyendo una búsqueda de línea entonces este problema desaparecería. ◇

4.11 Ejercicios adicionales

••Ejercicio 4.3. Encuentra el mínimo de la función $g(\lambda) = 6e^{-2\lambda} + 2\lambda^2$ en el intervalo $[0, 10]$ usando tres de los métodos de búsqueda de línea vistos en este tema. Representa en una tabla los resultados obtenidos iteración a iteración.

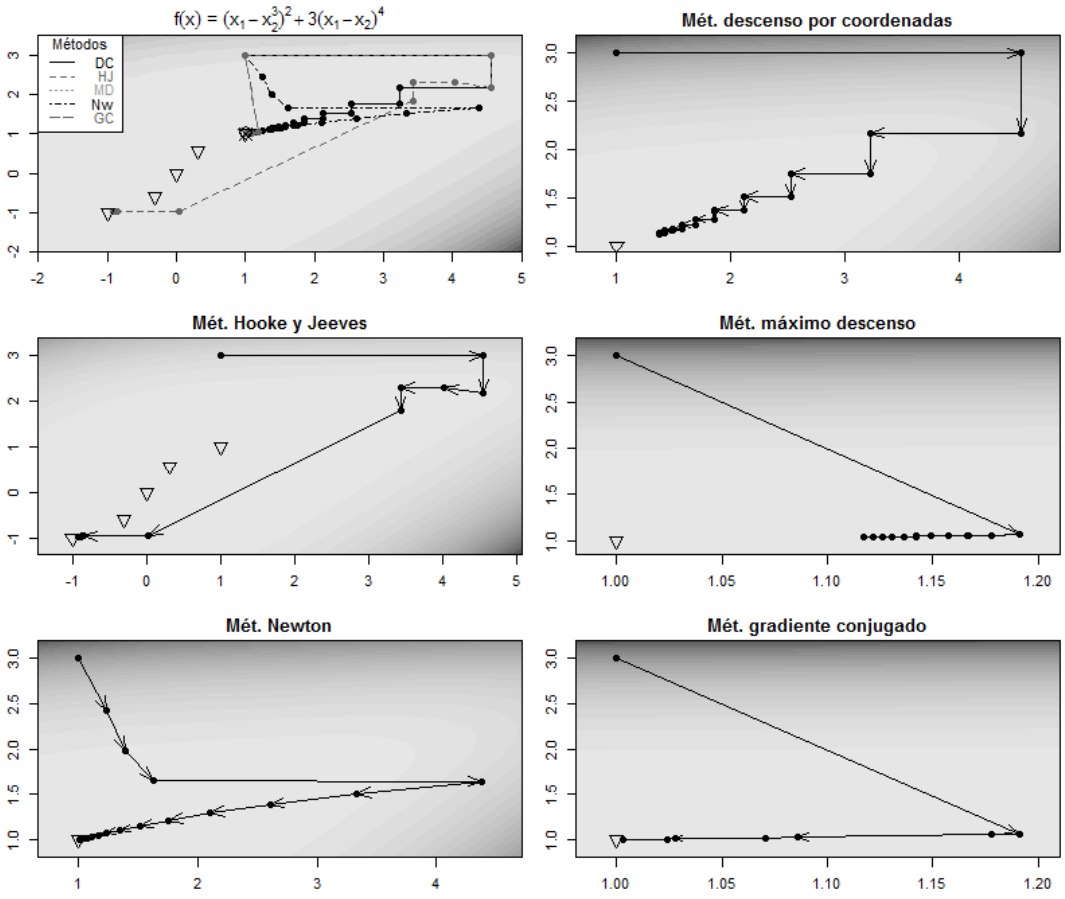


Figura 4.22: Función $f(\boldsymbol{x}) = (x_1 - x_2^3)^2 + 3(x_1 - x_2)^4$ con iterante inicial $\boldsymbol{x}^1 = (1, 3)$.

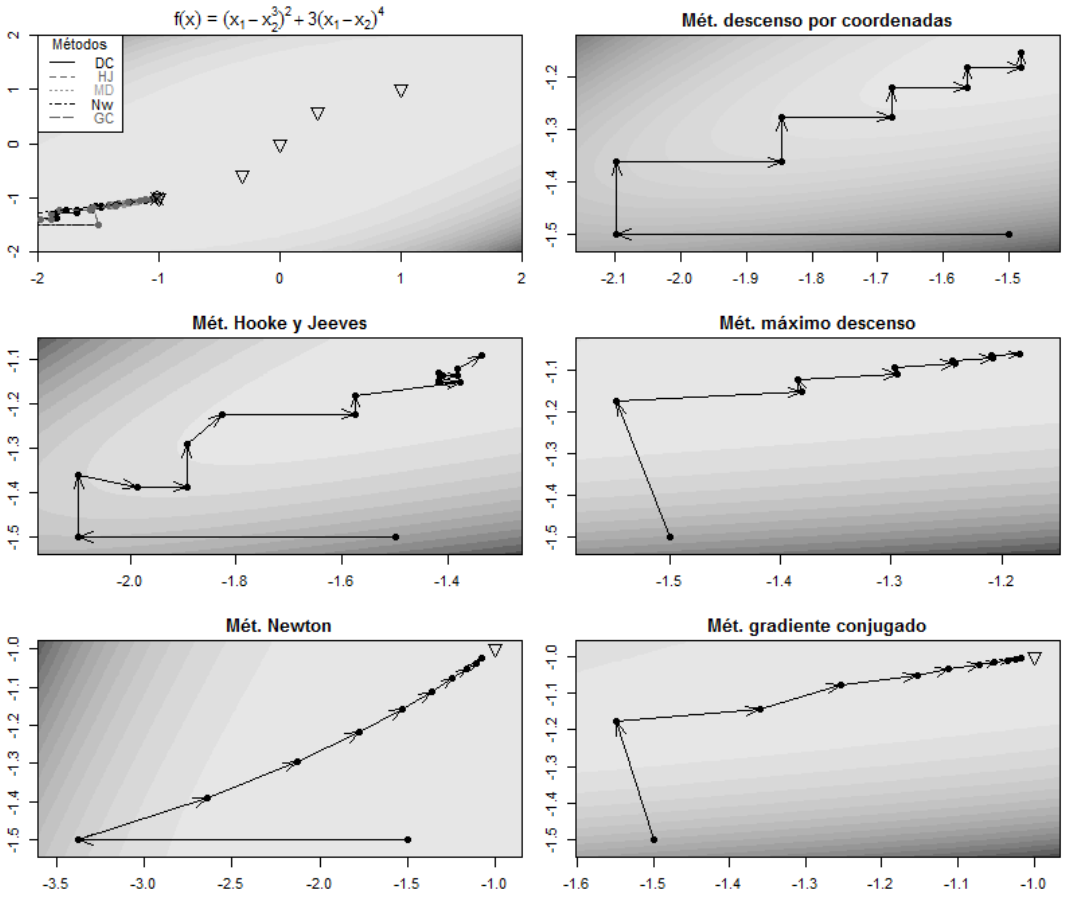


Figura 4.23: Función $f(\boldsymbol{x}) = (x_1 - x_2^3)^2 + 3(x_1 - x_2)^4$ con iterante inicial $\boldsymbol{x}^1 = (-1.5, -1.5)$.

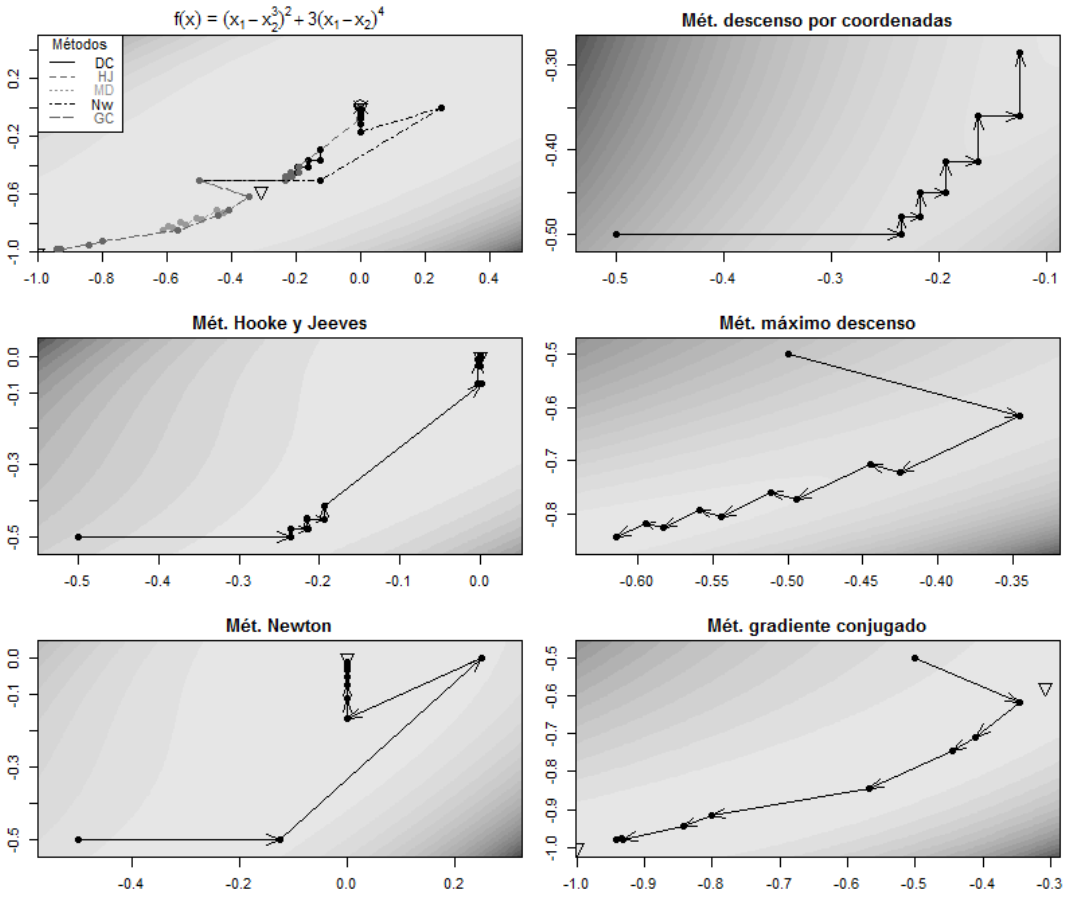


Figura 4.24: Función $f(\boldsymbol{x}) = (x_1 - x_2^3)^2 + 3(x_1 - x_2)^4$ con iterante inicial $\boldsymbol{x}^1 = (-0.5, -0.5)$.

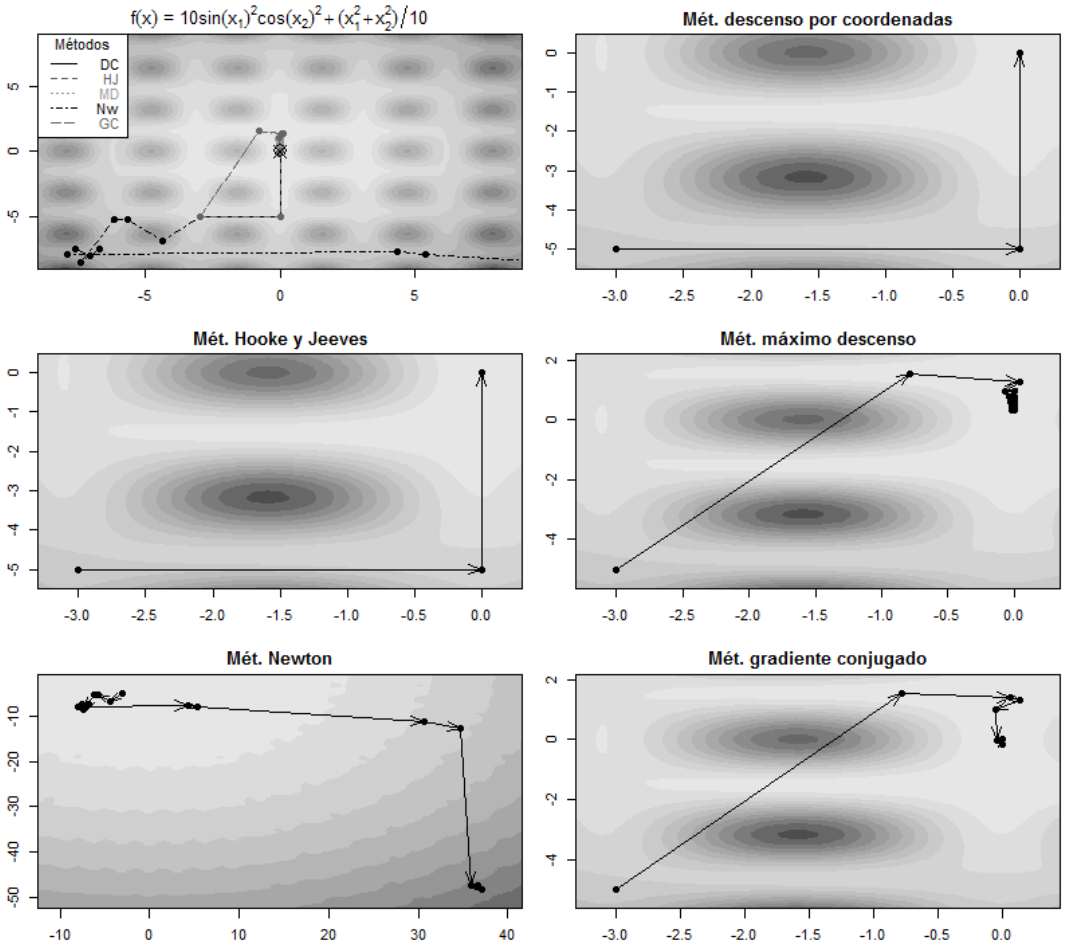


Figura 4.25: Función $f(\boldsymbol{x}) = 10\,\mathrm{sin}(x_1)^2\,\mathrm{cos}(x_2)^2+(x_1^2+x_2^2)/10$ con iterante inicial $\boldsymbol{x}^1 = (-3, 5)$.

Discute brevemente el rendimiento de los tres métodos elegidos en el ejemplo en cuestión y relaciónalos, en la medida de lo posible, con lo que cabría esperar dado lo visto en este tema. ◁

•**Ejercicio 4.4.** Para los métodos de búsqueda uniforme, dicotomía, sección áurea y Fibonacci, calcula el número de evaluaciones funcionales requeridas para $\alpha \in \{0.1, 0.01, 0.001, 0.0001\}$, donde α es el ratio entre la longitud del intervalo de incertidumbre final y el inicial.

Discute brevemente el incremento observado para los distintos métodos a medida que se reduce α y relaciona este rendimiento, en la medida de lo posible, con lo que cabría esperar dado lo visto en este tema. ◁

••**Ejercicio 4.5.** Considera la función f definida por

$$f(\boldsymbol{x}) = (x_1 + x_2^3)^2 + 2(x_1 - x_2 - 4)^4.$$

Dado un punto $\boldsymbol{x}$ y una dirección $\boldsymbol{d}$, definimos $g(\lambda) = f(\boldsymbol{x} + \lambda \boldsymbol{d})$.

(i) Para $\boldsymbol{x} = (0, 0)$ y $\boldsymbol{d} = -\nabla f(\boldsymbol{x})$, usando el método de la bisección, encuentra el valor que resuelve el problema de minimizar $g(\lambda)$.

(ii) Para $\boldsymbol{x} = (5, 4)$ y $\boldsymbol{d} = -\nabla f(\boldsymbol{x})$, usando el método de la sección áurea, encuentra el valor que resuelve el problema de minimizar $g(\lambda)$.

Discute los resultados obtenidos. ◁

••**Ejercicio 4.6.** Demuestra que el método de Fibonacci se aproxima al método de la sección áurea cuando el número de evaluaciones funcionales se hace suficientemente grande. ◁

Tema 5

Optimización con restricciones. Conceptos teóricos

Contenidos

5.1 Introducción

Este tema se puede considerar la columna vertebral de la optimización con restricciones: el estudio de los fundamentos matemáticos de las condiciones de optimalidad y de la dualidad en optimización no lineal. Trabajaremos sin apoyarnos en supuestos de convexidad, aunque siempre mantendremos los problemas convexos como punto de comparación. Por otro lado, durante prácticamente todo el análisis en este tema trabajaremos bajo supuestos de **diferenciabilidad**. Este tema se estructura en dos partes:

- Inicialmente estudiaremos las **condiciones de optimalidad** para problemas con restricciones, con el objetivo último de presentar formalmente las condiciones de Karush-Kuhn-Tucker (y mencionando de pasada las condiciones de Fritz John).
- Después pasaremos a presentar los conceptos básicos de **dualidad**. Más concretamente hablaremos de **dualidad lagrangiana**. Después presentaremos los teoremas de dualidad y sus implicaciones en el estudio de problemas de programación matemática.

5.2 Condiciones de optimalidad

Vamos a empezar esta sección recordando algunas de las condiciones de optimalidad que obtuvimos para el caso **problemas de programación convexa** con **función objetivo diferenciable**.

Corolario 2.5. $\bar{\boldsymbol{x}} \in S$ es un mínimo global si y sólo si $\nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x} - \bar{\boldsymbol{x}}) \geq 0$, para todo $\boldsymbol{x} \in S$.

Teorema 2.9. $\bar{\boldsymbol{x}} \in S$ es un mínimo global si y sólo si $D^D \cap D^F = \emptyset$.

Recordemos que, según la Definición 2.1 y la Definición 2.2, tenemos:

$$D^D(f, \boldsymbol{x}) = \{\boldsymbol{d} \in \mathbb{R}^n : f(\boldsymbol{x} + \lambda \boldsymbol{d}) < f(\boldsymbol{x}) \text{ para todo } \lambda \in (0, \delta), \text{para un cierto } \delta > 0\} \text{ y}$$
$$D^F(S, \boldsymbol{x}) = \{\boldsymbol{d} \in \mathbb{R}^n : \boldsymbol{x} + \lambda \boldsymbol{d} \in S \text{ para todo } \lambda \in (0, \delta), \text{para un cierto } \delta > 0\}.$$

Como ya comentamos en su momento, cuando no haya ambigüedad, conjuntos del tipo $D^D(f, \boldsymbol{x})$ y $D^F(S, \boldsymbol{x})$ se denotarán simplemente por D^D y D^F.

Las condiciones de optimalidad que acabamos de recordar, aunque importantes, no son fáciles de manejar en la práctica, pues su verificación resulta cuanto menos difícil. En esta sección vamos a ver cómo reemplazar estas condiciones por otras más manejables, que en particular nos servirán para estudiar optimalidad local en el caso no convexo. Además, en el caso convexo serán esencialmente equivalentes a las originales.

Recordemos también que, en el caso diferenciable, la Proposición 2.8 nos daba una condición suficiente para que una dirección sea de descenso:

Proposición 2.8. Si $\boldsymbol{d} \in \mathbb{R}^n$ es tal que $\nabla f(\boldsymbol{x})^\intercal \boldsymbol{d} < 0$, entonces $\boldsymbol{d}$ es una dirección de descenso. Si además f es convexa el recíproco también es cierto.

Este resultado da lugar a la definición de dos conjuntos asociados con el cono de direcciones de descenso, D^D:

$$D_-^D(f, \boldsymbol{x}) = \{\boldsymbol{d} \in \mathbb{R}^n : \nabla f(\boldsymbol{x})^\intercal \boldsymbol{d} < 0\} \text{ y}$$
$$D_+^D(f, \boldsymbol{x}) = \{\boldsymbol{d} \in \mathbb{R}^n : \nabla f(\boldsymbol{x})^\intercal \boldsymbol{d} \leq 0\}.$$

Es fácil ver que se tiene la siguiente relación

$$D_-^D \subseteq D^D \subseteq D_+^D.$$

La primera desigualdad se sigue inmediatamente de la Proposición 2.8 ($\nabla f(\boldsymbol{x})^\intercal \boldsymbol{d} < 0$ implica dirección de descenso). Además, este mismo resultado asegura que si f es convexa $D_-^D = D^D$. La segunda desigualdad se sigue de un resultado análogo, que nos diría que $\nabla f(\boldsymbol{x})^\intercal \boldsymbol{d} > 0$ implica que tenemos una dirección de ascenso, con lo que, necesariamente, $D^D \subseteq D_+^D$.

Ejercicio 5.1. Presenta algún ejemplo en el que las inclusiones en $D_-^D \subseteq D^D \subseteq D_+^D$ sean estrictas. ◁

A continuación presentamos una primera condición necesaria de optimalidad local. Este resultado, cuya ilustración gráfica puede verse en la Figura 5.1, nos dice que si en un punto dado tenemos que $D_-^D \cap D^F \neq \emptyset$, entonces hay alguna dirección que es simultáneamente de descenso y factible, con lo que ese punto no puede ser un mínimo local. Nótese que el enunciado de este resultado es similar al del Teorema 2.9, pero ahora no se asume convexidad, ni de f ni de S. Por otro lado, sí que se asume diferenciabilidad de f en el punto en estudio para poder hablar del conjunto D_-^D.

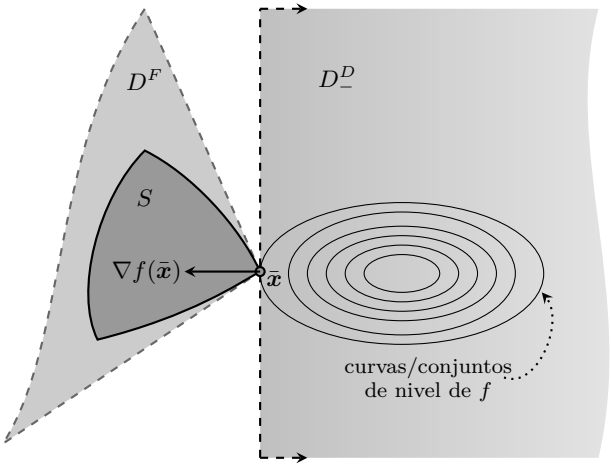


Figura 5.1: Condición necesaria de optimalidad local.

Teorema 5.1. *Dada la función $f : \mathbb{R}^n \to \mathbb{R}$ y el conjunto $S \subseteq \mathbb{R}$ un conjunto no vacío. Consideremos el siguiente problema de optimización:*

$$\begin{aligned} \text{minimizar} \quad & f(\boldsymbol{x}) \\ \text{sujeto a} \quad & \boldsymbol{x} \in S. \end{aligned}$$

Si f es diferenciable en el punto $\bar{\boldsymbol{x}} \in S$ y $\bar{\boldsymbol{x}}$ es un mínimo local, entonces $D_-^D \cap D^F = \emptyset$.

Demostración. Supongamos que $\bar{\boldsymbol{x}}$ es un mínimo local y que existe $\boldsymbol{d} \in D_-^D \cap D^F$. Entonces, por la Proposición 2.8, $\boldsymbol{d} \in D^D$, con lo que existe $\delta_1 > 0$ tal que, para todo $\lambda \in (0, \delta_1)$,

$$f(\boldsymbol{x} + \lambda \boldsymbol{d}) < f(\boldsymbol{x}).$$

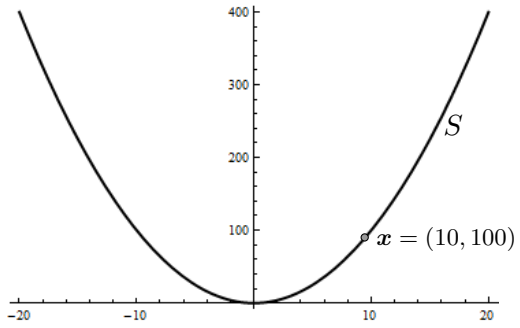
Además, como $\boldsymbol{d} \in D^F$, existe $\delta_2 > 0$ tal que, para todo $\lambda \in (0, \delta_2)$,

$$\boldsymbol{x} + \lambda \boldsymbol{d} \in S,$$

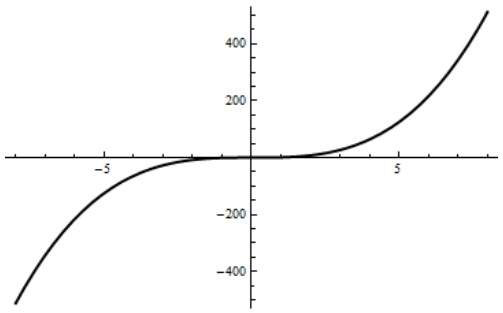
con lo que tenemos una contradicción con que $\bar{\boldsymbol{x}}$ sea un mínimo local. □

Nótese que estas condiciones en general no son suficientes, como ilustramos con dos ejemplos en la Figura 5.2:

- En la Figura 5.2(a), tenemos la función $f(\boldsymbol{x}) = x_2$, definida sobre el conjunto $S = \{\boldsymbol{x} \in \mathbb{R}^2 : x_2 = x_1^2\}$. El dominio es una curva en $\mathbb{R}^2$ que en ningún punto tiene direcciones factibles, $D^F = \emptyset$. Por tanto, cualquier punto cumple que $D_-^D \cap D^F = \emptyset$, y únicamente $\boldsymbol{x} = (0, 0)$ es un mínimo local (y global).
- En la Figura 5.2(b) tenemos la función $f(x) = x^3$, que en $x = 0$ cumple que $\nabla f(0) = 0$, con lo que $D_-^D = \emptyset$. Por tanto, $x = 0$ cumple que $D_-^D \cap D^F = \emptyset$ pero no es un mínimo local.



(a) Conjunto $S = \{\boldsymbol{x} \in \mathbb{R}^2 : x_2 = x_1^2\}$, dominio de $f : S \to \mathbb{R}$ dada por $f(\boldsymbol{x}) = x_2$.



(b) Función $f : \mathbb{R} \to \mathbb{R}$ dada por $f(x) = x^3$.

Figura 5.2: La condición del Teorema 5.1 no es suficiente.

Un aspecto importante del Teorema 5.1, de cara al diseño de algoritmos, es que D_-^D es más fácil de manejar que D^D, pues D_-^D se define a través de una condición directa sobre el gradiente. En la Figura 5.1 puede verse que D_-^D no es más que un semiespacio. El siguiente paso es conseguir algo parecido para el conjunto D^F, para pasar de expresiones puramente geométricas relativas a conos de direcciones de descenso y factibles, a condiciones puramente algebraicas.

Para facilitar la exposición, supondremos por el momento que tenemos un problema de programación matemática con restricciones de desigualdad únicamente. Es decir, problemas de la forma:

$$\begin{array}{ll} \text{minimizar} & f(\boldsymbol{x}) \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 \qquad i = 1, \ldots, m. \end{array}$$

Por tanto, tenemos el conjunto factible $S = \{\boldsymbol{x} \in \mathbb{R}^n : g_i(\boldsymbol{x}) \leq 0 \text{ para todo } i \in \{1, \dots, m\}\}$. Además, dado $\boldsymbol{x} \in S$, denotamos por $I(\boldsymbol{x}) = \{i : g_i(\boldsymbol{x}) = 0\}$, el conjunto de las restricciones activas o saturadas en $\boldsymbol{x}$. Dado un punto $\boldsymbol{x} \in S$, las restricciones asociadas a $I(\boldsymbol{x})$ son las que limitan los movimientos desde el punto $\boldsymbol{x}$, pues estamos en el interior del conjunto definido por las otras restricciones. Como caso extremo, si $I(\boldsymbol{x}) = \emptyset$ quiere decir que $\boldsymbol{x} \in \mathring{S}$ y cualquier dirección será una dirección factible.

Estamos ahora en condiciones de definir dos nuevos conjuntos que, como veremos posteriormente, están muy estrechamente ligados a D^F.

$$D_-^F(S, \boldsymbol{x}) = \{\boldsymbol{d} \in \mathbb{R}^n : \nabla g_i(\boldsymbol{x})^\intercal \boldsymbol{d} < 0 \text{ para todo } i \in I(\boldsymbol{x})\} \text{ y}$$
$$D_+^F(S, \boldsymbol{x}) = \{\boldsymbol{d} \in \mathbb{R}^n : \nabla g_i(\boldsymbol{x})^\intercal \boldsymbol{d} \leq 0 \text{ para todo } i \in I(\boldsymbol{x})\}.$$

El siguiente resultado establece una relación análoga de estos conjuntos con respecto al conjunto D^F a la de los conjuntos D_-^D y D_+^D con respecto al conjunto D^D.

Proposición 5.2. *Consideremos la región factible $S = \{\boldsymbol{x} \in \mathbb{R}^n : g_i(\boldsymbol{x}) \leq 0$ para todo $i \in \{1, \dots, m\}\}$. Dado un punto $\bar{\boldsymbol{x}} \in S$, si*

- *las funciones g_i con $i \in I(\bar{\boldsymbol{x}})$ son diferenciables en $\bar{\boldsymbol{x}}$ y*
- *las funciones g_i con $i \notin I(\bar{\boldsymbol{x}})$ son continuas en $\bar{\boldsymbol{x}}$,*

entonces

$$D_-^F \subseteq D^F \subseteq D_+^F.$$

Además, si las funciones g_i con $i \in I(\bar{\boldsymbol{x}})$ son estrictamente convexas en $\bar{\boldsymbol{x}}$, entonces $D_-^F = D^F$.[1]

Demostración. Supongamos que $\boldsymbol{d} \in D_-^F$. Para las restricciones $i \notin I(\bar{\boldsymbol{x}})$ tenemos que $g_i(\bar{\boldsymbol{x}}) < 0$. Entonces, por la continuidad estas funciones, existe $\delta_1 > 0$ tal que, para todo $\lambda \in (0, \delta_1)$ y todo $i \notin I(\bar{\boldsymbol{x}})$,

$$g_i(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) < 0.$$

Además, como $\boldsymbol{d} \in D_-^F$, para todo $i \in I(\bar{\boldsymbol{x}})$ tenemos $\nabla g_i(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} < 0$. Aplicando ahora la Proposición 2.8 tenemos que $\boldsymbol{d}$ es una dirección de descenso para las g_i. Es decir, existe δ_2 tal que, para todo $\lambda \in (0, \delta_2)$ y todo $i \in I(\bar{\boldsymbol{x}})$,

$$g_i(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) < g_i(\bar{\boldsymbol{x}}) = 0.$$

Combinando las dos ecuaciones anteriores, tomando $\delta = \min\{\delta_1, \delta_2\}$, tenemos que, para todo $\lambda \in (0, \delta)$, los puntos de la forma $\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}$ son factibles. Por tanto, $\boldsymbol{d} \in D^F$ con lo que hemos probado que $D_-^F \subseteq D^F$.

La demostración de que $D^F \subseteq D_+^F$ es similar. Si $\boldsymbol{d} \in D^F$ y $\boldsymbol{d} \notin D_+^F$, entonces tendríamos que $\nabla g_i(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} > 0$ para algún $i \in I(\bar{\boldsymbol{x}})$. La Proposición 2.8 implicaría ahora que $\boldsymbol{d}$ es una dirección de ascenso para g_i en $\bar{\boldsymbol{x}}$. Por tanto, tomando $\lambda > 0$ suficientemente pequeño tendríamos que $g_i(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) > g_i(\bar{\boldsymbol{x}}) = 0$, contradiciendo que $\boldsymbol{d} \in D^F$.

Para terminar, supongamos que las funciones g_i con $i \in I(\bar{\boldsymbol{x}})$ son estrictamente convexas en $\bar{\boldsymbol{x}}$. Tomemos ahora $\boldsymbol{d} \in D^F$. Supongamos ahora que $\boldsymbol{d} \notin D_-^F$. Entonces existe $i \in I(\bar{\boldsymbol{x}})$ tal que $\nabla g_i(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} \geq 0$. Como g_i es estrictamente convexa tenemos que, para todo $\lambda > 0$, $g_i(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) > g_i(\bar{\boldsymbol{x}}) + \nabla g_i(\bar{\boldsymbol{x}})^\intercal \lambda \boldsymbol{d} > g_i(\bar{\boldsymbol{x}}) = 0$. Pero esto contradice que $\boldsymbol{d} \in D^F$. □

[1] Para este último resultado sería suficiente con pedir seudoconvexidad estricta de las g_i con $i \in I(\bar{\boldsymbol{x}})$.

Es fácil ver que en el ejemplo de la Figura 5.1 se tiene que $D_-^F = D^F \subsetneq D_+^F$.

•**Ejercicio 5.2.** Presenta ejemplos mostrando las distintas posibilidades para las inclusiones $D_-^F \subseteq D^F \subseteq D_+^F$ (estrictas y no estrictas). ◁

Ahora estamos en condiciones de presentar una nueva condición necesaria de optimalidad, que recoge la esencia de las condiciones de Karush-Kuhn-Tucker que veremos en la siguiente sección.

Teorema 5.3. *Consideremos el siguiente problema de optimización:*

$$\begin{array}{rl} \text{minimizar} & f(\boldsymbol{x}) \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 \qquad i = 1, \ldots, m. \end{array}$$

Supongamos que $\bar{\boldsymbol{x}}$ es un punto factible tal que

- *las funciones f y g_i con $i \in I(\bar{\boldsymbol{x}})$ son diferenciables en $\bar{\boldsymbol{x}}$ y*
- *las funciones g_i con $i \notin I(\bar{\boldsymbol{x}})$ son continuas en $\bar{\boldsymbol{x}}$.*

Si $\bar{\boldsymbol{x}}$ es un mínimo local, entonces $D_-^D \cap D_-^F = \emptyset$. Además, si f es convexa y las g_i con $i \in I(\bar{\boldsymbol{x}})$ son estrictamente convexas en $\bar{\boldsymbol{x}}$, $D_-^D \cap D_-^F = \emptyset$ implica que $\bar{\boldsymbol{x}}$ es un mínimo local (y global).

Demostración. Si $\bar{\boldsymbol{x}}$ es un mínimo local, por el Teorema 5.1 tenemos que $D_-^D \cap D^F = \emptyset$. Además, la Proposición 5.2 nos dice que $D_-^F \subseteq D^F$. Por tanto, $D_-^D \cap D_-^F = \emptyset$.

Si f es convexa sabemos que $D_-^D = D^D$ y, además, la convexidad estricta de las funciones g_i con $i \in I(\bar{\boldsymbol{x}})$ implica que $D_-^F = D^F$ (Proposición 5.2). Por tanto, $D^D \cap D^F = D_-^D \cap D_-^F = \emptyset$ y el Teorema 2.9 nos asegura que $\bar{\boldsymbol{x}}$ es un óptimo local (y global). □

Recapitulemos brevemente antes de pasar a la siguiente sección. Las definiciones de los conjuntos D_-^D y D_-^F son tales que, en todo mínimo local, ha de cumplirse que $D_-^D \cap D_-^F = \emptyset$. Además, sabemos que en general el recíproco no es cierto como podemos ver, por ejemplo, con las funciones de la Figura 5.2. En la siguiente sección veremos que, a cambio, la condición $D_-^D \cap D_-^F = \emptyset$ es mucho más manejable que la condición $D^D \cap D^F = \emptyset$.

En la Figura 5.3 tenemos una representación de un punto $\bar{\boldsymbol{x}}$ en el que se cumple la condición $D_-^D \cap D_-^F = \emptyset$, y en la que además hemos representado los gradientes de las restricciones activas en el punto junto con el gradiente de la función objetivo.

El hecho de que en el punto $\bar{\boldsymbol{x}}$ se cumpla la condición $D_-^D \cap D_-^F = \emptyset$ nos dice que no hay ningún vector $\boldsymbol{d}$ que forme un ángulo mayor de 90° con $\nabla f(\bar{\boldsymbol{x}})$ y, simultáneamente, mayor de 90° con todos los gradientes de las restricciones activas en $\bar{\boldsymbol{x}}$. Geométricamente, esto está muy relacionado con el hecho de que $-\nabla f(\bar{\boldsymbol{x}})$ esté en el cono generado por los gradientes de dichas restricciones, como se puede ver en la Figura 5.3. Son justamente estas intuiciones la que formalizaremos en la siguiente sección, primero a través de las condiciones de Fritz John y después a través de las condiciones de KKT.

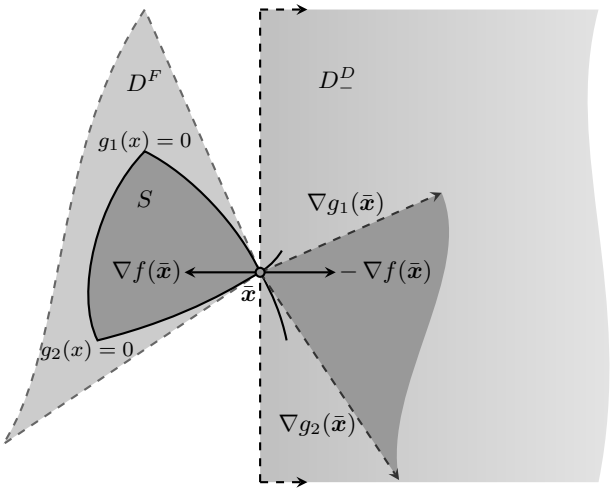


Figura 5.3: Ilustración de un punto en el que se cumple que $D^D_- \cap D^F_- = \emptyset$.

5.2.1 Ejemplos

Ejemplo 5.1. Consideremos el siguiente problema de optimización:

$$\begin{aligned} \text{minimizar} \quad & (x_1-3)^2 + (x_2-2)^2 \\ \text{sujeto a} \quad & x_1^2 + x_2^2 \leq 5 \\ & x_1 + x_2 \leq 3 \\ & x_1 \geq 0 \\ & x_2 \geq 0. \end{aligned}$$

En este problema tenemos $f(\boldsymbol{x}) = (x_1-3)^2 + (x_2-2)^2$ y cuatro restricciones de desigualdad que, con la formulación $g_i(\boldsymbol{x}) \leq 0$ se corresponderían con:

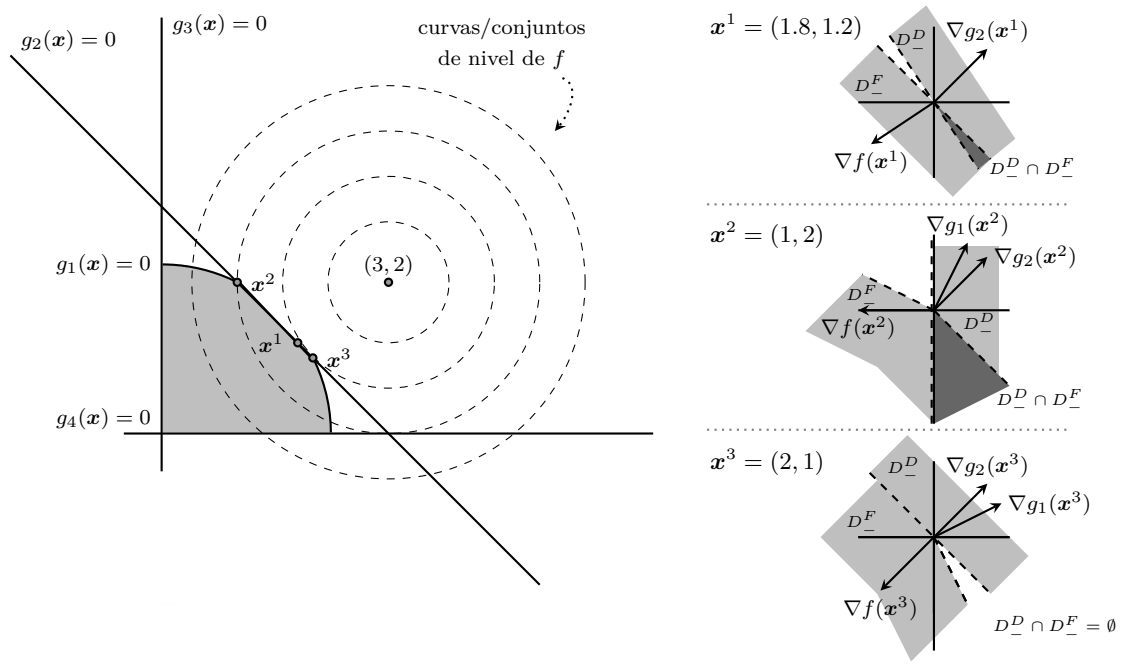
$$\begin{aligned} g_1(\boldsymbol{x}) &= x_1^2 + x_2^2 - 5, & g_3(\boldsymbol{x}) &= -x_1, \\ g_2(\boldsymbol{x}) &= x_1 + x_2 - 3, & g_4(\boldsymbol{x}) &= -x_2. \end{aligned}$$

Este problema aparece representado en la Figura 5.4(a). En ella se puede ver que el punto $(2,1)$ será el óptimo global del problema, pues, como marcan las curvas de nivel de f, es el punto factible que más está cerca de $(3,2)$. La función objetivo en este punto es $f(2,1) = 2$ Vamos a estudiar tres puntos factibles y evaluar en ellos la condición necesaria de optimalidad: $D^D_- \cap D^F_- = \emptyset$. Previamente, veamos la expresión de los gradientes de función objetivo y restricciones:

$$\begin{aligned} \nabla f(\boldsymbol{x}) &= (2x_1 - 6, 2x_2 - 4), \\ \nabla g_1(\boldsymbol{x}) &= (2x_1, 2x_2), & \nabla g_3(\boldsymbol{x}) &= (-1, 0), \\ \nabla g_2(\boldsymbol{x}) &= (1,1), & \nabla g_4(\boldsymbol{x}) &= (0,-1). \end{aligned}$$

Primer punto: $\boldsymbol{x}^1 = (1.8, 1.2)$. La función objetivo es $f(1.8, 1.2) = 2.08$ con $I(\boldsymbol{x}^1) = \{2\}$, pues únicamente la restricción $g_2(\boldsymbol{x}) = x_1 + x_2 - 3$ está activa en $(1.8, 1.2)$. Para calcular D^D_- y D^F_- necesitamos los gradientes, que vienen dados por:

$$\nabla f(\boldsymbol{x}^1) = (-2.4, -1.6) \quad \text{y} \quad \nabla g_2(\boldsymbol{x}^1) = (1,1).$$



(a) Región factible y curvas de nivel.

(b) Estudiando si $D_-^D \cap D_-^F = \emptyset$.

Figura 5.4: Ilustración del Ejemplo 5.1.

Por tanto, $D_-^D = \{\boldsymbol{d} : (-2.4, -1.6)\cdot\boldsymbol{d} < 0\}$ y $D_-^F = \{\boldsymbol{d} : (1,1)\cdot\boldsymbol{d} < 0\}$. En la parte superior de la Figura 5.4(b) podemos ver los conjuntos D_-^D y D_-^F y vemos que $D_-^D \cap D_-^F \neq \emptyset$, y tenemos direcciones que nos permiten acercarnos hacia x^3 mateniéndonos dentro del conjunto y, al mismo tiempo, mejorando la función objetivo.

Segundo punto: $\boldsymbol{x}^2 = (1,2)$. La función objetivo es $f(1,2) = 4$, y ahora tenemos que $I(\boldsymbol{x}^2) = \{1,2\}$ y los gradientes asociados a $\boldsymbol{x}^2 = (1,2)$ son:

$$\nabla f(\boldsymbol{x}^2) = (-4, 0), \quad \nabla g_1(\boldsymbol{x}^2) = (2, 4), \quad \text{y} \quad \nabla g_2(\boldsymbol{x}^2) = (1, 1).$$

Podemos ver en la parte intermedia de la Figura 5.4(b) que, nuevamente $D_-^D \cap D_-^F \neq \emptyset$. Por tanto, $\boldsymbol{x}^2 = (1,2)$ no cumple la condición necesaria de optimalidad local.

Tercer punto: $\boldsymbol{x}^3 = (2,1)$. Veamos ahora que pasa con el punto que ya hemos anticipado que es el óptimo global con $f(2,1) = 2$. Tenemos que $I(\boldsymbol{x}^3) = \{1,2\}$ y los gradientes asociados a $\boldsymbol{x}^3 = (2,1)$ son:

$$\nabla f(\boldsymbol{x}^3) = (-2, -2), \quad \nabla g_1(\boldsymbol{x}^3) = (4, 2), \quad \text{y} \quad \nabla g_2(\boldsymbol{x}^3) = (1, 1).$$

Ahora podemos ver en la parte inferior de la Figura 5.4(b) que $D_-^D \cap D_-^F = \emptyset$ con lo que la condición necesaria de optimalidad se cumple en el óptimo, $\boldsymbol{x}^3 = (2,1)$. ◇

Para terminar esta sección presentamos otro ejemplo que ilustra que la utilidad de la condición $D_-^D \cap D_-^F = \emptyset$ para identificar óptimos locales puede depender de cómo estén expresadas

las restricciones. Esto supone una limitación más para esta condición necesaria, ya que su efectividad, para un problema dado, puede ser dependiente de la formulación escogida para el mismo. En el ejemplo siguiente se podría saber a priori la formulación que sería preferible, pero este no tiene por qué ser el caso en general.

Ejemplo 5.2. Consideremos el siguiente problema de optimización:

$$\begin{aligned} \text{minimizar} \quad & (x_1-1)^2+(x_2-1)^2 \\ \text{sujeto a} \quad & (x_1+x_2-1)^3 \leq 0 \\ & x_1 \geq 0 \\ & x_2 \geq 0. \end{aligned}$$

Para este problema tenemos que, dado un punto factible $\boldsymbol{x}$,

$$\nabla f(\boldsymbol{x}) = (2x_1-2, 2x_2-2) \quad \text{y} \quad \nabla g_1(\boldsymbol{x}) = \Big(3(x_1+x_2-1)^2, 3(x_1+x_2-1)^2\Big).$$

Ahora, para cualquier punto factible $\boldsymbol{x}$ en el que $x_1+x_2=1$ tendremos que $\nabla g_1(\boldsymbol{x}) = (0,0)$. Por tanto, $D_-^F = \emptyset$, con lo que $D_-^D \cap D_-^F = \emptyset$ y la condición necesaria no es de gran ayuda.

Por otro lado, la restricción $(x_1+x_2-1)^3 \leq 0$ es claramente equivalente a $x_1+x_2-1 \leq 0$, con lo que ahora el gradiente de la restricción asociada sería $(1,1)$. Con esta nueva formulación, no sería difícil comprobar que la condición $D_-^D \cap D_-^F = \emptyset$ sólo se cumplirá en el punto $(0.5, 0.5)$, que además resulta ser el único óptimo global del problema. ◇

Este ejemplo ilustra una importante limitación de la condición necesaria de optimalidad $D_-^D \cap D_-^F = \emptyset$. Vamos a dividir esta limitación en dos casos, de muy distinta relevancia:

- La condición se cumplirá siempre que $D_-^D = \emptyset$, independientemente de cómo sea D_-^F en el punto dado. Esto no es un gran problema, aunque, al igual que pasa en problemas sin restricciones, podemos encontrarnos con que la condición se cumple en máximos y puntos de silla, que no es lo que estamos buscando. Al igual que en el caso de problemas sin restricciones, uno puede recurrir a condiciones sobre las derivadas de segundo orden (la hessiana) para evaluar qué pasa exactamente en el punto en cuestión. Por supuesto, el problema también desaparece bajo ciertos supuestos de convexidad.

- La condición también se cumplirá siempre que $D_-^F = \emptyset$. Este ya es un problema más grande, pues estamos hablando de que una condición de optimalidad se cumplirá independientemente de cómo sea el gradiente de la función objetivo, lo que parece no deseable. Esto es exactamente lo que pasa con todos los puntos de la forma $x_1+x_2=1$ en el Ejemplo 5.2. Tenemos infinitos puntos que cumplen la condición necesaria y únicamente uno de ellos es un óptimo local.

- La utilidad de la condición $D_-^D \cap D_-^F = \emptyset$ puede variar entre formulaciones equivalentes de un mismo problema.

Las limitaciones que acabamos de comentar las heredarán las condiciones de Fritz John que veremos en la siguiente sección, pues se trata de unas condiciones que son equivalentes a $D_-^D \cap D_-^F = \emptyset$. Sin embargo, gracias a las condiciones de Karush-Kuhn-Tucker, estas limitaciones se verán sustancialmente atenuadas.

5.3 Condiciones de Karush-Kuhn-Tucker

En esta sección seguimos avanzando en nuestro objetivo de obtener condiciones de optimalidad manejables. Cuando decimos manejables nos referimos a condiciones que no sólo aporten intuiciones geométricas, sino que también puedan ser aprovechadas para diseñar algoritmos y para hacer desarrollos teóricos sobre distintas clases de problemas de optimización.

5.3.1 Condiciones de Fritz John para problemas sin restricciones de igualdad

Comenzaremos formalizando las intuiciones desarrolladas al final de la sección anterior, para lo cual nos será de gran importancia el siguiente resultado, conocido como Teorema de Gordan y cuya demostración se puede obtener como un corolario bastante sencillo del Lema de Farkas.

Teorema 5.4 (Teorema de Gordan)**.** *Tomemos $\boldsymbol{A}_{m\times n}$. Entonces uno y sólo uno de los siguientes sistemas tiene solución:*

$$\textit{Sistema 1. } \boldsymbol{Ax} < \mathbf{0} \textit{ para algún } \boldsymbol{x} \in \mathbb{R}^n.$$
$$\textit{Sistema 2. } \boldsymbol{A}^\intercal \boldsymbol{y} = \mathbf{0} \textit{ e } \boldsymbol{y} \geq \mathbf{0},\ \boldsymbol{y} \neq \mathbf{0}, \textit{ para algún } \boldsymbol{y} \in \mathbb{R}^m.$$

••Ejercicio 5.3. Demuestra el Teorema de Gordan apoyándote en el Lema de Farkas. ◁

Ahora estamos en condiciones de presentar las condiciones de Fritz John (John, 1948), que no son más que una reformulación de la condición $D_-^D \cap D_-^F = \emptyset$. Como veremos en los ejemplos de la Sección 5.3.2, estas condiciones resultan más manejables para trabajar de modo analítico.

Teorema 5.5 (Condiciones necesarias de Fritz John)**.** *Consideremos el siguiente problema de optimización:*

$$\begin{aligned} \text{minimizar} \quad & f(\boldsymbol{x}) \\ \text{sujeto a} \quad & g_i(\boldsymbol{x}) \leq 0 \qquad i = 1, \ldots, m. \end{aligned}$$

Supongamos que $\bar{\boldsymbol{x}}$ es un punto factible tal que

- *las funciones f y g_i con $i \in I(\bar{\boldsymbol{x}})$ son diferenciables en $\bar{\boldsymbol{x}}$ y*
- *las funciones g_i con $i \notin I(\bar{\boldsymbol{x}})$ son continuas en $\bar{\boldsymbol{x}}$.*

Si $\bar{\boldsymbol{x}}$ es un mínimo local, entonces existen escalares u_0 y u_i para todo $i \in I(\bar{\boldsymbol{x}})$ tales que

$$\begin{gathered} u_0 \nabla f(\bar{\boldsymbol{x}}) + \sum_{i \in I(\bar{\boldsymbol{x}})} u_i \nabla g_i(\bar{\boldsymbol{x}}) = \mathbf{0} \\ u_0 \geq 0, u_i \geq 0 \quad \textit{para todo } i \in I(\bar{\boldsymbol{x}}) \\ (u_0, \boldsymbol{u}_{I(\bar{\boldsymbol{x}})}) \neq (0, \mathbf{0}). \end{gathered}$$

Si las funciones g_i con $i \notin I(\bar{\boldsymbol{x}})$ son también diferenciables en $\bar{\boldsymbol{x}}$, entonces las condiciones

anteriores pueden ser escritas, de modo equivalente, con respecto a $(u_0, \boldsymbol{u}) \in \mathbb{R}^{m+1}$ *como*

$$\begin{aligned} u_0 \nabla f(\bar{\boldsymbol{x}}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{\boldsymbol{x}}) &= \mathbf{0} \\ u_i g_i(\bar{\boldsymbol{x}}) &= 0 \quad \textit{para todo } i \in \{1, \dots, m\} \\ u_0 \geq 0, u_i &\geq 0 \quad \textit{para todo } i \in \{1, \dots, m\} \\ (u_0, \boldsymbol{u}_{I(\bar{\boldsymbol{x}})}) &\neq (0, \mathbf{0}). \end{aligned}$$

Demostración. Como $\bar{\boldsymbol{x}}$ es un óptimo local, el Teorema 5.3 nos asegura que $D_-^D \cap D_-^F = \emptyset$. Equivalentemente, no existe ningún $\boldsymbol{d} \in \mathbb{R}^n$ tal que $\nabla f(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} < 0$ y $\nabla g_i(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} < 0$ para todo $i \in I(\bar{\boldsymbol{x}})$. Si llamamos $\boldsymbol{A}$ a la matriz cuyas filas son dichos gradientes (traspuestos), entonces la condición $D_-^D \cap D_-^F = \emptyset$ es equivalente a que el sistema $\boldsymbol{A}\boldsymbol{d} < \mathbf{0}$ no tenga solución. Por el Teorema de Gordan (Teorema 5.4), existirá un vector $\boldsymbol{y}$ con tantas componentes como filas tiene la matriz $\boldsymbol{A}$ tal que $\boldsymbol{y} \geq \mathbf{0}$, $\boldsymbol{y} \neq \mathbf{0}$ y $\boldsymbol{A}^\intercal \boldsymbol{y} = \mathbf{0}$. Si denotamos la primera componente de $\boldsymbol{y}$ por u_0 y el resto por u_i con $i \in I(\bar{\boldsymbol{x}})$, tenemos la primera formulación del resultado.

La segunda formulación se obtiene simplemente definiendo $u_i = 0$ para todo $i \notin I(\bar{\boldsymbol{x}})$ (la diferenciabilidad de las g_i con $i \notin I(\bar{\boldsymbol{x}})$ en $\bar{\boldsymbol{x}}$ es necesaria únicamente para poder hablar de los gradientes de estas restricciones en $\bar{\boldsymbol{x}}$). □

Las condiciones necesarias de FJ, al igual que pasará con las condiciones de KKT que veremos más adelante, se pueden convertir en condiciones suficientes bajo supuestos adicionales de convexidad (seudoconvexidad en el caso de las condiciones de FJ).

Nótese que, con respecto a la intuición desarrollada en la sección anterior, estas condiciones no implican que $-\nabla f(\bar{\boldsymbol{x}})$ sea una combinación cónica de los gradientes de las restricciones activas en $\bar{\boldsymbol{x}}$, pues u_0 puede valer 0. Esto es justamente lo que sucede en situaciones como la del Ejemplo 5.2 y, como comentamos al discutir dicho ejemplo, resulta problemático para las condiciones de Fritz John (y es solventado de modo bastante satisfactorio por las condiciones de Karush-Kuhn-Tucker).

Los escalares u_0 y u_i se conocen como *multiplicadores de Lagrange* y, como discutiremos más adelante, nos permiten obtener información acerca de ciertas características del punto $\bar{\boldsymbol{x}}$ relevantes para el problema de optimización en estudio. La condición de que $\bar{\boldsymbol{x}}$ sea factible se conoce como *condición de factibilidad del primal (FP)*, las condiciones $u_i g_i(\bar{\boldsymbol{x}}) = 0$ son las *condiciones de holguras complementarias (HC)*, y las condiciones sobre los gradientes y sobre $(u_0, \boldsymbol{u})$ se conocen como *condiciones de factibilidad del dual (FD)*. El papel de las holguras complementarias es exactamente el mismo que en los problemas de programación lineal. En particular, el multiplicador de Lagrange asociado a las restricciones que no se saturan en $\bar{\boldsymbol{x}}$ será cero. Como comentaremos más adelante y veremos con más detalle cuando introduzcamos la dualidad, los multiplicadores de Lagrange se corresponden justamente con la solución óptima del problema dual asociado a nuestro problema de optimización.

Cualquier punto $\bar{\boldsymbol{x}}$ para el que existen multiplicadores de Lagrange $(\bar{u}_0, \bar{\boldsymbol{u}})$ cumpliendo las condiciones FP, FD y HC se llamará punto Fritz John.

5.3.2 Condiciones de Fritz John: Discusión y ejemplos

Ejemplo 5.3. Consideremos el siguiente problema de optimización, que es una pequeña variación del visto en el Ejemplo 5.1:

$$\begin{array}{rl} \text{minimizar} & (x_1-3)^2+(x_2-2)^2 \\ \text{sujeto a} & x_1^2+x_2^2 \le 5 \\ & x_1+2x_2 \le 4 \\ & x_1 \ge 0 \\ & x_2 \ge 0. \end{array}$$

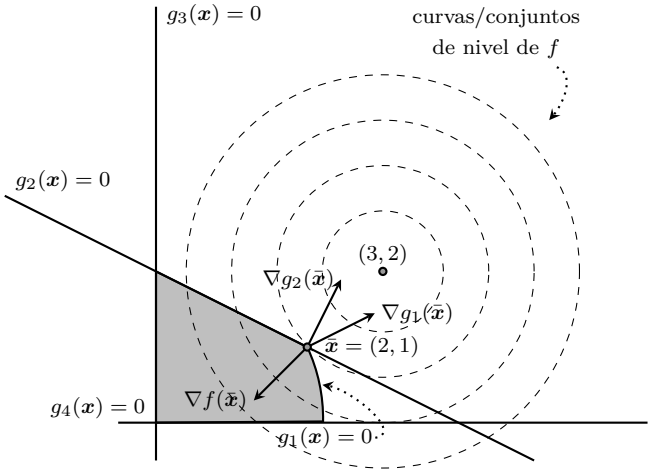


Figura 5.5: Región factible y curvas de nivel del problema del Ejemplo 5.3.

La región factible y las curvas de nivel de este problema están representadas en la Figura 5.5. El óptimo sigue siendo el punto $\bar{\boldsymbol{x}}=(2,1)$, aunque los gradientes de las restricciones en dicho punto, también representados en la figura, cambian ligeramente con respecto al Ejemplo 5.1. En este caso tenemos:

$$\nabla f(\bar{\boldsymbol{x}})=(-2,-2), \quad \nabla g_1(\bar{\boldsymbol{x}})=(4,2) \quad \text{y} \quad \nabla g_2(\bar{\boldsymbol{x}})=(1,2).$$

Del mismo modo que hicimos en el Ejemplo 5.1, podríamos comprobar que $D_-^D \cap D_-^F = \emptyset$, y obtendríamos una situación muy similar a la de la parte inferior de la Figura 5.4(b). Sin embargo, es más fácil comprobar las condiciones de Fritz John, que son mucho más manejables analíticamente. En este caso, se reducen a buscar un vector no nulo $(u_0,u_1,u_2)\ge 0$ tal que

$$u_0\begin{pmatrix}-2\\-2\end{pmatrix}+u_1\begin{pmatrix}4\\2\end{pmatrix}+u_2\begin{pmatrix}1\\2\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}.$$

Es fácil comprobar que $u_0=3$, $u_1=1$ y $u_2=2$ es una solución de dicho sistema y, por tanto, el punto $(2,1)$ cumple la condición de factibilidad del dual de FJ. De hecho, tenemos tres variables y dos ecuaciones, con lo que podemos despejar u_1 y u_2 en función de u_0, obteniendo que cualquier punto en el que $u_0>0$, $u_1=\frac{u_0}{3}$ y $u_2=\frac{2u_0}{3}$ cumple las condiciones requeridas.

Podemos coger otro punto distinto, como el $(0,0)$ y ver qué pasa con las condiciones de FJ. En $\boldsymbol{x} = (0,0)$ tenemos que $I(\boldsymbol{x}) = \{3,4\}$. Nótese que las restricciones asociadas, expresadas como restricciones $g_i(\boldsymbol{x}) \leq 0$ serían $-x_1 \leq 0$ y $-x_2 \leq 0$. Por tanto, los gradientes asociados al punto $\boldsymbol{x} = (0,0)$ son

$$\nabla f(\boldsymbol{x}) = (-6,-4), \quad \nabla g_3(\boldsymbol{x}) = (-1,0), \quad \text{y} \quad \nabla g_4(\boldsymbol{x}) = (0,-1).$$

El sistema dado por las condiciones de FJ queda ahora

$$u_0 \begin{pmatrix} -6 \\ -4 \end{pmatrix} + u_3 \begin{pmatrix} -1 \\ 0 \end{pmatrix} + u_4 \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

La primera ecuación implica que $u_3 = -6u_0$ y la segunda que $u_4 = -4u_0$. Por tanto, si $u_0 > 0$ romperíamos la restricción de no negatividad para u_3 y u_4. Por otro lado, si $u_0 = 0$ tendríamos $u_3 = u_4 = 0$, pero el vector (u_0, u_1, u_2) debe ser no nulo. En cualquiera de los dos casos, la condición FD de FJ no se cumpliría en $(0,0)$ y, por tanto, no es un óptimo local. ◇

Ejemplo 5.4. Consideremos el siguiente problema de optimización:

$$\begin{array}{rl} \text{minimizar} & -x_1 \\ \text{sujeto a} & x_2 - (1-x_1)^3 \leq 0 \\ & x_2 \geq 0. \end{array}$$

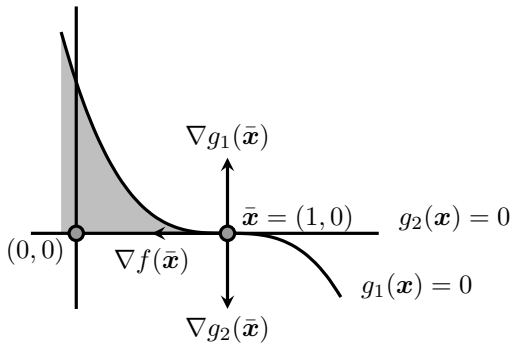


Figura 5.6: Región factible y curvas de nivel del problema del Ejemplo 5.4.

Este problema está representado en la Figura 5.6. Claramente, el óptimo del problema es el punto $\bar{\boldsymbol{x}} = (1,0)$. En este punto están activas las dos restricciones y, además, sus gradientes apuntan en direcciones opuestas $\nabla g_1(\bar{\boldsymbol{x}}) = -\nabla g_2(\bar{\boldsymbol{x}})$. Esto implica que $D_-^F = \emptyset$, pues ningún vector puede formar un ángulo de menos de 90° con un vector dado y con su opuesto. Por tanto, sabemos que las condiciones de FJ se cumplirán en este punto. Más concretamente, la condición FD en este caso queda:

$$u_0 \begin{pmatrix} -1 \\ 0 \end{pmatrix} + u_1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} + u_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

De la primera ecuación sacamos que, necesariamente, $u_0 = 0$. De la segunda sacamos que $u_1 = u_2$. Por tanto, cualquier vector de la forma $(0,k,k)$ con $k > 0$ servirá para verificar la condición FD de FJ. ◇

Obsérvese que en el ejemplo anterior tenemos algo parecido a lo que comentamos en el Ejemplo 5.2. La condición se cumple porque $D^F_- = \emptyset$, pero sin ninguna información de la función objetivo. En este caso coincide que es un óptimo local, pero como vimos en el Ejemplo 5.2, no siempre será el caso. A continuación mostramos un ejemplo de que esto puede ser problemático incluso para problemas convexos. Más aún, mostramos que incluso para problemas de programación lineal puedo tener puntos de FJ no son óptimos.

Ejemplo 5.5. Consideremos el siguiente problema de programación lineal:

$$\begin{array}{rl} \text{minimizar} & -x_2 \\ \text{sujeto a} & x_1 + 2x_2 \leq 6 \\ & x_1 - x_2 \leq 0 \\ & x_1 - x_2 \geq 0 \\ & x_1 \geq 0 \\ & x_2 \geq 0. \end{array}$$

La región factible de este problema aparece representada en la Figura 5.7. En este problema tenemos $g_1(\boldsymbol{x}) = x_1 + 2x_2 - 6$, $g_2(\boldsymbol{x}) = x_1 - x_2$, $g_3(\boldsymbol{x}) = -x_1 + x_2$, $g_4(\boldsymbol{x}) = -x_1$ y $g_5(\boldsymbol{x}) = -x_2$.

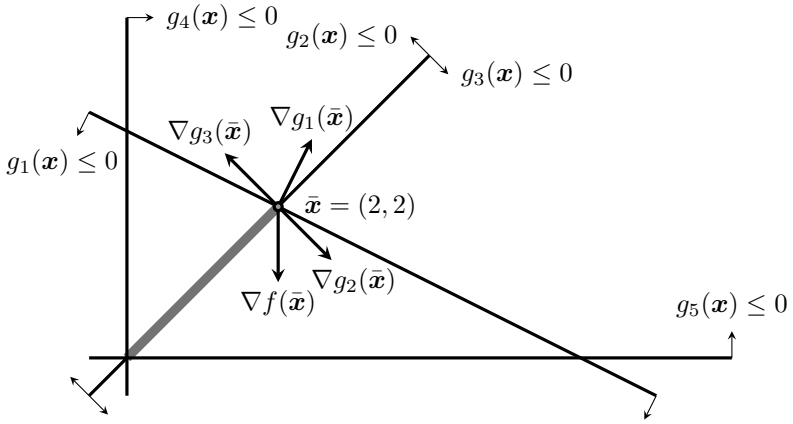


Figura 5.7: Ilustración de que las condiciones de FJ pueden no ser suficientes en problemas de programación lineal.

La región factible se corresponde con el segmento que une los puntos $(0, 0)$ y $(2, 2)$. Como el objetivo es minimizar $-x_2$, el único óptimo de este problema es el punto $\bar{\boldsymbol{x}} = (2, 2)$. Sin embargo, como en toda la región factible las restricciones 2 y 3 están activas y sus gradientes apuntan en direcciones opuestas, $\nabla g_2(\boldsymbol{x}) = -\nabla g_3(\boldsymbol{x})$, tendremos que, para todo punto factible, la condición de FD de FJ se cumplirá tomando, por ejemplo, $u_2 = u_3 = 1$ y todos los demás multiplicadores igual a 0.

Nuevamente, nos encontramos con que la condición necesaria de FJ selecciona demasiados puntos, con lo que está lejos de ser suficiente. ◇

A continuación presentamos, a modo de resumen, algunas observaciones que se puede extraer relativas a las condiciones de Fritz John:

- Se trata de condiciones muy manejables analíticamente.
- Se pueden convertir en condiciones suficientes, pero bajo supuestos adicionales de convexidad.
- Las condiciones de FJ son necesarias para optimalidad local, pero pueden cumplirse y que el punto no sea óptimo de ningún tipo.
- Cualquier punto en el que se anule ∇f o algún gradiente de una restricción activa en ese punto será automáticamente un punto FJ.
  - Siempre que $D_-^F = \emptyset$ tendré un punto Fritz John, independientemente de la función objetivo. Esto es un problema, pues puedo tener $D_-^F = \emptyset$ pero $D^F \neq \emptyset$, con lo que podría tener direcciones factibles. Si dichas direcciones también son de descenso, entonces no tengo optimalidad. Esto es lo que ilustramos con los ejemplos 5.2 y 5.5.
  - La aparición o no de los problemas anteriores puede ser sensible a formulaciones equivalentes del mismo problema (Ejemplo 5.2).
  - Las condiciones de KKT permiten mitigar estos problemas de las condiciones de FJ.

5.3.3 Condiciones de KKT para problemas sin restricciones de igualdad

Como hemos visto en la demostración del Teorema 5.5, el Teorema de Gordan nos asegura la equivalencia entre que un punto $\bar{\boldsymbol{x}}$ sea un punto Fritz John y la condición $D_-^D \cap D_-^F = \emptyset$. En particular, como hemos comentado repetidamente en el apartado anterior, esto se cumplirá siempre que $D_-^F = \emptyset$, independientemente de la función f.

Recordemos que $D_-^F = \emptyset$ quiere decir que no existe ningún $\boldsymbol{d} \in \mathbb{R}^n$ tal que $\nabla g_i(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} < 0$ para todo $i \in I(\bar{\boldsymbol{x}})$. Razonemos ahora de modo similar a la demostración del Teorema 5.5. Llamemos $\boldsymbol{A}$ a la matriz cuyas filas son dichos gradientes (traspuestos); pero a diferencia de entonces, ahora no incluimos el gradiente de f. La condición $D_-^F = \emptyset$ es equivalente a que el sistema $\boldsymbol{Ad} < \boldsymbol{0}$ no tenga solución. Por el Teorema de Gordan (Teorema 5.4), existirá un vector $\boldsymbol{y}$ con tantas componentes como filas tiene la matriz $\boldsymbol{A}$ tal que $\boldsymbol{y} \geq \boldsymbol{0}$, $\boldsymbol{y} \neq \boldsymbol{0}$ y $\boldsymbol{A}^\intercal \boldsymbol{y} = 0$. Por tanto, con bastante generalidad, dado un punto factible $\boldsymbol{x}$, el Teorema de Gordan nos permite decir lo siguiente:

$$D_-^F \neq \emptyset \iff \textit{el vector } \boldsymbol{0} \textit{ no se puede obtener mediante una combinación cónica no trivial de los gradientes } \nabla g_i(\boldsymbol{x}), \textit{ con } i \in I(\boldsymbol{x}).$$

Una condición como la anterior se conoce como "*constraint qualification*" o *condición de regularidad*. Las condiciones de Karush-Kuhn-Tucker se basan en incluir alguna condición de regularidad que asegure que $D_-^F \neq \emptyset$, y una gran cantidad de condiciones de regularidad han sido estudiadas en la literatura buscando un equilibrio entre la facilidad de manejo de dichas condiciones y la proximidad del conjunto de puntos resultantes al conjunto de óptimos locales. El estudio de condiciones de regularidad es un tema de investigación complejo en sí mismo, y va más allá de los objetivos de estas notas.[2] Es por ello que aquí nos centraremos en la condición

[2] El Capítulo 5 del libro de referencia Bazaraa y otros (2006) está íntegramente dedicado al estudio de "*constraint qualifications*".

de regularidad que impone que los gradientes $\nabla g_i(\boldsymbol{x})$, con $i \in I(\boldsymbol{x})$, sean linealmente independientes. Como toda combinación cónica es también una combinación lineal, esta condición implica la que mencionamos anteriormente sobre las combinaciones cónicas.

Una vez que tengamos una condición de regularidad para la cual $D_-^F \neq \emptyset$, tendremos asegurado que en todo punto FJ el multiplicador u_0 deberá ser distinto de cero (pues ninguna combinación de los gradientes $\nabla g_i(\boldsymbol{x})$ dará lugar al vector $\mathbf{0}$). Esto garantiza que la función objetivo siempre tenga algo que decir en la condición de optimalidad resultante.

Tras las reflexiones anteriores, ya podemos presentar las condiciones de Karush-Kuhn-Tucker, que fueron obtenidas independientemente por Karush (1939) y por Kuhn y Tucker (1951).

Teorema 5.6 (Condiciones necesarias de Karush-Kuhn-Tucker)**.** *Consideremos el siguiente problema de optimización:*

$$\begin{array}{rll} \text{minimizar} & f(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 & i = 1, \ldots, m. \end{array}$$

Supongamos que $\bar{\boldsymbol{x}}$ es un punto factible tal que

- *las funciones f y g_i con $i \in I(\bar{\boldsymbol{x}})$ son diferenciables en $\bar{\boldsymbol{x}}$,*
- *las funciones g_i con $i \notin I(\bar{\boldsymbol{x}})$ son continuas en $\bar{\boldsymbol{x}}$ y*
- *los vectores $\nabla g_i(\bar{\boldsymbol{x}})$, con $i \in I(\bar{\boldsymbol{x}})$, son linealmente independientes.*

Si $\bar{\boldsymbol{x}}$ es un mínimo local, entonces existen escalares únicos u_i para todo $i \in I(\bar{\boldsymbol{x}})$ tales que

$$\begin{aligned} \nabla f(\bar{\boldsymbol{x}}) + \sum_{i \in I(\bar{\boldsymbol{x}})} u_i \nabla g_i(\bar{\boldsymbol{x}}) &= \mathbf{0} \\ u_i &\geq 0 \quad \textit{para todo } i \in I(\bar{\boldsymbol{x}}). \end{aligned}$$

Si las funciones g_i con $i \notin I(\bar{\boldsymbol{x}})$ son también diferenciables en $\bar{\boldsymbol{x}}$, entonces las condiciones anteriores pueden ser escritas, de modo equivalente, como

$$\begin{aligned} \nabla f(\bar{\boldsymbol{x}}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{\boldsymbol{x}}) &= \mathbf{0} \\ u_i g_i(\bar{\boldsymbol{x}}) &= 0 \quad \textit{para todo } i \in \{1, \ldots, m\} \\ u_i &\geq 0 \quad \textit{para todo } i \in \{1, \ldots, m\}. \end{aligned}$$

Demostración. Por el Teorema 5.5 tenemos u_0 y $\bar{u}_i$ con $i \in I(\bar{\boldsymbol{x}})$ tales que

$$\begin{aligned} u_0 \nabla f(\bar{\boldsymbol{x}}) + \sum_{i \in I(\bar{\boldsymbol{x}})} \bar{u}_i \nabla g_i(\bar{\boldsymbol{x}}) &= \mathbf{0} \\ u_0, \bar{u}_i &\geq 0 \quad \text{para todo } i \in I(\bar{\boldsymbol{x}}) \\ (u_0, \bar{\boldsymbol{u}}_{I(\bar{\boldsymbol{x}})}) &\neq (0, \mathbf{0}). \end{aligned}$$

Además, la independencia de los vectores $\nabla g_i(\bar{\boldsymbol{x}})$, con $i \in I(\bar{\boldsymbol{x}})$ asegura que $u_0 > 0$. Por tanto, el resultado se sigue de tomar $u_i = \frac{\bar{u}_i}{u_0}$ con $i \in I(\bar{\boldsymbol{x}})$. Para la formulación equivalente basta definir $u_i = 0$ para aquellos $i \notin I(\bar{\boldsymbol{x}})$.

Por último, para probar la unicidad de los multiplicadores supongamos que

$$\nabla f(\bar{\boldsymbol{x}}) + \sum_{i \in I(\bar{\boldsymbol{x}})} u_i \nabla g_i(\bar{\boldsymbol{x}}) = \mathbf{0} \quad \text{y}$$
$$\nabla f(\bar{\boldsymbol{x}}) + \sum_{i \in I(\bar{\boldsymbol{x}})} \bar{u}_i \nabla g_i(\bar{\boldsymbol{x}}) = \mathbf{0}.$$

Entonces, $\mathbf{0} = \sum_{i \in I(\bar{\boldsymbol{x}})} u_i \nabla g_i(\bar{\boldsymbol{x}}) - \sum_{i \in I(\bar{\boldsymbol{x}})} \bar{u}_i \nabla g_i(\bar{\boldsymbol{x}}) = \sum_{i \in I(\bar{\boldsymbol{x}})} (u_i - \bar{u}_i) \nabla g_i(\bar{\boldsymbol{x}})$. La independencia los vectores $\nabla g_i(\bar{\boldsymbol{x}})$, con $i \in I(\bar{\boldsymbol{x}})$, implica que, para todo $i \in I(\bar{\boldsymbol{x}})$, $u_i = \bar{u}_i$. □

Al igual que con las condiciones de FJ, los u_i se llaman multiplicadores de Lagrange y las restricciones que componen las condiciones de KKT se suelen llamar factibilidad del dual (FD) y holguras complementarias (HC). Cualquier punto $\bar{\boldsymbol{x}}$ para el que existen multiplicadores de Lagrange $\bar{\boldsymbol{u}}$ tales que $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}})$ cumple las condiciones de KKT se denomina *punto KKT*.

Es habitual ver las condiciones de KKT expresadas matricialmente, apoyándose en la matriz $\nabla \boldsymbol{g}(\bar{\boldsymbol{x}})$, una matriz $m \times n$ que representa la matriz jacobiana de la función $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$, cuyas componentes son las funciones $g_i(\boldsymbol{x})$. En este caso, además de la condición de factibilidad de $\bar{\boldsymbol{x}}$, tendríamos

$$\begin{aligned} \nabla f(\bar{\boldsymbol{x}}) + \nabla \boldsymbol{g}(\bar{\boldsymbol{x}})^\intercal \boldsymbol{u} &= \mathbf{0} \\ \boldsymbol{u}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) &= 0 \\ \boldsymbol{u} &\geq \mathbf{0}. \end{aligned}$$

Al igual que pasaba en el caso de las condiciones de FJ, las condiciones necesarias de KKT también se pueden pasar a condiciones suficientes bajo unos mínimos supuestos de convexidad.

5.3.4 Condiciones de KKT: Discusión y ejemplos

Comenzaremos revisando los mismos ejemplos que estudiamos para las condiciones de FJ:

Ejemplo 5.3 En este primer ejemplo teníamos que $(2, 1)$ era un punto FJ y que los multiplicadores tenían que cumplir que $u_0 > 0$. Por tanto, es inmediato pasar a las condiciones KKT obteniendo $u_1 = \frac{1}{3}$ y $u_2 = \frac{2}{3}$.

Ejemplo 5.4 El óptimo de este problema era el punto $\bar{\boldsymbol{x}} = (1, 0)$, que vimos era un punto FJ. Sin embargo, tenemos que $\nabla g_1(\bar{\boldsymbol{x}}) = -\nabla g_2(\bar{\boldsymbol{x}})$, con lo que no cumple la condición de regularidad y no podemos invocar al Teorema 5.6 para asegurar que nos encontramos ante un punto KKT. De hecho, vimos que la condición de FD de FJ sólo se podía cumplir con $u_0 = 0$, nunca con $u_0 > 0$. Por tanto, estamos ante un óptimo local (y global) que no es un KKT.

Ejemplo 5.5 No es difícil comprobar que en este ejemplo, de todos los puntos FJ (que son todos los puntos factibles), sólo $\bar{\boldsymbol{x}} = (2, 2)$ es un punto KKT. La razón es que es el único en el que entra en juego el gradiente $\nabla g_1(\bar{\boldsymbol{x}})$ y que permitirá expresar $-\nabla f(\bar{\boldsymbol{x}})$ como combinación cónica de $\nabla g_1(\bar{\boldsymbol{x}})$ y $\nabla g_3(\bar{\boldsymbol{x}})$:

$$\nabla f(\bar{\boldsymbol{x}}) + u_1 \nabla g_1(\bar{\boldsymbol{x}}) + u_2 \nabla g_2(\bar{\boldsymbol{x}}) + u_3 \nabla g_3(\bar{\boldsymbol{x}}) = \begin{pmatrix} 0 \\ -1 \end{pmatrix} + u_1 \begin{pmatrix} 1 \\ 2 \end{pmatrix} + u_2 \begin{pmatrix} 1 \\ -1 \end{pmatrix} + u_3 \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

La condición KKT se cumple tomando $u_1 = \frac{1}{3}$, $u_2 = 0$ y $u_3 = \frac{1}{3}$. Nótese que en este caso tenemos que $\bar{\boldsymbol{x}} = (2, 2)$ es un óptimo que no cumple la condición de regularidad y, sin embargo, es un KKT.

Los dos últimos ejemplos ponen de manifiesto la principal diferencia entre las condiciones de FJ y las de KKT. A cambio de descartar muchos puntos FJ que no son óptimos locales, como en el caso del Ejemplo 5.5, las condiciones de regularidad detrás de la motivación de los puntos KKT también pueden dejar fuera óptimos locales y globales, como en el Ejemplo 5.4. En la práctica, se considera más positivo el primer efecto y por eso las condiciones de KKT son las condiciones de referencia en el diseño de algoritmos.

Por otro lado, el último ejemplo nos permite ver que un punto $\bar{\boldsymbol{x}}$ puede ser un punto KKT aunque en él no se cumpla ninguna condición de regularidad. Es decir, aunque no podamos apoyarnos en resultados como el Teorema 5.6 para asegurar que, efectivamente, ha de ser un punto KKT.

Como ya hemos comentado, el Ejemplo 5.4 muestra que, en aquellos puntos donde no se cumplen las condiciones de regularidad, las condiciones de KKT pueden no ser condiciones necesarias de optimalidad. El siguiente ejercicio nos pide que construyamos un ejemplo en el que esto suceda incluso bajo supuestos de convexidad.

••Ejercicio 5.4. Presenta un ejemplo de un problema de programación convexa en el que el óptimo global no es un punto KKT. ◁

5.3.5 Interpretación económica de los multiplicadores de Lagrange

Por último, es importante destacar que los multiplicadores de Lagrange tienen importante utilidad práctica en el análisis de sensibilidad, basada en una interpretación económica de los mismos. Esta interpretación está muy conectada con la dualidad y la discutiremos de nuevo en la Sección 5.4, donde primero probaremos la correspondencia entre multiplicadores de Lagrange y variables óptimas del problema dual (Teorema 5.21 y Teorema 5.23) y después formalizaremos la interpretación económica que a continuación describimos (Proposición 5.26). Al igual que sucedía en el caso de la programación lineal, las variables duales pueden ser interpretadas como precios sombra. Más concretamente, de la misma manera que el gradiente me dice cuánto varía la función objetivo por cada unidad que me mueva siguiendo su dirección, cada multiplicador de Lagrange me indica cuánto mejoraría la función objetivo por cada unidad en la que relaje la restricción asociada. En general, para problemas no lineales estas tasas de variación son únicamente locales, pero pueden ser muy informativas a la hora de tomar decisiones, ya que me dicen "cuánto me molesta" cada restricción en el punto KKT en cuestión.

Para ilustrar esto volveremos al problema del Ejemplo 5.3. Recordamos que el problema allí planteado era

$$\begin{aligned} \text{minimizar} \quad & (x_1-3)^2+(x_2-2)^2 \\ \text{sujeto a} \quad & x_1^2+x_2^2 \leq 5 \\ & x_1+2x_2 \leq 4 \\ & x_1 \geq 0 \\ & x_2 \geq 0. \end{aligned}$$

El óptimo de este problema es el punto $\bar{\boldsymbol{x}} = (2,1)$, en el que están activas las dos primeras restricciones. Vimos que los multiplicadores de Lagrange asociados son $u_1 = \frac{1}{3}$ y $u_2 = \frac{2}{3}$. Esto sugiere que relajar la restricción $g_2(\boldsymbol{x}) \leq 0$ debería tener un mayor impacto que relajar la restricción $g_1(\boldsymbol{x}) \leq 0$. Para comprobar esto vamos a hacer un pequeño análisis de sensibilidad, cambiando los lados derechos de las restricciones 1 y 2, que llamaremos b_1 y b_2, respectivamente.[3]

Caso base: $b_1 = 5$ **y** $b_2 = 4$**.** El óptimo es $\bar{\boldsymbol{x}} = (2,1)$ y $f(\bar{\boldsymbol{x}}) = 2$.

Variación 1: $b_1 = 5.1$ **y** $b_2 = 4$**.** El nuevo óptimo es $\bar{\boldsymbol{x}} = (2.033, 0.984)$ y $f(\bar{\boldsymbol{x}}) = 1.968$. La mejora en la función objetivo es 0.032, mientras que lo predicho por el multiplicador de Lagrange sería $u_1 \cdot 0.1 = \frac{1}{3}0.1 \approx 0.033$. Por tanto vemos que, efectivamente, el multiplicador de Lagrange nos da una estimación bastante fiable de lo que podemos mejorar localmente si relajamos la restricción asociada (o lo que podemos empeorar si la endurecemos).

Variación 2: $b_1 = 5$ **y** $b_2 = 4.1$**.** Hemos cambiado la segunda restricción, que anticipaba una tasa de variación dos veces mayor. En este caso el nuevo óptimo es $\bar{\boldsymbol{x}} = (1.965, 1.068)$ y $f(\bar{\boldsymbol{x}}) = 1.941$. La mejora en la función objetivo es 0.059, mientras que lo predicho por el multiplicador de Lagrange sería $u_2 \cdot 0.1 = \frac{2}{3}0.1 \approx 0.067$. La aproximación vuelve a ser aceptable.

Variación 3: $b_1 = 6$ **y** $b_2 = 4$**.** Ahora hemos incrementado b_1 en un 20 %, que ya es un cambio suficientemente grande como para que la tasa de variación predicha por el multiplicador deje de ser fiable. En este caso el nuevo óptimo es $\bar{\boldsymbol{x}} = (2.297, 0.852)$ y $f(\bar{\boldsymbol{x}}) = 1.813$. La mejora en la función objetivo es 0.187, mientras que lo predicho por el multiplicador de Lagrange sería $u_1 \cdot 1 = \frac{1}{3} \approx 0.333$. La aproximación no es mala, pero es bastante peor que antes.

Variación 4: $b_1 = 5$ **y** $b_2 = 5$**.** Ahora hemos incrementado b_2 en un 25 %. El nuevo óptimo es $\bar{\boldsymbol{x}} = (1.861, 1.240)$ y $f(\bar{\boldsymbol{x}}) = 1.875$. La mejora en la función objetivo es 0.125, mientras que lo predicho por el multiplicador de Lagrange sería $u_2 \cdot 1 = \frac{2}{3} \approx 0.667$. Vemos que en este caso la aproximación es bastante mala.

Viendo los dos últimos casos tenemos que el cambio de b_2 de 4 a 5 da lugar a una mejora menor que el cambio de b_1 de 5 a 6, a pesar de que el multiplicador de Lagrange de la segunda restricción es mayor. Esto ilustra perfectamente el carácter local de la interpretación de los multiplicadores de Lagrange como precios sombra.

[3] Las variaciones aquí presentadas han sido resueltas mediante un simple script en AMPL.

5.3.6 Condiciones de KKT para problemas generales

En esta sección vamos a ver cómo obtener las condiciones KKT en problemas en los que puede haber tanto restricciones de desigualdad como de igualdad. Las condiciones de Fritz John también se pueden generalizar para este caso, siendo dicha generalización debida a Mangasarian y Fromovitz (1967), pero como dichas condiciones las usamos principalmente como vehículo para motivar las condiciones de Karush-Kuhn-Tucker, en esta sección presentaremos directamente estas últimas.

Inicialmente, uno podría pensar que la extensión a problemas con restricciones de igualdad es inmediata, sin más que descomponer toda restricción de la forma $h_j(\boldsymbol{x}) = 0$ en dos restricciones de desigualdad $g_{j_1}(\boldsymbol{x}) = h_j(\boldsymbol{x}) \leq 0$ y $g_{j_2}(\boldsymbol{x}) = -h_j(\boldsymbol{x}) \leq 0$ y aplicar entonces el Teorema 5.6. Esto resultaría en que habría un multiplicador $u_{j_1} \geq 0$ asociado a $\nabla g_{j_1}(\bar{\boldsymbol{x}}) = \nabla h_j(\bar{\boldsymbol{x}})$ y otro $u_{j_2} \geq 0$ asociado a $\nabla g_{j_2}(\bar{\boldsymbol{x}}) = -\nabla h_j(\bar{\boldsymbol{x}})$. Juntando ambos nos quedaría un sumando de la forma

$$(u_{j_1} - u_{j_2})\, \nabla h_j(\bar{\boldsymbol{x}}),$$

lo que es equivalente a tener un único multiplicador $u_j = u_{j_1} - u_{j_2}$ asociado a la restricción de igualdad, pero que puede tener cualquier signo (ya que es la diferencia de dos números no negativos).

Aunque la intuición que acabamos de mostrar no va desencaminada y es, en cierto modo, lo que recoge el siguiente resultado, la demostración formal requiere mucho más cuidado. La razón es que la transformación de cada restricción de igualdad en dos restricciones de desigualdad resulta en problemas que no cumplen la condición de regularidad (pues $\nabla g_{j_1}(\bar{\boldsymbol{x}}) = -\nabla g_{j_2}(\bar{\boldsymbol{x}})$, con lo que no son independientes).

Teorema 5.7 (Condiciones necesarias de Karush-Kuhn-Tucker)**.** *Consideremos el siguiente problema de optimización:*

$$\begin{array}{rll} \text{minimizar} & f(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 & i = 1, \ldots, m \\ & h_j(\boldsymbol{x}) = 0 & j = 1, \ldots, l. \end{array}$$

Supongamos que $\bar{\boldsymbol{x}}$ es un punto factible tal que

- *las funciones f y g_i con $i \in I(\bar{\boldsymbol{x}})$ son diferenciables en $\bar{\boldsymbol{x}}$,*
- *las funciones g_i con $i \notin I(\bar{\boldsymbol{x}})$ son continuas en $\bar{\boldsymbol{x}}$,*
- *las funciones $h_j(\bar{\boldsymbol{x}})$ con $j \in \{1, \ldots, l\}$ son continuamente diferenciables.*
- *los vectores $\nabla g_i(\bar{\boldsymbol{x}})$, con $i \in I(\bar{\boldsymbol{x}})$, y $\nabla h_j(\bar{\boldsymbol{x}})$ con $j \in \{1, \ldots, l\}$ son linealmente independientes.*

Si $\bar{\boldsymbol{x}}$ es un mínimo local, entonces existen escalares únicos u_i para todo $i \in I(\bar{\boldsymbol{x}}) = I$ y v_j para todo $j \in \{1, \ldots, l\}$ tales que

$$\nabla f(\bar{\boldsymbol{x}}) + \sum_{i \in I} u_i \nabla g_i(\bar{\boldsymbol{x}}) + \sum_{j=1}^{l} v_j \nabla h_j(\bar{\boldsymbol{x}}) = \mathbf{0}$$

$$u_i \geq 0 \quad \textit{para todo } i \in I.$$

Si las funciones g_i con $i \notin I(\bar{\boldsymbol{x}})$ son también diferenciables en $\bar{\boldsymbol{x}}$, entonces las condiciones anteriores pueden ser escritas, de modo equivalente, como

$$\nabla f(\bar{\boldsymbol{x}}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{\boldsymbol{x}}) + \sum_{j=1}^{l} v_j \nabla h_j(\bar{\boldsymbol{x}}) = \mathbf{0}$$
$$u_i g_i(\bar{\boldsymbol{x}}) = 0 \quad \text{para todo } i \in \{1, \dots, m\}$$
$$u_i \geq 0 \quad \text{para todo } i \in \{1, \dots, m\}.$$

La demostración de este resultado es significativamente más complicada que la demostración del resultado sin restricciones de igualdad (Teorema 5.6). De hecho, las demostraciones más comunes requieren trabajar con sistemas de ecuaciones diferenciales los cuales, aun sin ser muy complejos, incrementan la dificultad de la demostración. Una de estas demostraciones puede verse en la Sección 4.3 del libro Bazaraa y otros (2006).

•••**Ejercicio 5.5.** Presenta una demostración del Teorema 5.7. ◁

5.3.7 Condiciones KKT de segundo orden

En este último apartado relativo a las condiciones de KKT presentamos las condiciones de segundo orden, cuyo papel es análogo al desempeñado por las condiciones de segundo orden en el caso de la optimización sin restricciones y que fueron inicialmente estudiadas por McCormick (1967) y posteriormente refinadas en Ben-Tal (1980) y Fletcher (1987). En el estudio de problemas de optimización sin restricciones tenemos la condición necesaria de optimalidad $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$, basada en en gradiente, y con la ayuda de las derivadas de segundo orden, basadas en la hessiana, podemos conseguir condiciones suficientes (sin necesidad de apoyarse en supuestos de convexidad, que es la otra vía para garantizar suficiencia).

Para introducir las condiciones de segundo orden es útil introducir el concepto de *función lagrangiana*, que será de gran importancia en la Sección 5.4 para desarrollar la teoría de dualidad y también es de gran utilidad en el desarrollo de algunos algoritmos para problemas con restricciones como el método del lagrangiano aumentado que veremos en la Sección 7.3.

Definición 5.1. Dado problema de optimización de la forma

$$\begin{array}{rll} \text{minimizar} & f(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 & i = 1, \dots, m \\ & h_j(\boldsymbol{x}) = 0 & j = 1, \dots, l, \end{array}$$

donde todas las funciones son dos veces diferenciables, la *función lagrangiana* se define como

$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} u_i g_i(\boldsymbol{x}) + \sum_{j=1}^{l} v_j h_j(\boldsymbol{x}).$$

Esta función está muy relacionada con las condiciones de KKT pues, dado $\bar{\boldsymbol{x}}$ un punto KKT con multiplicadores $\bar{\boldsymbol{u}} \geq \mathbf{0}$ y $\bar{\boldsymbol{v}}$ podemos hablar de la función lagrangiana restringida al primal:

$$L^P(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \bar{u}_i g_i(\boldsymbol{x}) + \sum_{j=1}^{l} \bar{v}_j h_j(\boldsymbol{x}).$$

Ahora, podemos ver que la condición de factibilidad del dual de KKT en el punto $\bar{\boldsymbol{x}}$ es equivalente a que $\nabla L^P(\bar{\boldsymbol{x}}) = \mathbf{0}$. Además, es fácil ver que, para todo punto factible $\boldsymbol{x}$, $L^P(\boldsymbol{x}) \leq f(\boldsymbol{x})$, mientras que $L^P(\bar{\boldsymbol{x}}) = f(\bar{\boldsymbol{x}})$ (por la condición de holguras complementarias). Simplemente basta notar que en un punto factible se anulan las funciones h_j y que las g_i serán no positivas lo cual, unido a la no negatividad de los $\bar{u}_i$ da lugar a la propiedad mencionada.

Por lo tanto, si $\bar{\boldsymbol{x}}$ es un mínimo local para L^P, tendremos que existirá un entorno de $\bar{\boldsymbol{x}}$ tal que, para todo $\boldsymbol{x}$ en dicho entorno, $f(\bar{\boldsymbol{x}}) = L^P(\bar{\boldsymbol{x}}) \leq L^P(\boldsymbol{x}) \leq f(\boldsymbol{x})$. Entonces, $\bar{\boldsymbol{x}}$ será un mínimo local del problema de optimización de partida. Como es habitual, bajo convexidad ambos mínimos serían globales. En particular, la Proposición 4.3 nos asegurará lo siguiente. Si $\bar{\boldsymbol{x}}$ es un punto KKT ($\nabla L^P(\bar{\boldsymbol{x}}) = \mathbf{0}$) y, además, la hessiana de la función $L^P(\bar{\boldsymbol{x}})$, $\nabla^2 L^P(\bar{\boldsymbol{x}})$, es definida positiva, entonces $\bar{\boldsymbol{x}}$ es un mínimo local estricto de $L^P(\bar{\boldsymbol{x}})$ y, por tanto, también del problema de optimización de partida.

A continuación presentamos la condición de segundo orden de KKT, que está muy relacionada con la propiedad de que $\nabla^2 L^P(\bar{\boldsymbol{x}})$ sea definida positiva, pero siendo mucho más débil, ya que sólo requiere que $\boldsymbol{d}^\intercal \nabla^2 L^P(\bar{\boldsymbol{x}})\boldsymbol{d}$ sea positivo para un subconjunto específico de direcciones y no para todo $\boldsymbol{d} \in \mathbb{R}^n$, como es el caso en la condición que enunciamos arriba.

Previamente necesitamos dividir en dos el conjunto de restricciones activas en un punto KKT $\bar{\boldsymbol{x}}$, $I(\bar{\boldsymbol{x}}) = \{i : g_i(\bar{\boldsymbol{x}}) = 0\}$. Más concretamente, si los multiplicadores asociados a $\bar{\boldsymbol{x}}$ vienen dados por $\boldsymbol{u}$ y $\boldsymbol{v}$, definimos

Restricciones fuertemente activas. $I^+(\bar{\boldsymbol{x}}) = \{i \in I(\bar{\boldsymbol{x}}) : \bar{u}_i > 0\}$.

Restricciones débilmente activas. $I^0(\bar{\boldsymbol{x}}) = \{i \in I(\bar{\boldsymbol{x}}) : \bar{u}_i = 0\}$.

En las condiciones de KKT de segundo orden trabajaremos con la hessiana de la función lagrangiana primal para un punto $\bar{\boldsymbol{x}}$ con multiplicadores dados por $\bar{\boldsymbol{u}}$ y $\bar{\boldsymbol{v}}$:

$$\nabla^2 L^P(\bar{\boldsymbol{x}}) = \nabla^2 f(\bar{\boldsymbol{x}}) + \sum_{i \in I(\bar{\boldsymbol{x}})} \bar{u}_i \nabla^2 g_i(\bar{\boldsymbol{x}}) + \sum_{j=1}^{l} \bar{v}_i \nabla^2 h_j(\bar{\boldsymbol{x}}),$$

donde $\nabla^2 f(\bar{\boldsymbol{x}})$, $\nabla^2 g_i(\bar{\boldsymbol{x}})$ y $\nabla^2 h_j(\bar{\boldsymbol{x}})$ son las hessianas de las funciones que definen el problema evaluadas en el punto $\bar{\boldsymbol{x}}$.

Además, trabajaremos con el cono de direcciones dado por

$$\begin{aligned} C(\bar{\boldsymbol{x}}) = \Big\{ \boldsymbol{d} \in \mathbb{R}^n : \boldsymbol{d} \neq \mathbf{0} \text{ y } & \nabla g_i(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} = 0 \quad \forall i \in I^+(\bar{\boldsymbol{x}}), \\ & \nabla g_i(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} \leq 0 \quad \forall i \in I^0(\bar{\boldsymbol{x}}), \text{ y} \\ & \nabla h_j(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} = 0 \quad \forall j \in \{1, \ldots, l\} \Big\}. \end{aligned}$$

Teorema 5.8 (Condiciones suficientes de Karush-Kuhn-Tucker de segundo orden)**.** *Consideremos el siguiente problema de optimización:*

$$\begin{aligned} \text{minimizar} \quad & f(\boldsymbol{x}) \\ \text{sujeto a} \quad & g_i(\boldsymbol{x}) \leq 0 \qquad i = 1, \ldots, m \\ & h_j(\boldsymbol{x}) = 0 \qquad j = 1, \ldots, l. \end{aligned}$$

Supongamos que las funciones f, g_i con $i \in \{1,\dots,m\}$ y h_j con $j \in \{1,\dots,l\}$ son dos veces diferenciables. Supongamos que $\bar{\boldsymbol{x}}$ es un punto KKT con multiplicadores de Lagrange dados por $\bar{\boldsymbol{u}}$ y $\bar{\boldsymbol{v}}$. Dada la hessiana de la función lagrangiana primal

$$\nabla^2 L^P(\bar{\boldsymbol{x}}) = \nabla^2 f(\bar{\boldsymbol{x}}) + \sum_{i \in I(\bar{\boldsymbol{x}})} \bar{u}_i \nabla^2 g_i(\bar{\boldsymbol{x}}) + \sum_{j=1}^{l} \bar{v}_i \nabla^2 h_j(\bar{\boldsymbol{x}}).$$

Entonces, si $\boldsymbol{d}^\intercal \nabla^2 L^P(\bar{\boldsymbol{x}})\boldsymbol{d} > 0$ para todo $\boldsymbol{d} \in C(\bar{\boldsymbol{x}})$, el punto $\bar{\boldsymbol{x}}$ es un mínimo local estricto.

Nótese que, en el caso de que tengamos un problema sin restricciones, el enunciado del Teorema 5.8 nos dice que si $\nabla f(\bar{\boldsymbol{x}}) = \boldsymbol{0}$ y $\boldsymbol{H}(\bar{\boldsymbol{x}})$ es definida positiva entonces $\bar{\boldsymbol{x}}$ es un mínimo local, que es justamente el enunciado de la Proposición 4.3, que por tanto es un caso particular del teorema. De hecho, la demostración del Teorema 5.8 sigue una línea parecida a la de dicha proposición. Los cálculos necesarios, que son algo más laboriosos en este caso son el objetivo del siguiente ejercicio y pueden encontrarse nuevamente en el libro Bazaraa y otros (2006).

•••**Ejercicio 5.6.** Presenta una demostración del Teorema 5.8. ◁

El siguiente corolario contiene una condición suficiente para que $C(\bar{\boldsymbol{x}}) = \emptyset$, que se traduce inmediatamente en una condición suficiente de optimalidad local.

Corolario 5.9. *En las condiciones del Teorema 5.8, si los vectores $\nabla g_i(\bar{\boldsymbol{x}})$, con $i \in I^+(\bar{\boldsymbol{x}})$, y $\nabla h_j(\bar{\boldsymbol{x}})$, con $j \in \{1,\dots,l\}$, contienen un subconjunto de n vectores linealmente independientes, entonces $\bar{\boldsymbol{x}}$ es un mínimo local estricto.*

Demostración. En los supuestos del corolario claramente $C(\bar{\boldsymbol{x}}) = \emptyset$, pues el vector $\boldsymbol{0}$ es el único que puede ser ortogonal a todos los vectores de una base. Por tanto, la condición del Teorema 5.8 se cumple trivialmente. □

Para terminar esta sección vamos a presentar las condiciones necesarias de Karush-Kuhn-Tucker de segundo orden.

Teorema 5.10 (Condiciones necesarias de Karush-Kuhn-Tucker de segundo orden)**.** *Consideremos el siguiente problema de optimización:*

$$\begin{array}{rll} \text{minimizar} & f(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 & i = 1,\dots,m \\ & h_j(\boldsymbol{x}) = 0 & j = 1,\dots,l. \end{array}$$

Supongamos que se cumplen las condiciones necesarias de KKT de primer orden y que las funciones f, g_i con $i \in \{1,\dots,m\}$ y h_j con $j \in \{1,\dots,l\}$ son dos veces diferenciables. Supongamos que $\bar{\boldsymbol{x}}$ es un mínimo local, que entonces también es KKT y tendrá multiplicadores de Lagrange dados por $\bar{\boldsymbol{u}}$ y $\bar{\boldsymbol{v}}$. Sea $\nabla^2 L^P(\bar{\boldsymbol{x}})$ la hessiana de la función lagrangiana primal,

$$\nabla^2 L^P(\bar{\boldsymbol{x}}) = \nabla^2 f(\bar{\boldsymbol{x}}) + \sum_{i \in I(\bar{\boldsymbol{x}})} \bar{u}_i \nabla^2 g_i(\bar{\boldsymbol{x}}) + \sum_{j=1}^{l} \bar{v}_i \nabla^2 h_j(\bar{\boldsymbol{x}}).$$

Entonces, $\boldsymbol{d}^\intercal \nabla^2 L^P(\bar{\boldsymbol{x}})\boldsymbol{d} \geq 0$ para todo $\boldsymbol{d} \in C(\bar{\boldsymbol{x}})$.

•••**Ejercicio 5.7.** Presenta una demostración del Teorema 5.10. ◁

En este caso, si tenemos un problema de optimización sin restricciones, el enunciado del Teorema 5.10 nos dice que, si tenemos un mínimo local, entonces $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$ y $\boldsymbol{H}(\bar{\boldsymbol{x}})$ es semidefinida positiva. Este es justamente el encunciado de la Proposición 4.2, que por tanto es un caso particular del teorema.

5.3.8 Condiciones de KKT y problemas de programación lineal

En este apartado vamos a explorar las implicaciones de las condiciones de KKT en el caso especial de los problemas de programación lineal. En particular, el problema ilustrado en el Ejemplo 5.5, en el que vimos que las condiciones necesarias de FJ no son suficientes ni siquiera en el caso de problemas lineales, no afecta a las condiciones de KKT.

Antes de continuar, es conveniente recordar un par de conceptos básicos de dualidad en programación lineal, que serán de utilidad en esta sección. Recordemos la formulación de un problema de programación lineal en forma estándar y su dual:

Problema primal P		**Problema dual D**	
minimizar	$\boldsymbol{c}^\top \boldsymbol{x}$	maximizar	$\boldsymbol{b}^\top \boldsymbol{w}$
sujeto a	$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$	sujeto a	$\boldsymbol{A}^\top \boldsymbol{w} \leq \boldsymbol{c}$.
	$\boldsymbol{x} \geq \mathbf{0}$		

Supongamos ahora que P tiene n variables y m restricciones, es decir, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Denotemos las filas de $\boldsymbol{A}$ por $\boldsymbol{A}_i^f$ y las columnas por $\boldsymbol{A}_k^c$. El *Teorema de holguras complementarias* nos dice que si $\boldsymbol{x}$ es una solución factible de P y $\boldsymbol{w}$ es una solución factible de D, entonces $\boldsymbol{x}$ y $\boldsymbol{w}$ son un par primal-dual óptimo si y sólo si

$$\begin{aligned} w_i(\boldsymbol{A}_i^f \boldsymbol{x} - b_i) = 0 \quad & \text{para todo } i \in \{1,\dots,m\} \text{ y} \\ (c_j - \boldsymbol{w}^\top \boldsymbol{A}_j^c)x_j = 0 \quad & \text{para todo } j \in \{1,\dots,n\}. \end{aligned}$$

Aquellas restricciones del primal en las que haya holgura ($\boldsymbol{A}_j^f \boldsymbol{x} - b_j \neq 0$) tendrán variable dual asociada nula ($w_j = 0$). Análogamente, aquellas restricciones del dual en las que haya holgura ($c_i - \boldsymbol{w}^\top \boldsymbol{A}_i^c \neq 0$) tendrán variable primal asociada nula ($x_i = 0$).

En el caso que nos ocupa, en el que el problema primal ha sido expresado en forma estándar, tenemos que, por ser $\boldsymbol{x}$ factible, $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, con lo que el primer conjunto de condiciones de holguras complementarias se cumplirá trivialmente y únicamente tenemos que preocuparnos de las igualdades $(c_i - \boldsymbol{w}^\top \boldsymbol{A}_i^c)x_i = 0$.

Tras este pequeño recordatorio estamos en condiciones de presentar el resultado que relaciona las condiciones de Karush-Kuhn-Tucker con la optimalidad en los problemas P y D.

Proposición 5.11. *Dado un problema de programación lineal en forma estándar, entonces $\bar{\boldsymbol{x}}$ es un punto KKT con multiplicadores de Lagrange $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$, con $\bar{\boldsymbol{u}} \geq \mathbf{0}$ asociado a las restricciones $-\boldsymbol{x} \leq \mathbf{0}$ (o $\boldsymbol{x} \geq 0$) si y sólo si $\bar{\boldsymbol{x}}$ es óptimo de P y $\bar{\boldsymbol{v}}$ es óptimo de D.*

Demostración. Si expresamos el problema P poniendo primero las restricciones de desigualdad en forma $g_i(\boldsymbol{x}) \leq 0$ y cambiando de signo las restricciones de igualdad nos queda:

$$\begin{array}{rl} \text{minimizar} & \boldsymbol{c}^\intercal \boldsymbol{x} \\ \text{sujeto a} & -\boldsymbol{x} \leq \boldsymbol{0} \\ & -\boldsymbol{A}\boldsymbol{x} = -\boldsymbol{b}. \end{array}$$

Recordemos que $\boldsymbol{e}^i$ denota al i-ésimo vector de la base canónica. Entonces, un punto factible $\bar{\boldsymbol{x}}$ es un KKT si existen multiplicadores $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ con $\bar{\boldsymbol{u}} \geq \boldsymbol{0}$ tales que

$$\begin{array}{lrcl} \text{(FD)} & \boldsymbol{c} + \sum_{i=1}^n \bar{u}_i(-\boldsymbol{e}^i) - \sum_{j=1}^m \bar{v}_j \boldsymbol{A}_j^f & = & \boldsymbol{0} \\ \text{(HC)} & \bar{u}_i(-\bar{x}_i) & = & 0 \quad \text{para todo } i \in \{1, \ldots, n\}. \end{array}$$

Estas ecuaciones pueden escribirse de modo equivalente como

$$\begin{array}{lrcl} \text{(FD)} & \boldsymbol{A}^\intercal \bar{\boldsymbol{v}} + \bar{\boldsymbol{u}} & = & \boldsymbol{c} \\ \text{(HC)} & \bar{u}_i \bar{x}_i & = & 0 \quad \text{para todo } i \in \{1, \ldots, n\}. \end{array}$$

Como $\bar{\boldsymbol{u}} \geq \boldsymbol{0}$, la llamada condición de factibilidad del dual de KKT (FD), efectivamente es equivalente a que $\boldsymbol{A}^\intercal \bar{\boldsymbol{v}} \leq \boldsymbol{c}$, es decir, $\bar{\boldsymbol{v}}$ es una solución factible de D.

Además, como $\bar{u}_i = c_i - (\boldsymbol{A}^\intercal \bar{\boldsymbol{v}})_i = (c_i - \bar{\boldsymbol{v}}^\intercal \boldsymbol{A}_i^c)$, tenemos que las llamadas condiciones de holguras complementarias de KKT (HC) efectivamente implican que $\bar{\boldsymbol{v}}$ cumple las condiciones de holguras complementarias pues, para todo $i \in \{1, \ldots, n\}$,

$$\bar{u}_i \bar{x}_i = 0 \quad \text{si y sólo si} \quad (c_i - \bar{\boldsymbol{v}}^\intercal \boldsymbol{A}_i^c)\bar{x}_i = 0.$$

Nótese que las condiciones $\bar{v}_j(\boldsymbol{A}_j^{\mathrm{f}} \boldsymbol{x} - b_j) = 0$ para todo $j \in \{1, \ldots, m\}$ se cumplen trivialmente por estar el problema P en forma estándar.

Por tanto, tenemos probada la equivalencia entre las condiciones KKT asociadas al punto $\bar{\boldsymbol{x}}$ con multiplicadores $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ y la optimalidad del par primal-dual dado por $\bar{\boldsymbol{x}}$ y $\bar{\boldsymbol{v}}$. □

Esta equivalencia entre condiciones de KKT y optimalidad para problemas de programación lineal es importante para el diseño de ciertas familias de algoritmos de optimización. Más concretamente, algoritmos que se basan en resolver iterativamente aproximaciones lineales del problema no lineal en cuestión, que se fundamentan en el siguiente resultado.

Teorema 5.12. *Supongamos que las funciones f y g_i con $i \in \{1, \ldots, m\}$ son diferenciables. Consideremos el problema de optimización que definen y su aproximación lineal de primer orden en un punto factible $\bar{\boldsymbol{x}}$, LP($\bar{\boldsymbol{x}}$):*

	Problema P		*Problema LP($\bar{\boldsymbol{x}}$)*
minimizar	$f(\boldsymbol{x})$	minimizar	$f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}})$
sujeto a	$g_i(\boldsymbol{x}) \leq 0 \quad i = 1, \ldots, m$	sujeto a	$g_i(\bar{\boldsymbol{x}}) + \nabla g_i(\bar{\boldsymbol{x}})^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq 0 \quad i = 1, \ldots, m.$

Entonces, $\bar{\boldsymbol{x}}$ es un punto KKT de P si y sólo si $\bar{\boldsymbol{x}}$ es un óptimo de LP($\bar{\boldsymbol{x}}$).

Demostración. Claramente, $\bar{\boldsymbol{x}}$ es una solución factible de LP($\bar{\boldsymbol{x}}$), pues los términos $(\boldsymbol{x} - \bar{\boldsymbol{x}})$ se anulan en $\bar{\boldsymbol{x}}$ y $\bar{\boldsymbol{x}}$ es factible del problema original. Si obviamos los términos constantes de la función objetivo, $f(\bar{\boldsymbol{x}})$ y $\nabla f(\bar{\boldsymbol{x}})^\intercal \bar{\boldsymbol{x}}$, LP($\bar{\boldsymbol{x}}$) y su dual, DLP($\bar{\boldsymbol{x}}$) se pueden expresar como

$$\begin{aligned} \text{maximizar} \quad & -\nabla f(\bar{\boldsymbol{x}})^\intercal \boldsymbol{x} \\ \text{sujeto a} \quad & \nabla g_i(\bar{\boldsymbol{x}})^\intercal \boldsymbol{x} \leq \nabla g_i(\bar{\boldsymbol{x}})^\intercal \bar{\boldsymbol{x}} - g_i(\bar{\boldsymbol{x}}) \quad i = 1, \ldots, m \end{aligned}$$

y

$$\begin{aligned} \text{minimizar} \quad & \sum_{i=1}^{m} u_i \Big(\nabla g_i(\bar{\boldsymbol{x}})^\intercal \bar{\boldsymbol{x}} - g_i(\bar{\boldsymbol{x}}) \Big) \\ \text{sujeto a} \quad & \sum_{i=1}^{m} u_i \nabla g_i(\bar{\boldsymbol{x}}) = -\nabla f(\bar{\boldsymbol{x}}) \qquad i = 1, \ldots, m. \\ & \boldsymbol{u} \geq \boldsymbol{0} \end{aligned}$$

En este caso, el Teorema de holguras complementarias nos dice que $\bar{\boldsymbol{x}}$ es una solución óptima de LP($\bar{\boldsymbol{x}}$) si y sólo si existe $\bar{\boldsymbol{u}}$ solución factible de DLP($\bar{\boldsymbol{x}}$) que cumpla las condiciones de holguras complementarias. Pero la factibilidad de DLP($\bar{\boldsymbol{x}}$) es justamente la condición FD de KKT. Por otro lado, las condiciones de holguras complementarias se pueden expresar como:

$$\bar{\boldsymbol{u}}_i \Big(\nabla g_i(\bar{\boldsymbol{x}})^\intercal \bar{\boldsymbol{x}} - \nabla g_i(\bar{\boldsymbol{x}})^\intercal \bar{\boldsymbol{x}} + g_i(\bar{\boldsymbol{x}}) \Big) = \bar{\boldsymbol{u}}_i g_i(\bar{\boldsymbol{x}}) = 0, \quad i = 1, \ldots, m.$$

Nótese que las condiciones de holguras complementarias asociadas a las restricciones del dual no son necesarias pues dichas restricciones son de igualdad. Ahora bien, las condiciones $\bar{\boldsymbol{u}}_i g_i(\bar{\boldsymbol{x}}) = 0$ son justamente las condiciones HC de KKT. □

5.4 Dualidad

En esta sección estudiaremos formalmente los resultados principales de dualidad en problemas de optimización no lineal y discutiremos las implicaciones de los mismos a la hora de resolver problemas de optimización, con especial atención a las implicaciones en el desarrollo de algoritmos. Se tratará de un desarrollo principalmente conceptual, aunque se intentará incluir ejemplos que ayuden a entender las intuiciones detrás de los distintos resultados. Además, en la Sección 5.5 se presentarán tres aplicaciones de la dualidad y los multiplicadores de Lagrange en el estudio y resolución de distintos problemas de optimización.

5.4.1 Recordatorio de dualidad en programación lineal

Antes de presentar los conceptos de dualidad en optimización no lineal, es conveniente presentar un rápido recordatorio de los conceptos y resultados básicos en optimización lineal.[4] Después veremos que muchos de los resultados e ideas subyacentes a la dualidad en optimización lineal tienen sus análogos en el caso no lineal.

Supondremos que el problema primal tiene n variables y m restricciones, con lo que la matriz de restricciones será $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, el vector de costes $\boldsymbol{c} \in \mathbb{R}^n$ y el de lados derechos

[4] Para un análisis más profundo de la dualidad en programación lineal, así como para las demostraciones de los resultados aquí presentados, referirse a cualquier libro de programación lineal, como Bazaraa y otros (2009).

$\boldsymbol{b} \in \mathbb{R}^m$. Las variables del primal se corresponderán con $\boldsymbol{x} \in \mathbb{R}^n$ y las del dual con $\boldsymbol{w} \in \mathbb{R}^m$. Denotamos las filas de $\boldsymbol{A}$ por $\boldsymbol{A}_i^{\mathrm{f}}$ y las columnas por $\boldsymbol{A}_j^{\mathrm{c}}$.

Comenzamos recordando las formulaciones del problema dual cuando el primal está en forma estándar y cuando está en forma canónica:

(forma estándar)

Problema primal P		**Problema dual D**	
minimizar	$\boldsymbol{c}^\intercal \boldsymbol{x}$	maximizar	$\boldsymbol{b}^\intercal \boldsymbol{w}$
sujeto a	$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$	sujeto a	$\boldsymbol{A}^\intercal \boldsymbol{w} \leq \boldsymbol{c}$.
	$\boldsymbol{x} \geq \boldsymbol{0}$		

(forma canónica)

Problema primal P		**Problema dual D**	
minimizar	$\boldsymbol{c}^\intercal \boldsymbol{x}$	maximizar	$\boldsymbol{b}^\intercal \boldsymbol{w}$
sujeto a	$\boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}$	sujeto a	$\boldsymbol{A}^\intercal \boldsymbol{w} \leq \boldsymbol{c}$
	$\boldsymbol{x} \geq \boldsymbol{0}$		$\boldsymbol{w} \geq \boldsymbol{0}$.

A modo de resumen, de las anteriores formulaciones podemos extraer las siguientes relaciones entre un problema y su dual:

- Si el primal es un problema de minimización, entonces el dual lo es de maximización.
- Matricialmente, en el problema primal trabajamos con la matriz $\boldsymbol{A}$ y en el dual con $A^\intercal$. Por este motivo en vez de $\boldsymbol{A}\boldsymbol{x}$ tenemos $\boldsymbol{A}^\intercal \boldsymbol{w}$.
- Cada variable del primal se corresponde con una restricción del dual.
- Cada restricción del primal se corresponde con una variable del dual.
- El vector $\boldsymbol{b}$ de lados derechos del primal es el vector de costes del dual.
- El vector de costes del primal $\boldsymbol{c}$ es el vector de lados derechos del dual.
- Variables no restringidas se corresponden con restricciones de igualdad y variables no negativas se corresponden con restricciones de menor o igual.
- Restricciones de igualdad se corresponden con variables no restringidas y restricciones de mayor o igual se corresponden con variables no negativas.

Recordamos a continuación los principales resultados relativos a la dualidad:

Proposición. *El dual del dual es el primal.*

Teorema (Teorema de dualidad débil)**.** *Dado un par soluciones factibles $\boldsymbol{x}$ y $\boldsymbol{w}$ de los problemas P y D, respectivamente, entonces $\boldsymbol{c}^\intercal \boldsymbol{x} \geq \boldsymbol{b}^\intercal \boldsymbol{w}$.*

Este sencillo resultado, cuya demostración es inmediata, tiene varias implicaciones prácticas. La primera de ellas es que las soluciones factibles del primal (problema de minimización) siempre nos darán cotas superiores para el dual (problema de maximización). Análogamente, las soluciones factibles del dual siempre nos darán cotas inferiores para el primal.

Corolario. *Si $\boldsymbol{x}$ y $\boldsymbol{w}$ son soluciones factibles del primal y del dual tales que $\boldsymbol{c}^\top \boldsymbol{x} = \boldsymbol{b}^\top \boldsymbol{w}$, entonces $\boldsymbol{x}$ y $\boldsymbol{w}$ son soluciones óptimas de primal y dual, respectivamente.*

Corolario. *Dado un par de problemas duales P y D, si la función objetivo de uno de ellos es no acotada, entonces el otro no tiene soluciones factibles.*

Teorema (Teorema de dualidad fuerte)**.** *Dado un par de problemas duales P y D, si uno de ellos tiene una solución óptima, entonces también el otro tiene solución óptima y los valores óptimos de la función objetivo coinciden.*

En particular, si $\bar{\boldsymbol{x}}$ y $\bar{\boldsymbol{w}}$ son soluciones óptimas de P y D, respectivamente, tenemos que

$$\boldsymbol{c}^\top \bar{\boldsymbol{x}} = \boldsymbol{b}^\top \bar{\boldsymbol{w}}.$$

Dado un problema de programación lineal, este puede caer en una de las siguientes categorías: (I) El problema tiene un óptimo finito, (II) El problema es no acotado y (III) No tiene solución. Si combinamos este hecho con los resultados que acabamos de presentar, no es difícil las relaciones en la Tabla 5.1, que ilustra las distintas categorías en las que se puede encontrar el par primal-dual.

Primal \ Dual	Óptimo finito	No acotado	Sin soluciones factibles
Óptimo finito	(1)	imposible	imposible
No acotado	imposible	imposible	(2)
Sin soluciones factibles	imposible	(2)	(3)

Tabla 5.1: Posibles categorías para un par primal-dual.

Para terminar, presentamos un último resultado, el conocido como Teorema de las holguras complementarias.

Teorema (Teorema de las holguras complementarias)**.** *Dado un par de problemas duales P y D y un par de soluciones factibles $\boldsymbol{x}$ y $\boldsymbol{w}$. Entonces, $\boldsymbol{x}$ y $\boldsymbol{w}$ forman un par de soluciones óptimas si y sólo si*

$$\begin{aligned} &\boldsymbol{w}_j(\boldsymbol{A}_j^f \boldsymbol{x} - b_j) = 0 \quad \textit{para todo } j \in \{1, \dots, m\} \textit{ y} \\ &(c_i - \boldsymbol{w}^\top \boldsymbol{A}_i^c) x_i = 0 \quad \textit{para todo } i \in \{1, \dots, n\}. \end{aligned}$$

Este resultado nos dice que, bajo optimalidad, aquellas restricciones del primal en las que haya holgura, $\boldsymbol{A}_j^{\mathrm{f}} \boldsymbol{x} - b_j \neq 0$, tendrán variable dual asociada nula, $w_j = 0$. Análogamente, aquellas restricciones del dual en las que haya holgura, $c_i - \boldsymbol{w}^\top \boldsymbol{A}_i^{\mathrm{c}} \neq 0$, tendrán variable primal asociada nula, $x_i = 0$.

5.4.2 El dual lagrangiano

Durante esta sección hablaremos continuamente de los problemas primal y dual asociados a un problema de optimización no lineal. En particular, trabajaremos con el *dual lagrangiano*, que se apoya en la función lagrangiana, con la cual ya trabajamos en la Sección 5.3.7. Más concretamente, trabajaremos con problemas de la forma

$$\begin{array}{rll} \text{minimizar} & f(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 & i = 1, \ldots, m \\ & h_j(\boldsymbol{x}) = 0 & j = 1, \ldots, l \\ & \boldsymbol{x} \in S, & \end{array} \tag{5.1}$$

donde el conjunto S vendrá dado a su vez por restricciones adicionales, que pueden ser tanto lineales como no lineales. La idea es que este conjunto contendrá aquellas restricciones que no se desean "dualizar" (subir a la función objetivo), típicamente porque son restricciones "fáciles" (como por ejemplo restricciones lineales o de cota). La función lagrangiana asociada a este problema viene dada por

$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} u_i g_i(\boldsymbol{x}) + \sum_{j=1}^{l} v_j h_j(\boldsymbol{x}).$$

Se trata de una función que depende tanto de las variables del primal, $\boldsymbol{x}$, como de las que en esta sección llamaremos *variables duales*, $\boldsymbol{u}$ y $\boldsymbol{v}$. Naturalmente, existe una fuerte relación entre las variables duales y los multiplicadores de Lagrange estudiados en la Sección 5.3, y que analizaremos más adelante.

Asociadas a la función lagrangiana tendremos las funciones primal y dual, dadas por $\mathcal{L}^P(\boldsymbol{x})$ y $\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v})$ respectivamente, y cuyas definiciones son

$$\mathcal{L}^P(\boldsymbol{x}) = \sup_{(\boldsymbol{u},\boldsymbol{v}),\boldsymbol{u}\geq 0} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \qquad \text{y} \qquad \mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v}) = \inf_{\boldsymbol{x}\in S} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}).$$

Nótese que la función $\mathcal{L}^P(\boldsymbol{x})$ toma valores en $\mathbb{R}\cup\{+\infty\}$ y $\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v})$ en $\mathbb{R}\cup\{-\infty\}$. A continuación definimos los problemas primal y dual:

Problema primal P			**Problema dual D**	
minimizar	$\mathcal{L}^P(\boldsymbol{x})$	y	maximizar	$\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v})$
sujeto a	$\boldsymbol{x} \in S$		sujeto a	$\boldsymbol{u} \geq \boldsymbol{0}$.

En la literatura se han propuesto muchas formulaciones alternativas de los llamados *problemas duales*, y la que acabamos de presentar, que se conoce con el nombre de *dual lagrangiano* es quizá la variante más estudiada y en la que se apoyan más algoritmos de optimización. La teoría de la dualidad investiga las relaciones entre el problema primal y el dual. La función primal $\mathcal{L}^P(\boldsymbol{x})$ es fácil de evaluar. Si un punto $\boldsymbol{x}$ es factible en el problema de optimización original, entonces $\mathcal{L}^P(\boldsymbol{x}) = f(\boldsymbol{x})$ pues las restricciones asociadas a las variables no negativas u_i toman valores no positivos y las restricciones asociadas a las variables v_j toman valor cero; es decir, para minimizar $f(\boldsymbol{x}) + \sum_{i=1}^{m} u_i g_i(\boldsymbol{x}) + \sum_{j=1}^{l} v_j h_j(\boldsymbol{x})$ dado un $\boldsymbol{x}$ factible lo mejor que se puede hacer es tomar $\boldsymbol{u} = \boldsymbol{0}$, obteniendo justamente $f(\boldsymbol{x})$. Por otro lado, si $\boldsymbol{x}$ no es factible y

se viola una cierta restricción, podremos hacer tender $\mathcal{L}^P(\boldsymbol{x})$ a $+\infty$ sin más que incrementar arbitrariamente la variable dual asociada. Por tanto, la función primal viene dada por:

$$\mathcal{L}^P(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & \text{si } \boldsymbol{x} \text{ es factible para el problema (5.1), y} \\ +\infty & \text{en otro caso,} \end{cases}$$

de donde se concluye inmediatamente que el problema de optimización original es equivalente al problema primal P. La función dual es más compleja evaluar y en general no es fácil obtener una expresión explícita para la misma. Como mostraremos más adelante en esta sección, una excepción son los problemas de programación lineal. Tras esta discusión podemos presentar las formulaciones finales de los problemas P y D, con las que trabajaremos en el resto de la sección:

Problema primal P

$$\begin{array}{lll} \text{minimizar} & f(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 & i = 1, \ldots, m \\ & h_j(\boldsymbol{x}) = 0 & j = 1, \ldots, l \\ & \boldsymbol{x} \in S & \end{array} \tag{5.2}$$

y

Problema dual D

$$\begin{array}{ll} \text{maximizar} & \mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v}) = \inf\limits_{\boldsymbol{x} \in S} f(\boldsymbol{x}) + \sum\limits_{i=1}^{m} u_i g_i(\boldsymbol{x}) + \sum\limits_{j=1}^{l} v_j h_j(\boldsymbol{x}) \\ \text{sujeto a} & \boldsymbol{u} \geq \boldsymbol{0}. \end{array} \tag{5.3}$$

El dual lagragiano consiste en pasar penalizadas a la función objetivo todas las restricciones del problema P, salvo aquellas recogidas en el conjunto S. Por tanto, a un mismo problema de programación matemática pueden corresponderle distintos duales lagrangianos, según las restricciones que se incluyan en el conjunto S. Esta elección será relevante, pues afectará a la dificultad de resolución del problema dual resultante y a su utilidad para resolver el problema P. Las penalizaciones asociadas a las restricciones dualizadas vienen dadas por las *variables duales*, u_i y v_j.

Fijados $\boldsymbol{u}$ y $\boldsymbol{v}$, el problema de minimización que hay que resolver para evaluar la función dual $\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v})$ se llama *subproblema dual lagrangiano*. Siendo totalmente formales, en la Ecuación (5.3) deberíamos hablar del supremo de $\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v})$, pues el problema de optimización dual podría ser no acotado (aunque el Teorema de dualidad débil implicará que esto no será posible si el problema P tiene alguna solución factible).

Dualidad lagrangiana y teoría de juegos

Existe una fuerte relación entre la teoría de la dualidad y la teoría de juegos. Desde el punto de vista histórico, la dualidad en programación lineal surgió cuando el brillante matemático John von Neumann estaba estudiando condiciones de equilibrio en juegos bipersonales matriciales, y también en discusiones con George Dantzig, padre del método símplex, cuando este último estaba desarrollando dicho algoritmo. Veamos un poco la intuición detrás de esta relación, en la que profundizaremos algo más en la Sección 5.4.5.

Tenemos un juego en el que un jugador quiere minimizar la función lagrangiana y otro quiere maximizarla. Para ello, el jugador 1 tiene que elegir las variables del primal, el jugador 2 las variables del dual y $L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$ representa el pago que tiene que hacerle el jugador 1 al jugador 2 para una combinación de estrategias dada.

La función objetivo del jugador 1 será, para cada $\boldsymbol{x} \in S$, $u_1(\boldsymbol{x}, (\boldsymbol{u}, \boldsymbol{v})) = -L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$, y la del jugador 2 será, dados $\boldsymbol{u} \geq \boldsymbol{0}$ y $\boldsymbol{v}$, $u_2(\boldsymbol{x}, (\boldsymbol{u}, \boldsymbol{v})) = L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = -u_1(\boldsymbol{x}, (\boldsymbol{u}, \boldsymbol{v}))$. Estamos entonces ante un juego de suma nula. De este modo, dada una elección de "estrategia" $(\boldsymbol{u}, \boldsymbol{v})$ con $\boldsymbol{u} \geq 0$ por parte del jugador 2, el jugador 1 tiene que intentar maximizar $u_1(\boldsymbol{x}, (\boldsymbol{u}, \boldsymbol{v}))$ o, equivalentemente, minimizar $f(\boldsymbol{x}) + \sum_{i=1}^{m} u_i g_i(\boldsymbol{x}) + \sum_{j=1}^{l} v_j h_j(\boldsymbol{x})$. Independientemente de la elección del jugador 2, el jugador 1 siempre puede asegurarse un valor al menos tan bajo como el del óptimo del problema primal P. Esto es porque en toda solución factible del primal $\boldsymbol{x}$, por ser $\boldsymbol{u} \geq \boldsymbol{0}$, tenemos que

$$f(\boldsymbol{x}) + \sum_{i=1}^{m} u_i g_i(\boldsymbol{x}) + \sum_{j=1}^{l} v_j h_j(\boldsymbol{x}) \leq f(\boldsymbol{x}).$$

Esta observación es la base del Teorema de dualidad débil (Teorema 5.13) que veremos en la Sección 5.4.4. El objetivo del jugador 1 es intentar aprovecharse de que puede elegir "estrategias" $\boldsymbol{x}$ que no se correspondan con soluciones factibles del primal para así conseguir un valor todavía inferior de la función lagrangiana. Lo que nos dice el Teorema de dualidad fuerte (Teorema 5.19) que también veremos en la Sección 5.4.4 es que, cuando P es un problema de optimización convexa, si el jugador 2 elige adecuadamente su "estrategia" $(\boldsymbol{u}, \boldsymbol{v})$, el jugador 1 no puede hacer nada mejor que responder con la estrategia dada por la solución óptima de P. En general esto no tiene por qué ser así, pues el jugador 1 podría ser capaz de mejorar el valor del óptimo del primal para cualquier elección de estrategias del jugador 2.

En particular, cuando estamos ante un problema de optimización convexa, tenemos que las elecciones óptimas de ambos jugadores, $\bar{\boldsymbol{x}}$ y $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$, son un equilibrio de Nash del juego que acabamos de describir. Aún más, en la Sección 5.4.5 probaremos que el recíproco también es cierto: hay una correspondencia entre pares de soluciones óptimas de los problemas P y D y equilibrios de Nash de este juego.

Dualidad lagrangiana y métodos de penalización

A continuación discutimos brevemente la relación entre la dualidad lagrangiana y uno de los algoritmos clásicos de optimización con restricciones que se discutirán en la Sección 7.2: los métodos de penalización exterior. Estos métodos, a la hora de resolver un problema de optimización como el de la Ecuación (5.2), proceden de la siguiente manera: toman una sucesión $\{\rho^t\}_{t \in \mathbb{N}} \subset \mathbb{R}^+$ tal que $\rho^t \to \infty$ y, para cada ρ^t, resuelven el problema:

$$\underset{\boldsymbol{x} \in S}{\text{minimizar}}\ f(\boldsymbol{x}) + \sum_{i=1}^{m} \rho^t \max\{0, g_i(\boldsymbol{x})\} + \sum_{j=1}^{l} \rho^t |h_j(\boldsymbol{x})|.$$

La idea es que, cuando ρ^t sea suficientemente grande, los óptimos del problema de la iteración t serán necesariamente puntos factibles del problema P (pues las infactibilidades están muy

penalizadas). En particular, dentro de estos puntos factibles, los óptimos para el problema penalizado coincidirán con los óptimos del problema original. En el caso particular de que $S = \emptyset$, esto permite atacar problemas de optimización con restricciones utilizando toda la maquinaria disponible para la resolución de problemas de optimización sin restricciones. Aunque se trata de un enfoque muy natural, en la Sección 7.2 veremos que tiene importantes limitaciones.

A la hora de trabajar con el dual lagrangiano tenemos que las variables duales reemplazan a la penalización ρ^t de las distintas restricciones, pero el papel es similar. El método del lagrangiano aumentado que veremos en la Sección 7.3 enriquece la idea de los métodos de penalización, pero trabajando explícitamente con las variables duales en vez de usar la sucesión $\{\rho^t\}_{t\in\mathbb{N}}$. Informalmente la idea se puede describir como sigue: si el método de penalización busca trabajar con penalizaciones suficientemente grandes como para que ninguna solución no factible pueda ser óptimo del subproblema a resolver en la iteración t, el método de lagrangiano aumentado busca que las penalizaciones $(\boldsymbol{u}^t, \boldsymbol{v}^t)$ de la iteración t se aproximen cada vez más al valor óptimo del problema dual. Esto permite equilibrar la tensión entre optimalidad y factibilidad sin necesitar inclinar la balanza totalmente hacia el lado de la factibilidad como hacen los métodos de penalización al hacer que $\rho^t \to \infty$.

Dualidad en programación lineal

A continuación vamos a recordar la formulación de la dualidad en programación lineal y mostrar que es un caso particular de la dualidad lagrangiana. Esto implicará, en particular, que cualquier resultado matemático que se obtenga para dualidad lagrangiana se puede aplicar inmediatamente para la dualidad clásica en programación lineal. La formulación clásica de los problemas primal y dual en programación lineal viene dada por

$$\begin{array}{ll} \textbf{Problema primal P} & \\ \text{minimizar} & \boldsymbol{c}^\intercal \boldsymbol{x} \\ \quad \text{sujeto a} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq 0 \end{array} \quad y \quad \begin{array}{ll} \textbf{Problema dual D} & \\ \text{maximizar} & \boldsymbol{v}^\intercal \boldsymbol{b} \\ \quad \text{sujeto a} & \boldsymbol{A}^\intercal \boldsymbol{v} \leq \boldsymbol{c}. \end{array}$$

Por otro lado, si tomamos $S = \{\boldsymbol{x} : \boldsymbol{x} \geq \boldsymbol{0}\}$ tenemos el siguiente dual lagrangiano:

$$\text{maximizar} \quad \mathcal{L}^D(\boldsymbol{v}) = \inf_{\boldsymbol{x}\geq\boldsymbol{0}} \{\boldsymbol{c}^\intercal \boldsymbol{x} + \boldsymbol{v}^\intercal(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})\}.$$

Claramente, $\mathcal{L}^D(\boldsymbol{v}) = \boldsymbol{v}^\intercal \boldsymbol{b} + \inf_{\boldsymbol{x}\geq\boldsymbol{0}}\{(\boldsymbol{c}^\intercal - \boldsymbol{v}^\intercal \boldsymbol{A})\boldsymbol{x}\}$. Y ahora obtenemos una formulación explícita para la función dual:

$$\mathcal{L}^D(\boldsymbol{v}) = \begin{cases} \boldsymbol{v}^\intercal \boldsymbol{b} & \text{si } (\boldsymbol{c}^\intercal - \boldsymbol{v}^\intercal \boldsymbol{A}) \geq \boldsymbol{0} \\ -\infty & \text{en otro caso.} \end{cases}$$

Con lo que el dual lagrangiano puede ser reescrito equivalentemente como

$$\begin{array}{ll} \text{maximizar} & \boldsymbol{v}^\intercal \boldsymbol{b} \\ \quad \text{sujeto a} & \boldsymbol{A}^\intercal \boldsymbol{v} \leq \boldsymbol{c}, \end{array}$$

que es justamente lo que queríamos mostrar.

5.4.3 Interpretación geométrica de la dualidad lagrangiana

En este apartado discutimos la intuición geométrica detrás del problema dual. Para ello trabajaremos con un problema con una única restricción de la forma $g(\boldsymbol{x}) \leq 0$, con lo que primal y dual serán de la siguiente forma:

$$\begin{array}{ll} \textbf{Problema primal P} & \\ \text{minimizar} & f(\boldsymbol{x}) \\ \text{sujeto a} & g(\boldsymbol{x}) \leq 0 \\ & \boldsymbol{x} \in S \end{array} \qquad y \qquad \begin{array}{ll} \textbf{Problema dual D} & \\ \text{maximizar} & \mathcal{L}^D(u) = \text{ínf } f(\boldsymbol{x}) + ug(\boldsymbol{x}) \\ \text{sujeto a} & u \geq 0. \end{array} \tag{5.4}$$

Ahora trabajaremos en $\mathbb{R}^2$, con vectores de la forma (y, z) y, en particular, con el conjunto $G = \{(y, z) \in \mathbb{R}^2 : y = g(\boldsymbol{x}) \text{ y } z = f(\boldsymbol{x}) \text{ para algún } \boldsymbol{x} \in S\}$. Este conjunto, representado en la Figura 5.8, es la imagen bajo la aplicación (g, f) del conjunto S. Los puntos factibles de P se corresponden con puntos (y, z) de G con un valor $y \leq 0$ y entonces z representará un valor factible de f en P. El óptimo de P se corresponderá con el punto de G con $y \leq 0$ y un menor valor de z, representado como $(\bar{y}, \bar{z})$ en la figura.

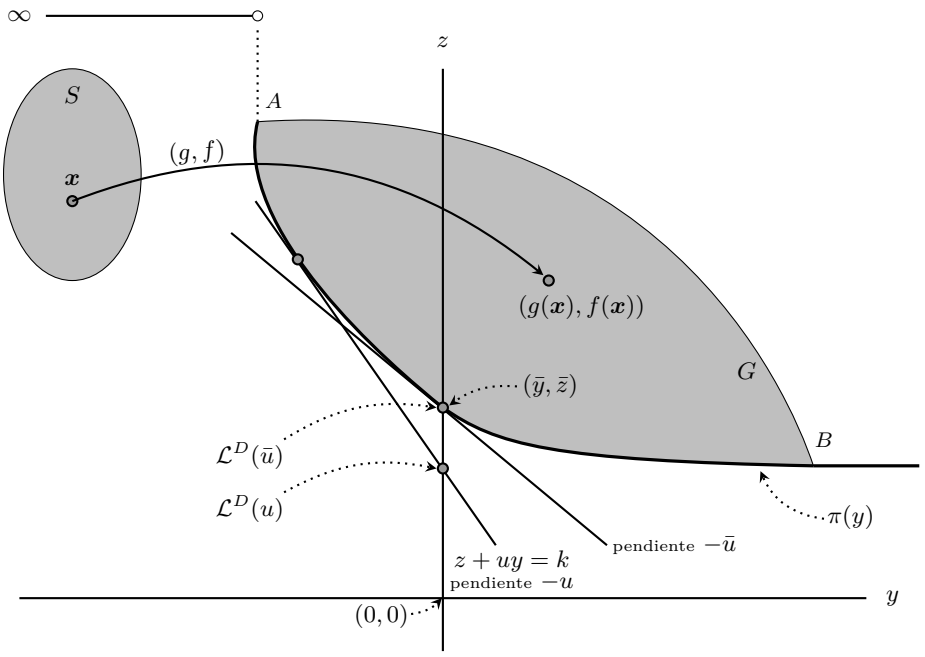


Figura 5.8: Problema primal. Encontrar el punto en G con $y \leq 0$ y menor ordenada z: $(\bar{y}, \bar{z})$ (adaptación de la Figura 6.1 en Bazaraa y otros (2006)).

Veamos ahora la cómo interpretar el dual, apoyándonos también en la Figura 5.8. Para ello empezamos con la evaluación de la función dual $\mathcal{L}^D(u)$. Dado $u \geq 0$, $\mathcal{L}^D(u)$ resulta de minimizar $f(\boldsymbol{x}) + ug(\boldsymbol{x})$ con $\boldsymbol{x} \in S$. Equivalentemente, queremos minimizar $z + uy$ con (y, z) en el conjunto G. Ahora bien, $z + uy = k$ representa una recta con pendiente $-u$ y ordenada en el origen k. Por tanto, $\mathcal{L}^D(u)$ se corresponde con el valor de k más pequeño para el cual la

recta $z + uy = k$ interseca a G. Equivalentemente, $\mathcal{L}^D(u)$ busca el hiperplano soporte de G con pendiente $-u$.

Ahora resulta sencillo interpretar el problema dual, que consiste maximizar la función $\mathcal{L}^D(u)$. El problema D consiste en encontrar la pendiente del hiperplano soporte de G que tiene una mayor ordenada en el origen, denotado por $-\bar{u}$ en la Figura 5.8. Además, en la Figura 5.8 vemos también que dicho hiperplano se corresponde con la recta $z + \bar{u}y = \bar{z}$, con lo que en el problema ahí representado tenemos que el objetivo en el óptimo del primal coincide con el objetivo en el óptimo del dual. Esto se debe a la convexidad del conjunto G y, como veremos más adelante, esta se obtendrá cuando el problema P sea un problema de optimización convexa.

Función perturbación

Para terminar esta discusión vamos a presentar una interpretación adicional, relacionada con la anterior, y que es relevante para el análisis de sensibilidad y la interpretación económica de los multiplicadores de Lagrange presentada en la Sección 5.3.5.

Definamos $\pi(y) = \text{mín}_{x \in S}\{f(\boldsymbol{x}) : g(\boldsymbol{x}) \leq y\}$, llamada *función perturbación* pues devuelve el valor óptimo del problema que se obtiene al perturbar la restricción $g(\boldsymbol{x}) \leq 0$ del problema P. En particular, el cálculo de $\pi(0)$ es equivalente a resolver el problema P. Claramente π es una función no creciente en y, pues a medida que y crece también crece la región factible de nuestro problema de minimización. En la Figura 5.8 vemos que para valores de y a la izquierda de A tenemos $\pi(y) = \infty$ y a la derecha de B la π permanece constante. Entre los puntos A y B, π se corresponde con la envolvente inferior de G, propiedad que se cumplirá siempre que tengamos que P es un problema de optimización convexa (sin convexidad π será la mayor función monótona decreciente que envuelve G inferiormente).

Fijémonos ahora en que si π es diferenciable tenemos que $\pi'(0) = -\bar{u}$, lo que da lugar a la siguiente interpretación económica del la solución óptima del dual. El opuesto de la solución óptima del dual, $-\bar{u}$, es la tasa de variación de la función objetivo a medida que perturbo la restricción $g(\boldsymbol{x}) \leq 0$. Equivalentemente, $-\bar{u}$ me dice cuanto gano, localmente, por cada unidad en la que relaje la restricción. Dado que $-\bar{u}$ es la pendiente de un hiperplano soporte de G y bajo convexidad de este conjunto π es una envolvente inferior del mismo, si π no es diferenciable pero sí convexa tenemos que $-\bar{u}$ es un subgradiente de π en 0, pues $\pi(y) \geq \pi(0) - \bar{u}y$ para todo $y \in \mathbb{R}$.[5] En la Sección 5.4.5 probaremos que las propiedades que acabamos de comentar de la función perturbación son ciertas en general.

Como ya hemos comentado, más adelante probaremos la correspondencia entre multiplicadores de Lagrange y soluciones óptimas del dual, con lo que la anterior interpretación económica aplica a los multiplicadores de Lagrange, lo que supone una justificación adicional de la discusión presentada en la Sección 5.3.5.

•**Ejercicio 5.8.** Demuestra que la función perturbación $\pi(y)$ asociada al problema P de la Ecuación (5.4) es la mayor función decreciente cuyo epigrafo contiene al conjunto G. ◁

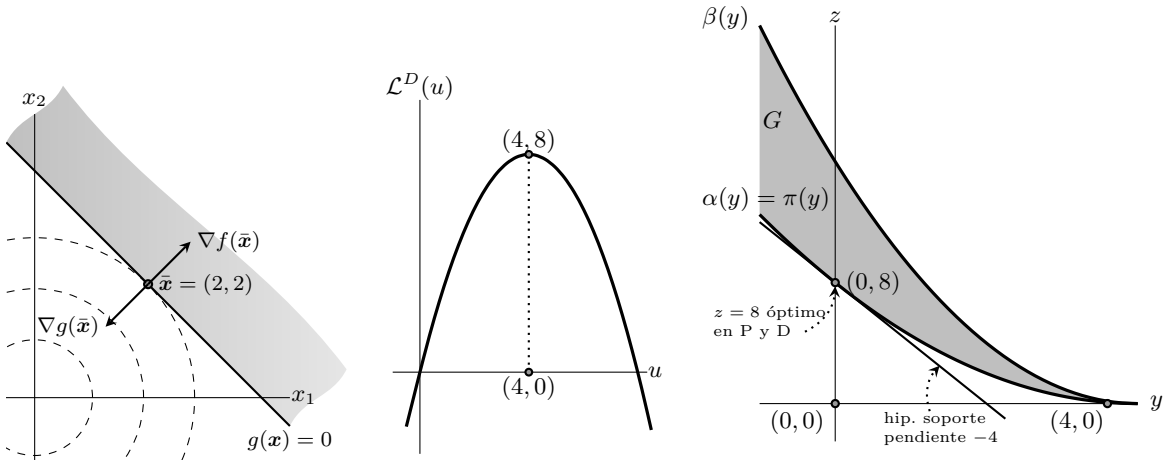
[5] Recordemos que, según la Definición 1.5, la condición para que $\boldsymbol{s}$ sea un subgradiente de una función convexa $f : S \to \mathbb{R}^n$ en un punto $\bar{\boldsymbol{x}}$ es que $f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}) + s^{\intercal}(\boldsymbol{x} - \bar{\boldsymbol{x}})$ para todo $\boldsymbol{x} \in S$.

Ejemplo ilustrativo

Ejemplo 5.6. Terminamos esta sección con un ejemplo ilustrativo de los conceptos geométricos que hemos descrito a través de la Figura 5.8. Consideremos el problema primal dado por

$$\begin{aligned} \text{minimizar} \quad & x_1^2 + x_2^2 \\ \text{sujeto a} \quad & -x_1 - x_2 + 4 \leq 0 \\ & \boldsymbol{x} \geq \mathbf{0} \end{aligned}$$

y cuya representación puede verse en la Figura 5.9(a). Claramente, el óptimo del primal es el punto $\bar{\boldsymbol{x}} = (2, 2)$ y el valor objetivo es $f(\bar{\boldsymbol{x}}) = 8$.



(a) Problema P. (b) Problema D. (c) Relación geométrica entre P y D.

Figura 5.9: Ilustración del Ejemplo 5.6 (adaptación de la Figura 6.2 en Bazaraa y otros (2006)).

Por otro lado, tomando $S = \{\boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{x} \geq \mathbf{0}\}$ tenemos que la función dual viene dada por

$$\mathcal{L}^D(u) = \inf_{x \in S} \{x_1^2 + x_2^2 + u(-x_1 - x_2 + 4)\} = \inf_{x_1 \geq 0} \{x_1^2 - ux_1\} + \inf_{x_2 \geq 0} \{x_2^2 - ux_2\} + 4u.$$

Los ínfimos de $\mathcal{L}^D(u)$ se alcanzan en $x_1 = x_2 = u/2$ siempre que $u \geq 0$. Si $u < 0$ el mínimo se alcanza en $x_1 = x_2 = 0$ con lo que obtenemos una expresión explícita para la función dual:

$$\mathcal{L}^D(u) = \begin{cases} -\frac{1}{2}u^2 + 4u & \text{si } u \geq 0 \\ 4u & \text{si } u < 0. \end{cases}$$

Esta función la tenemos representada en la la Figura 5.9(b). Se trata de una función cóncava y su máximo se alcanza en $\bar{u} = 4$ con función objetivo $\mathcal{L}^D(4) = 8$, con lo que el valor óptimo del problema dual coincide con el del primal. Más adelante veremos que la concavidad es una propiedad general de la función dual, lo cual es muy importante a la hora de resolver el problema dual, dado que este es un problema de maximizar. En otras palabras, expresado como un problema de minimización, el problema dual es un problema de optimización convexo, siendo esta propiedad independiente de si el primal lo es.

Si pasamos ahora a la construcción del conjunto G, nos encontramos que podemos caracterizarlo mediante su envoltura superior e inferior. Para cada valor y estamos interesados en los valores mínimo y máximo que puede tomar $f(\boldsymbol{x})$ sujeto a la restricción $-x_1 - x_2 + 4 = y$. A la función que representa los valores mínimos le llamaremos $\alpha(y)$ y a la que representa los máximos $\beta(y)$. Sustituyendo por ejemplo $x_2 = -x_1 + 4 - y$ en la función objetivo y derivando es fácil obtener que, para todo $y \leq 4$, $\alpha(y) = (4-y)^2/2$ y $\beta(y) = (4-y)^2$. Además, es fácil ver si $y > 4$ el problema no tiene soluciones factibles, pues ningún $\boldsymbol{x} \geq \mathbf{0}$ puede cumplir la restricción. El conjunto G y las funciones $\alpha(y)$ y $\beta(y)$ están representados en la Figura 5.9(c), donde se puede ver además que la función $\alpha(y)$ coincide con la función perturbación $\pi(y)$ cuando $y \leq 4$ y permanece constante en el valor 0 de ahí en adelante. La pendiente de la tangente a la función en $y = 0$ es $\pi'(0) = -4$, el opuesto de la solución óptima del problema dual. Por último, tenemos que dicha tangente está estrictamente por debajo del conjunto G o, lo que es lo mismo, $\pi(y) \geq \pi(0) - 4y$ para todo $y \in \mathbb{R}$, lo cual, aunque no lo probaremos, es una condición necesaria y suficiente para que el primal y el dual tengan el mismo valor óptimo. ◇

5.4.4 Teoremas de dualidad

A continuación presentamos algunos de los principales resultados teóricos relativos a la dualidad, conocidos como dualidad débil y dualidad fuerte. La dualidad débil se cumple siempre y la dualidad fuerte necesita supuestos de convexidad.

Durante toda esta sección trabajaremos con las formulaciones de los problemas primal, P, y dual, D, de las ecuaciones (5.2) y (5.3), respectivamente. Cuando sea conveniente trabajaremos con la formulación vectorial de las restricciones: $\boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}$ y $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$. Esto nos permite, en particular, escribir la función dual como $\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v}) = \inf_{\boldsymbol{x} \in S} f(\boldsymbol{x}) + \boldsymbol{u}^\top \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{v}^\top \boldsymbol{h}(\boldsymbol{x})$.

Dualidad débil

Teorema 5.13 (Teorema de dualidad débil)**.** *Dado un par soluciones factibles $\boldsymbol{x}$ y $(\boldsymbol{u}, \boldsymbol{v})$ de los problemas P y D, respectivamente, entonces $f(\boldsymbol{x}) \geq \mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v})$.*

Demostración. El resultado es consecuencia inmediata de la siguiente ecuación:

$$\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v}) = \inf_{\hat{\boldsymbol{x}} \in S} f(\hat{\boldsymbol{x}}) + \boldsymbol{u}^\top \boldsymbol{g}(\hat{\boldsymbol{x}}) + \boldsymbol{v}^\top \boldsymbol{h}(\hat{\boldsymbol{x}}) \leq f(\boldsymbol{x}) + \boldsymbol{u}^\top \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{v}^\top \boldsymbol{h}(\boldsymbol{x}) \leq f(\boldsymbol{x}),$$

donde la primera desigualdad no es más que aplicar la definición de ínfimo y la segunda usa que, dada la factibilidad de $\boldsymbol{x}$ y $(\boldsymbol{u}, \boldsymbol{v})$, sabemos que $\boldsymbol{u} \geq \mathbf{0}$, $\boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}$ y $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$. □

El Teorema de dualidad débil nos dice que la función objetivo de cualquier solución factible del primal siempre estará por encima de la función objetivo en cualquier solución factible del dual. En particular, como P es un problema de minimizar, cualquier solución factible del dual nos dará una cota inferior del valor óptimo del primal. Del mismo modo, cualquier solución factible del primal nos dará una cota superior del valor óptimo del dual. Esta y otras consideraciones están recogidas en los siguientes corolarios del Teorema de dualidad débil, cuyas demostraciones omitimos por ser inmediatas.

Corolario 5.14. $\inf_{\boldsymbol{x} \in S}\{f(\boldsymbol{x}) : \boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}, \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}\} \geq \sup\{\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v}), \boldsymbol{u} \geq \mathbf{0}\}$.

Corolario 5.15. *Si $f(\bar{\boldsymbol{x}}) = \mathcal{L}^D(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$, con $\bar{\boldsymbol{x}}$ factible en P y $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ factible en D, entonces $\bar{\boldsymbol{x}}$ es óptimo en P y $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ óptimo en D.*

Corolario 5.16. *Si $\inf_{\boldsymbol{x}\in S}\{f(\boldsymbol{x}) : \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}, \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}\} = -\infty$, entonces $\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v}) = -\infty$ para todo $\boldsymbol{u} \geq \boldsymbol{0}$ y para todo $\boldsymbol{v}$.*

Corolario 5.17. *Si $\sup\{\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v}), \boldsymbol{u} \geq \boldsymbol{0}\} = \infty$, entonces P no tiene soluciones factibles.*

Duality gap

El Corolario 5.15 nos da una condición suficiente para comprobar la optimalidad global de una solución factible del primal: basta con encontrar una solución factible del dual que tenga la misma función objetivo. Desafortunadamente, esto no resulta fácil, ya que es habitual que dicha solución no exista, en cuyo caso nos encontramos en situaciones en las que $\inf_{\boldsymbol{x}\in S}\{f(\boldsymbol{x}) : \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}, \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}\} > \sup\{\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v}), \boldsymbol{u} \geq \boldsymbol{0}\}$, y la diferencia entre estas dos cantidades se conoce como *duality gap*.

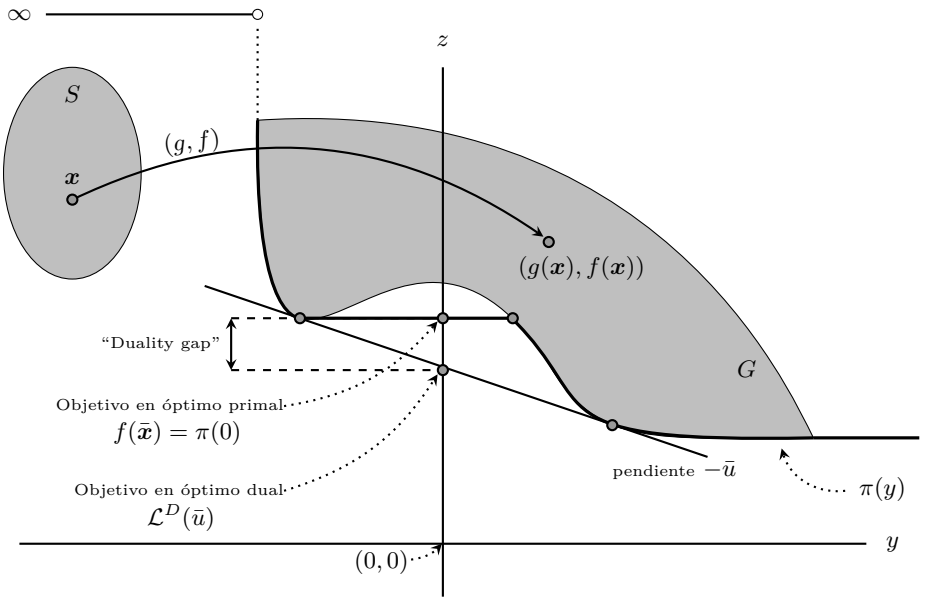


Figura 5.10: Ilustración de la existencia de *duality gap* (adaptación de la Figura 6.3 en Bazaraa y otros (2006)).

La Figura 5.10 representa una situación en la que el *duality gap* es estrictamente positivo. En la figura podemos ver también la función perturbación $\pi(y)$, que recordemos es la mayor función decreciente que envuelve G inferiormente (Ejercicio 5.8). Recordemos que el óptimo del problema primal es $\pi(0)$. Por otro lado, para el hiperplano soporte del conjunto G con una mayor ordenada en el origen dicha ordenada toma el valor $\mathcal{L}^D(\bar{\boldsymbol{u}}) < \pi(0)$. Además, también es fácil ver que no existe ningún $\boldsymbol{u}$ tal que $\pi(y) \geq \pi(0) - uy$ para todo $y \in \mathbb{R}$. En otras palabras, la falta de convexidad de G hace que el epigrafo de $\pi(y)$ no tenga ningún hiperplano soporte en $y = 0$.

Dualidad fuerte

En este apartado presentamos el Teorema de dualidad fuerte, que da condiciones suficientes para que el valor óptimo de P y D coincidan o, lo que es lo mismo, para que no exista *duality gap*. Previamente necesitamos un lema auxiliar, que guarda cierta similitud con el Lema de Farkas (Teorema 1.9) y que al igual que este requiere del uso de los teoremas de separación para su demostración; más concretamente, el Corolario 1.5 al Teorema del hiperplano soporte (Teorema 1.4). Recordemos que dicho corolario dice que, dado un conjunto no vacío y convexo $S \subseteq \mathbb{R}^n$ y $\bar{\boldsymbol{x}} \notin \mathring{S}$, entonces existe un hiperplano separando $\bar{\boldsymbol{x}}$ y $\bar{S}$. Equivalentemente, existe un vector $\boldsymbol{w} \neq \mathbf{0}$ tal que $\boldsymbol{w}^\top(\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq 0$ para todo $\boldsymbol{x}$ en la clausura de S; es decir, el vector $\boldsymbol{w}$ forma un ángulo de al menos 90° con todos los vectores de la forma $\boldsymbol{x} - \bar{\boldsymbol{x}}$.

Lema 5.18. *Sea $S \subseteq \mathbb{R}^n$ un conjunto convexo, $\alpha : \mathbb{R}^n \to \mathbb{R}$ y $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ convexas, y $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^l$ afín, i.e., $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$. Dados los sistemas:*

Sistema 1. *$\alpha(\boldsymbol{x}) < 0$, $\boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}$, $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$ con $\boldsymbol{x} \in S$.*

Sistema 2. *$u_0\alpha(\boldsymbol{x}) + \boldsymbol{u}^\top\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{v}^\top\boldsymbol{h}(\boldsymbol{x}) \geq 0$ para todo $\boldsymbol{x} \in S$ con $(u_0, \boldsymbol{u}) \geq \mathbf{0}$ y $(u_0, \boldsymbol{u}, \boldsymbol{v}) \neq \mathbf{0}$.*

Entonces, si el Sistema 1 no tiene una solución $\boldsymbol{x}$, el Sistema 2 tiene una solución $(u_0, \boldsymbol{u}, \boldsymbol{v})$. Además, el recíproco es cierto si $u_0 > 0$.

Demostración. "⇒" Supongamos que el Sistema 1 no tiene solución y definamos el conjunto

$$\Lambda = \{(p, \boldsymbol{q}, \boldsymbol{r}) : p > \alpha(\boldsymbol{x}),\ \boldsymbol{q} \geq \boldsymbol{g}(\boldsymbol{x}), \boldsymbol{r} = \boldsymbol{h}(\boldsymbol{x}) \text{ para algún } \boldsymbol{x} \in S\}.$$

La convexidad de S y de los epigrafos de α y $\boldsymbol{g}$, unida al carácter afín de $\boldsymbol{h}$, nos asegura que Λ es convexo. Como el Sistema 1 no tiene solución, $(0, \mathbf{0}, \mathbf{0}) \notin \Lambda$. Entonces, por el Corolario 1.5 existe $(u_0, \boldsymbol{u}, \boldsymbol{v}) \neq \mathbf{0}$ tal que

$$u_0 p + \boldsymbol{u}^\top\boldsymbol{q} + \boldsymbol{v}^\top\boldsymbol{r} \geq 0 \quad \text{para todo } (p, \boldsymbol{q}, \boldsymbol{r}) \text{ en la clausura de } \Lambda.^{6}$$

Por la definición de Λ, en la desigualdad anterior p y $\boldsymbol{q}$ se pueden hacer arbitrariamente grandes. Por tanto, necesariamente $(u_0, \boldsymbol{u}) \geq \mathbf{0}$. Fijemos ahora $\boldsymbol{x} \in S$. Entonces, $(p, \boldsymbol{q}, \boldsymbol{r}) = (\alpha(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{x}), \boldsymbol{h}(\boldsymbol{x}))$ está en la clausura de Λ con lo que la desigualdad $u_0\alpha(\boldsymbol{x}) + \boldsymbol{u}^\top\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{v}^\top\boldsymbol{h}(\boldsymbol{x}) \geq 0$ se cumple para todo $\boldsymbol{x} \in S$. Por tanto, el Sistema 2 tiene solución.

"⇐" Supongamos que el Sistema 2 tiene una solución con $u_0 > 0$. Es decir, existe $(u_0, \boldsymbol{u}, \boldsymbol{v})$ con $u_0 > 0$ y $\boldsymbol{u} \geq 0$ tal que

$$u_0\alpha(\boldsymbol{x}) + \boldsymbol{u}^\top\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{v}^\top\boldsymbol{h}(\boldsymbol{x}) \geq 0 \text{ para todo } \boldsymbol{x} \in S.$$

Supongamos que existe $\boldsymbol{x} \in S$ tal que $\boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}$ y $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$. Entonces, necesariamente, $u_0\alpha(\boldsymbol{x}) \geq 0$. Como $u_0 > 0$, tendremos $\alpha(\boldsymbol{x}) \geq 0$. Por tanto, el Sistema 1 no tiene solución. □

De la misma manera que el Lema de Farkas captura la esencia geométrica de la dualidad en programación lineal, el Lema 5.18 captura la esencia geométrica del Teorema de dualidad fuerte. En ambos casos, la condición recogida en el Sistema 1 tiene que ver con el problema primal pues incluye condiciones sobre su región factible $\boldsymbol{A}\boldsymbol{x} \leq \mathbf{0}$, $\boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}$, $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$ y su función objetivo

[6] Realmente el Corolario 1.5 asegura la existencia de $-(u_0, \boldsymbol{u}, \boldsymbol{v}) \neq \mathbf{0}$, que es tal que $-u_0 p - \boldsymbol{u}^\top\boldsymbol{q} - \boldsymbol{v}^\top\boldsymbol{r} \leq 0$.

a través de $\boldsymbol{c}^\intercal \boldsymbol{x} > 0$ y $\alpha(\boldsymbol{x}) < 0$; la relación entre la función $\alpha(\boldsymbol{x})$ y la función $f(\boldsymbol{x})$ quedará clara en la demostración del Teorema de dualidad fuerte. Del mismo modo, en el Sistema 2 aparece reflejado el dual mediante $\boldsymbol{A}^\intercal \boldsymbol{y} = \boldsymbol{c}$ con $\boldsymbol{y} \geq \boldsymbol{0}$ y $u_0\alpha(\boldsymbol{x}) + \boldsymbol{u}^\intercal \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{v}^\intercal \boldsymbol{h}(\boldsymbol{x}) \geq 0$ con $(u_0, \boldsymbol{u}) \geq \boldsymbol{0}$.

A continuación presentamos el Teorema de dualidad, cuyo principal mensaje es que en problemas de optimización convexa no hay *duality gap*. Aunque, al igual que sucedía con las condiciones necesarias de Karush-Kuhn-Tucker (Teorema 5.6), es necesario incluir una condición de regularidad que excluya ciertas situaciones "patológicas".

Teorema 5.19 (Teorema de dualidad fuerte)**.** *Sea P un un problema de optimización convexa en el que $\boldsymbol{0}$ pertenece al interior de $h(S)$ y se cumple la siguiente condición de regularidad: existe $\hat{\boldsymbol{x}}$ tal que $\boldsymbol{g}(\hat{\boldsymbol{x}}) < \boldsymbol{0}$ y $\boldsymbol{h}(\hat{\boldsymbol{x}}) = \boldsymbol{0}$. Entonces no existe* duality gap*, es decir,*

$$\inf_{\boldsymbol{x}\in S}\{f(\boldsymbol{x}) : \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0},\, \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}\} = \sup_{\boldsymbol{u}\geq\boldsymbol{0}} \mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v}).$$

Además, si el anterior ínfimo es finito tenemos que

(i) $\sup_{\boldsymbol{u}\geq\boldsymbol{0}} \mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v})$ *se alcanza para un par $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ con $\bar{\boldsymbol{u}} \geq \boldsymbol{0}$ y*

(ii) *si el ínfimo se alcanza en $\bar{\boldsymbol{x}}$, entonces $\bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) = 0$ (holguras complementarias).*

Demostración. Sea $\gamma = \inf_{x\in S}\{f(\boldsymbol{x}) : \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}, \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}\}$. Como $\hat{\boldsymbol{x}}$ es factible, $\gamma < \infty$. Si $\gamma = -\infty$, por el Corolario 5.16, $\sup_{\boldsymbol{u}\geq\boldsymbol{0}}\{\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v}) = -\infty$ y no hay *duality gap*. Suponemos entonces que γ es finito y consideremos el sistema

$$f(\boldsymbol{x}) - \gamma < 0,\ \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0},\, \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}, \quad \text{con } \boldsymbol{x} \in S.$$

Por definición de γ, este sistema no tiene solución. Los supuestos sobre S, f, $\boldsymbol{g}$ y $\boldsymbol{h}$ permiten aplicar el Lema 5.18. Por tanto, existe $(u_0, \boldsymbol{u}, \boldsymbol{v}) \neq \boldsymbol{0}$ con $(u_0, \boldsymbol{u}) \geq \boldsymbol{0}$ tal que

$$u_0(f(\boldsymbol{x}) - \gamma)) + \boldsymbol{u}^\intercal \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{v}^\intercal \boldsymbol{h}(\boldsymbol{x}) \geq 0 \quad \text{para todo } \boldsymbol{x} \in S. \tag{5.5}$$

A continuación probamos que la condición de regularidad asegura que $u_0 > 0$. Supongamos que $u_0 = 0$ y recordemos que $\hat{\boldsymbol{x}} \in S$ es tal que $\boldsymbol{g}(\hat{\boldsymbol{x}}) < \boldsymbol{0}$ y $\boldsymbol{h}(\hat{\boldsymbol{x}}) = \boldsymbol{0}$. Entonces, si ponemos $\hat{\boldsymbol{x}}$ en la Ecuación (5.5) obtenemos que $\boldsymbol{u}^\intercal \boldsymbol{g}(\boldsymbol{x}) \geq 0$. Como $\boldsymbol{g}(\hat{\boldsymbol{x}}) < \boldsymbol{0}$, necesariamente $\boldsymbol{u} = \boldsymbol{0}$. Entonces, $\boldsymbol{v}^\intercal \boldsymbol{h}(\boldsymbol{x}) \geq 0$ para todo $\boldsymbol{x} \in S$. En particular, como $\boldsymbol{0}$ está en el interior de $\boldsymbol{h}(S)$, existe $\boldsymbol{x}^* \in S$ tal que $\boldsymbol{h}(\boldsymbol{x}^*) = -\lambda \boldsymbol{v}$ con $\lambda > 0$. Por tanto, $0 \leq \boldsymbol{v}^\intercal \boldsymbol{h}(\boldsymbol{x}^*) = -\lambda\|\boldsymbol{v}\|^2$ lo que implica que $\boldsymbol{v} = \boldsymbol{0}$. Hemos probado que $u_0 = 0$ implica que $(u_0, \boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{0}$, con lo que tenemos una contradicción.

Dado que $u_0 > 0$, podemos dividir por u_0. Si definimos $\bar{\boldsymbol{u}} = \frac{\boldsymbol{u}}{u_0}$ y $\bar{\boldsymbol{v}} = \frac{\boldsymbol{v}}{u_0}$, tenemos

$$f(\boldsymbol{x}) + \bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\boldsymbol{x}) + \bar{\boldsymbol{v}}^\intercal \boldsymbol{h}(\boldsymbol{x}) \geq \gamma \quad \text{para todo } \boldsymbol{x} \in S.$$

Por tanto, $\mathcal{L}^D(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = \inf_{x\in S} f(\boldsymbol{x}) + \bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\boldsymbol{x}) + \bar{\boldsymbol{v}}^\intercal \boldsymbol{h}(\boldsymbol{x}) \geq \gamma$. Ahora bien, por el Teorema de dualidad débil (Teorema 5.13), $\mathcal{L}^D(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \leq \gamma$, con lo que $\mathcal{L}^D(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = \gamma$ y hemos demostrado que no hay *duality gap* y $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ resuelve el dual.

Por último, si $\bar{\boldsymbol{x}} \in S$ resuelve el primal, *i.e.*, $f(\bar{\boldsymbol{x}}) = \gamma$, $\boldsymbol{g}(\bar{\boldsymbol{x}}) \leq \boldsymbol{0}$ y $\boldsymbol{h}(\bar{\boldsymbol{x}}) = \boldsymbol{0}$, entonces la Ecuación (5.5) implica que $\bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) \geq 0$. Como $\bar{\boldsymbol{u}} \geq \boldsymbol{0}$ y $\boldsymbol{g}(\bar{\boldsymbol{x}}) \leq \boldsymbol{0}$, entonces $\bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) = 0$. □

La condición de que $\mathbf{0}$ pertenezca al interior de $\boldsymbol{h}(S)$ no es muy restrictiva. En particular, si $S = \mathbb{R}^n$ y $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$, entonces es fácil ver que si $\boldsymbol{A}$ tiene rango máximo (en otro caso podemos eliminar las restricciones afines redundantes) tenemos que $h(\mathbb{R}^n) = \mathbb{R}^l$ con lo que $\mathbf{0}$ está en el interior de $\boldsymbol{h}(S)$. Por otro lado, la condición de regularidad simplemente exige que haya algún punto factible que sea interior a todas las restricciones de desigualdad.

El Teorema de dualidad débil nos da una condición suficiente para poder obtener el valor óptimo del problema primal a partir de la resolución del dual. Desafortunadamente, dicha condición requiere que el problema primal sea un problema de optimización convexa, que es justo la clase de problemas donde tenemos garantizado que optimalidad local implica optimalidad global y resulta más sencillo resolver el primal. Aún así, se trata de un resultado de gran utilidad práctica pues muchos algoritmos iterativos resuelven simultáneamente P y D, lo que permite ir cerrando el *gap* entre las cotas superiores del valor óptimo de P obtenidas de las soluciones factibles de P y las cotas inferiores obtenidas a partir de soluciones factibles de D. La diferencia entre estas cotas da una información muy relevante acerca de la calidad de la mejor solución de P obtenida hasta una determinada iteración y permite establecer criterios de parada muy efectivos. Además, hay clases de problemas convexos donde la resolución simultánea de primal y dual ha probado ser especialmente efectiva, dando lugar a algoritmos muy eficientes en problemas de flujo en redes como el Algoritmo de Dijkstra para el problema del camino más corto y el Método húngaro para el problema de asignación.

5.4.5 Puntos de silla, dualidad y condiciones de KKT

El Teorema de dualidad fuerte (Teorema 5.19) presenta una condición suficiente para que el valor óptimo de los problemas primal y dual coincidan. Dicha condición consiste en la convexidad del problema P, complementada con una condición de regularidad. En este apartado vamos a presentar una condición necesaria y suficiente para dicha ausencia de *duality gap*: que la función lagrangiana $L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{x}) + \boldsymbol{u}^\intercal \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{v}^\intercal \boldsymbol{h}(\boldsymbol{x})$ tenga un *punto de silla*.

Definición 5.2. Dado el problema primal de la Ecuación (5.2), decimos que el punto $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ con $\bar{\boldsymbol{x}} \in S$ y $\bar{\boldsymbol{u}} \geq \mathbf{0}$ es un *punto de silla* de la función lagrangiana $L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$ si

$$L(\bar{\boldsymbol{x}}, \boldsymbol{u}, \boldsymbol{v}) \leq L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \leq L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \quad \text{para todo } \boldsymbol{x} \in S \text{ y todo } \boldsymbol{u} \geq \mathbf{0}.$$

Dado el punto $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$, si cambio la componente del dual reduzco L y si cambio la del primal aumento L. En otras palabras, $\bar{\boldsymbol{x}}$ resuelve $\mathcal{L}^D(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ y $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ resuelve $\mathcal{L}^P(\bar{\boldsymbol{x}})$. En la Figura 5.11 representamos un punto de silla de la función $f(x_1, x_2) = x_1^2 - x_2^2$.

Recordemos ahora la interpretación del dual que presentamos en la Sección 5.4.2 como un juego en el que un jugador quiere minimizar la función lagrangiana eligiendo las "estrategias" $\boldsymbol{x}$ y el otro maximizarla eligiendo las "estrategias" $(\boldsymbol{u}, \boldsymbol{v})$. Bajo esta interpretación, es equivalente hablar de puntos de silla de la función lagrangiana y de equilibrios de Nash de dicho juego: dado un punto de silla $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$, ninguno de los dos jugadores puede ganar desviándose unilateralmente.

La condición de punto de silla no resulta de gran utilidad en la práctica pero, como veremos a continuación, resulta de gran utilidad teórica. La principal razón es que permite obtener de modo relativamente sencillo relaciones entre dichos puntos y la ausencia de *duality gap*, las

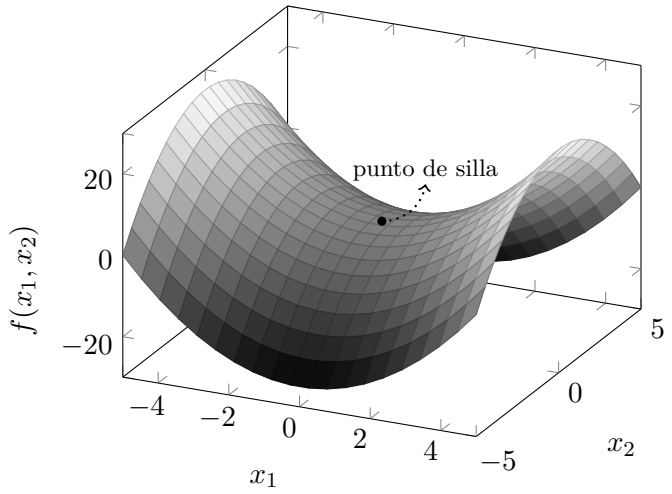


Figura 5.11: El $(0, 0)$ es un punto de silla de $f(x_1, x_2) = x_1^2 - x_2^2$; f crece en toda dirección $(d_1, 0)$, con $d_1 \neq 0$, y decrece en toda dirección $(0, d_2)$, con $d_2 \neq 0$.

condiciones de KKT y también la función perturbación. En cierta manera, nos encontramos ante cuatro formas complementarias de estudiar un problema primal controlando de forma explícita la tensión entre optimalidad y factibilidad.

Puntos de silla y ausencia de duality gap

El resultado siguiente me dice que una condición necesaria y suficiente para que $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ sea un punto de silla es que $\bar{\boldsymbol{x}}$ sea el óptimo del problema $\min_{\boldsymbol{x} \in S} L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ y además $\bar{\boldsymbol{x}}$ sea factible y cumpla holguras complementarias.

Teorema 5.20. *Sea $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ una solución con $\bar{\boldsymbol{x}} \in S$ y $\bar{\boldsymbol{u}} \geq \mathbf{0}$. Entonces, $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ es un punto de silla de $L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$ si y sólo si i) $L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = \min_{\boldsymbol{x} \in S} L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$, ii) $\boldsymbol{g}(\bar{\boldsymbol{x}}) \leq \mathbf{0}$ y $\boldsymbol{h}(\bar{\boldsymbol{x}}) = \mathbf{0}$ y iii) $\bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) = 0$.*

Demostración. "⇒" Si $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ es un punto de silla, entonces i) se cumple por definición. Trabajando ahora con la "optimalidad" de $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ dado $\bar{\boldsymbol{x}}$ tenemos que, para todo $(\boldsymbol{u}, \boldsymbol{v})$ con $\boldsymbol{u} \geq \mathbf{0}$, $f(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{v}}^\intercal \boldsymbol{h}(\bar{\boldsymbol{x}}) \geq f(\bar{\boldsymbol{x}}) + \boldsymbol{u}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) + \boldsymbol{v}^\intercal \boldsymbol{h}(\bar{\boldsymbol{x}})$. Como $\boldsymbol{u}$ y $\boldsymbol{v}$ pueden tomarse arbitrariamente grandes, $\boldsymbol{g}(\bar{\boldsymbol{x}}) \leq \mathbf{0}$ y $\boldsymbol{h}(\bar{\boldsymbol{x}}) = \mathbf{0}$, con lo que tenemos ii). Si además tomamos $\boldsymbol{u} = \mathbf{0}$ y $\boldsymbol{v} = \bar{\boldsymbol{v}}$ en la desigualdad anterior, obtenemos $\bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) \geq 0$. Ahora bien, como $\boldsymbol{u} \geq \mathbf{0}$ y $\boldsymbol{g}(\bar{\boldsymbol{x}}) \leq \mathbf{0}$, tenemos iii).

"⇐" Supongamos que $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ cumple i), ii) y iii). Por i) sabemos que, para todo $\boldsymbol{x} \in S$, $L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \leq L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$. Además,

$$L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = f(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{v}}^\intercal \boldsymbol{h}(\bar{\boldsymbol{x}}) \overset{ii),iii)}{=} f(\bar{\boldsymbol{x}}) \overset{ii),\boldsymbol{u} \geq \mathbf{0}}{\geq} f(\bar{\boldsymbol{x}}) + \boldsymbol{u}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) + \boldsymbol{v}^\intercal \boldsymbol{h}(\bar{\boldsymbol{x}}) = L(\bar{\boldsymbol{x}}, \boldsymbol{u}, \boldsymbol{v}),$$

con lo que $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ es un punto de silla. □

Nótese que a pesar de la similitud de las condiciones i), ii) y iii) con las condiciones de KKT, el resultado anterior no me dice que un punto de KKT se corresponda con un punto de silla. La razón es que, en general, el óptimo del problema $\min_{\boldsymbol{x}\in S} L(\boldsymbol{x},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ no tiene por qué ser un punto factible (las condiciones de KKT como mucho grantizan optimalidad local de este problema).

Teorema 5.21. *Sea $(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ una solución con $\bar{\boldsymbol{x}}\in S$ y $\bar{\boldsymbol{u}}\geq \mathbf{0}$. Entonces, $(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ es un punto de silla de $L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v})$ si y sólo si $\bar{\boldsymbol{x}}$ y $(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ son, respectivamente,* soluciones óptimas *de P y D sin* duality gap*, es decir, $f(\bar{\boldsymbol{x}})=\mathcal{L}^D(\boldsymbol{u},\boldsymbol{v})$.*

Demostración. "⇒" Supongamos que $(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ es un punto de silla. Como $\bar{\boldsymbol{u}}\geq\mathbf{0}$, $(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ es factible en D. A continuación usamos la equivalencia entre punto de silla y las condiciones i),ii) y iii) del Teorema 5.20. Como $\bar{\boldsymbol{x}}\in S$ y se cumple ii), $\bar{\boldsymbol{x}}$ es factible en P. Por iii), $L(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})=f(\bar{\boldsymbol{x}})$. Por otro lado, por i) tenemos que $\mathcal{L}^D(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})=\min_{\boldsymbol{x}\in S}L(\boldsymbol{x},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})=L(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ con lo que $f(\bar{\boldsymbol{x}})=\mathcal{L}^D(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ y el Corolario 5.15 asegura la optimalidad de $\bar{\boldsymbol{x}}$ y $(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ en P y D, respectivamente.

"⇐" Supongamos que $\bar{\boldsymbol{x}}$ y $(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ son soluciones óptimas en P y D, respectivamente con $f(\bar{\boldsymbol{x}})=\mathcal{L}^D(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$. Por la factiblidad de $\bar{\boldsymbol{x}}$ tenemos que $\boldsymbol{x}\in S$ y se cumple la condición ii) del Teorema 5.20. Además, usando la factibilidad de $\bar{\boldsymbol{x}}$ y $(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ tenemos

$$\mathcal{L}^D(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})=\min_{\boldsymbol{x}\in S}\{f(\boldsymbol{x})+\bar{\boldsymbol{u}}^\intercal\boldsymbol{g}(\boldsymbol{x})+\bar{\boldsymbol{v}}^\intercal\boldsymbol{h}(\boldsymbol{x})\}\leq f(\bar{\boldsymbol{x}})+\bar{\boldsymbol{u}}^\intercal\boldsymbol{g}(\bar{\boldsymbol{x}})+\bar{\boldsymbol{v}}^\intercal\boldsymbol{h}(\bar{\boldsymbol{x}})=f(\bar{\boldsymbol{x}})+\bar{\boldsymbol{u}}^\intercal\boldsymbol{g}(\bar{\boldsymbol{x}})\leq f(\bar{\boldsymbol{x}}).$$

Ahora bien, como sabemos que $f(\bar{\boldsymbol{x}})=\mathcal{L}^D(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$, las anteriores desigualdades son igualdades y tenemos que $\bar{\boldsymbol{u}}^\intercal\boldsymbol{g}(\bar{\boldsymbol{x}})=0$, con lo que se cumple iii) del Teorema 5.20. Además, $L(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})=f(\bar{\boldsymbol{x}})=\mathcal{L}^D(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})=\min_{\boldsymbol{x}\in S}L(\boldsymbol{x},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$, con lo que también se cumple iii). El Teorema 5.20 nos asegura ahora que $(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ es un punto de silla. □

Corolario 5.22. *Bajo las condiciones de convexidad y regularidad del Teorema de dualidad fuerte (Teorema 5.19), si $\bar{\boldsymbol{x}}$ es una solución óptima de P, entonces existe $(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ con $\bar{\boldsymbol{u}}\geq\mathbf{0}$ tal que $(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ es un punto de silla.*

Demostración. Por el Teorema de dualidad fuerte (Teorema 5.19), apartado (I), existe una solución óptima del problema D, $(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$, con $\bar{\boldsymbol{u}}\geq\mathbf{0}$. Además, no hay *duality gap*, con lo que $f(\bar{\boldsymbol{x}})=\mathcal{L}^D(\boldsymbol{u},\boldsymbol{v})$. Por tanto, el Teorema 5.21 nos asegura que $(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ es un punto de silla. □

Puntos de silla y condiciones de KKT

Recordemos que en el estudio de las condiciones de KKT resulta de gran utilidad el conjunto de restricciones activas en un cierto punto $\boldsymbol{x}$, $I(\boldsymbol{x})=\{i: g_i(\boldsymbol{x})=0\}$.

Teorema 5.23. *Supongamos que $\bar{\boldsymbol{x}}\in\mathring{S}$ cumple las condiciones de KKT del problema P de la Ecuación (5.2). Supongamos, además, que f y g_i con $i\in I(\bar{\boldsymbol{x}})$ son convexas y que cada h_j con $\bar{\boldsymbol{v}}_j\neq 0$ es afín. Entonces, $(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ es un punto de silla.*

Recíprocamente, si $(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$ con $\bar{\boldsymbol{x}}\in\mathring{S}$ y $\bar{\boldsymbol{u}}\geq\mathbf{0}$ es un punto de silla, entonces $\bar{\boldsymbol{x}}$ es factible en P y, además, $\bar{\boldsymbol{x}}$ es un punto KKT con multiplicadores dados por $(\bar{\boldsymbol{u}},\bar{\boldsymbol{v}})$.

Demostración. "⇒" Dado que $\bar{\boldsymbol{x}} \in \mathring{S}$, ninguna restricción que define S está activa en $\bar{\boldsymbol{x}}$ y las condiciones de KKT nos dicen que los multiplicadores asociados a dichas restricciones tomarán el valor 0. El Teorema 5.7 nos dice que existen $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ con $\bar{\boldsymbol{u}} \geq \mathbf{0}$ tales que

$$\nabla f(\bar{\boldsymbol{x}}) + \sum_{i=1}^{m} \bar{u}_i \nabla g_i(\bar{\boldsymbol{x}}) + \sum_{j=1}^{l} \bar{v}_j \nabla h_j(\bar{\boldsymbol{x}}) = \mathbf{0} \quad \text{y} \quad \bar{u}_i g_i(\bar{\boldsymbol{x}}) = 0 \ \text{ para todo } i \in \{1, \ldots, m\}.$$

En particular, la condición de holguras complementarias implica que, $\bar{u}_i = 0$ para todo $i \notin I(\bar{\boldsymbol{x}})$.

Por la convexidad de las funciones f y g_i con $i \in I(\bar{\boldsymbol{x}})$ y la afinidad de cada h_j con $\bar{v}_j \neq 0$ tenemos que, para todo $\boldsymbol{x} \in S$,

$$\begin{aligned} f(\boldsymbol{x}) &\geq f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}), \\ g_i(\boldsymbol{x}) &\geq g_i(\bar{\boldsymbol{x}}) + \nabla g_i(\bar{\boldsymbol{x}})^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}), \quad \text{para todo } i \in I(\bar{\boldsymbol{x}}), \text{ y} \\ h_j(\boldsymbol{x}) &= h_j(\bar{\boldsymbol{x}}) + \nabla h_j(\bar{\boldsymbol{x}})^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}), \quad \text{para todo } j \in \{1, \ldots, l\}, \text{ con } v_j \neq 0. \end{aligned}$$

Si ahora multiplicamos por $\bar{u}_i \geq 0$ la ecuación asociada a cada g_i y por $\bar{v}_j$ la ecuación asociada a cada h_j con $v_j \neq 0$, sumamos estas ecuaciones a la ecuación de la función f y agrupamos términos obtenemos que, para todo $\boldsymbol{x} \in S$,

$$L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) - L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \geq \Big(\nabla f(\bar{\boldsymbol{x}}) + \sum_{i=1}^{m} \bar{u}_i \nabla g_i(\bar{\boldsymbol{x}}) + \sum_{j=1}^{l} \bar{v}_j \nabla h_j(\bar{\boldsymbol{x}}) \Big)^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}) = 0,$$

donde la última igualdad no es más que la aplicación de la condición de KKT; con lo que ya tenemos que $L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \leq L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$. Por otro lado, dado que que $\boldsymbol{g}(\bar{\boldsymbol{x}}) \leq \mathbf{0}$, $\boldsymbol{h}(\bar{\boldsymbol{x}}) = \mathbf{0}$, y $\bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) = 0$, tenemos que

$$L(\bar{\boldsymbol{x}}, \boldsymbol{u}, \boldsymbol{v}) = f(\bar{\boldsymbol{x}}) + \boldsymbol{u}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) + \boldsymbol{v}^\intercal \boldsymbol{h}(\bar{\boldsymbol{x}}) \leq f(\bar{\boldsymbol{x}}) = L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}),$$

con lo que $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ es un punto de silla.

"⇐" Supongamos ahora que $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ es un punto de silla con $\bar{\boldsymbol{x}} \in \mathring{S}$ y $\bar{\boldsymbol{u}} \geq \mathbf{0}$. Por el Teorema 5.20, tenemos que $\boldsymbol{g}(\bar{\boldsymbol{x}}) \leq \mathbf{0}$, $\boldsymbol{h}(\bar{\boldsymbol{x}}) = \mathbf{0}$ y $\bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) = 0$, con lo que $\bar{\boldsymbol{x}}$ es factible para P. El Teorema 5.20 también nos dice que $L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = \min_{\boldsymbol{x} \in S} L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$. Como $\boldsymbol{x} \in \mathring{S}$, entonces

$$\mathbf{0} = \nabla_{\boldsymbol{x}} L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = \nabla f(\bar{\boldsymbol{x}}) + \sum_{i=1}^{m} \bar{u}_i \nabla g_i(\bar{\boldsymbol{x}}) + \sum_{j=1}^{l} \bar{v}_j \nabla h_j(\bar{\boldsymbol{x}}),$$

con lo que $\bar{\boldsymbol{x}}$ es un punto KKT con multiplicadores dados por $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$. □

Puntos de silla y función perturbación

Para terminar, nos apoyamos nuevamente en el concepto de punto de silla para formalizar alguna de las intuiciones presentadas en la Sección 5.4.3 para la función perturbación. Lo primero que tenemos que hacer es presentar la definición formal de función perturbación.

Definición 5.3. Dado el problema P de la Ecuación (5.2), definimos la función perturbación $\pi : \mathbb{R}^{m+l} \to \mathbb{R}$, para todo $(\boldsymbol{y}, \boldsymbol{q}) \in \mathbb{R}^{m+l}$, como

$$\pi(\boldsymbol{y}, \boldsymbol{q}) = \min_{\boldsymbol{x} \in S} \{f(\boldsymbol{x}) : \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{y},\ \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{q}\}.$$

El siguiente resultado muestra que, bajo el supuesto de que P tenga un óptimo, la ausencia de *duality gap* es equivalente a la existencia de un hiperplano soporte al epigrafo de π en el punto $(\mathbf{0}, \pi(\mathbf{0}))$.

Teorema 5.24. *Asumamos que el problema P tiene una solución óptima $\bar{\boldsymbol{x}}$. Entonces, $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ es un punto de silla si y sólo si, para todo $(\boldsymbol{y}, \boldsymbol{q}) \in \mathbb{R}^{m+l}$,*

$$\pi(\boldsymbol{y}, \boldsymbol{q}) \geq \pi(\mathbf{0}) - (\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})^\intercal (\boldsymbol{y}, \boldsymbol{q}), \tag{5.6}$$

o, equivalentemente, el hiperplano $H((\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}), z) = \{(\boldsymbol{y}, \boldsymbol{q}) \in \mathbb{R}^{m+l} : (\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})^\intercal (\boldsymbol{y}, \boldsymbol{q}) = z\}$ soporta $\mathrm{epi}(\pi)$ *en el punto $(\mathbf{0}, \pi(\mathbf{0}))$.*

Demostración. "⇒" Supongamos que $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ es un punto de silla. El Teorema 5.21 asegura que no hay *duality gap*, es decir, $f(\bar{\boldsymbol{x}}) = \mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v})$. Entonces, dado $(\boldsymbol{y}, \boldsymbol{q}) \in \mathbb{R}^{m+l}$, tenemos que

$$\begin{aligned}\pi(\mathbf{0}) = f(\bar{\boldsymbol{x}}) = \mathcal{L}^D(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) &= \min_{\boldsymbol{x} \in S} \{f(\boldsymbol{x}) + \bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\boldsymbol{x}) + \bar{\boldsymbol{v}}^\intercal \boldsymbol{h}(\boldsymbol{x})\} \\ &= (\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})^\intercal (\boldsymbol{y}, \boldsymbol{q}) + \min_{\boldsymbol{x} \in S} \Big\{ f(\boldsymbol{x}) + \sum_{i=1}^{m} \bar{\boldsymbol{u}}_i (g_i(\bar{\boldsymbol{x}}) - y_i) + \sum_{j=1}^{l} \bar{\boldsymbol{v}}_j (h_j(\bar{\boldsymbol{x}}) - q_j) \Big\}.\end{aligned}$$

Si aplicamos ahora el Teorema de dualidad débil (Teorema 5.13) al problema perturbado con $(\boldsymbol{y}, \boldsymbol{q})$, tenemos que el mínimo de la última igualdad es menor o igual que $\pi(\boldsymbol{y}, \boldsymbol{q})$. Por tanto, concluimos que $\pi(\boldsymbol{y}, \boldsymbol{q}) \geq \pi(\mathbf{0}) - (\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})^\intercal (\boldsymbol{y}, \boldsymbol{q})$.

"⇐" Supongamos ahora que $\bar{\boldsymbol{x}}$ es un óptimo de P y que $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ es tal que, para todo $(\boldsymbol{y}, \boldsymbol{q}) \in \mathbb{R}^{m+l}$, $\pi(\boldsymbol{y}, \boldsymbol{q}) \geq \pi(\mathbf{0}) - (\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})^\intercal (\boldsymbol{y}, \boldsymbol{q})$. Antes de nada veamos que $\bar{\boldsymbol{u}} \geq \mathbf{0}$. Supongamos que existe i tal que $\bar{u}_i < 0$. Entonces, tomando $(\boldsymbol{y}, \boldsymbol{q})$ tal que su única componente distinta de cero es $y_i > 0$ tendríamos $\pi(\mathbf{0}) \geq \pi(\boldsymbol{y}, \boldsymbol{q})$, por la monotonía de la función π. Por tanto, tenemos

$$\pi(\mathbf{0}) \geq \pi(\boldsymbol{y}, \boldsymbol{q}) \geq \pi(\mathbf{0}) - (\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})^\intercal (\boldsymbol{y}, \boldsymbol{q}),$$

de donde $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})^\intercal (\boldsymbol{y}, \boldsymbol{q}) \geq 0$, lo que a su vez implica que $\bar{u}_i y_i \geq 0$, lo que es imposible pues ambos términos son estrictamente negativos.

Tenemos ya que $\bar{\boldsymbol{u}} \geq \mathbf{0}$. A continuación vamos a probar que se cumplen las condiciones i), ii) y iii) del Teorema 5.20, que implican que $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ es un punto de silla. Como $\bar{\boldsymbol{x}}$ es factible, $\boldsymbol{g}(\bar{\boldsymbol{x}}) \leq \mathbf{0}$ y $\boldsymbol{h}(\bar{\boldsymbol{x}}) = \mathbf{0}$, con lo que se cumple ii). Además, si tomamos $(\bar{\boldsymbol{y}}, \bar{\boldsymbol{q}}) = (\boldsymbol{g}(\bar{\boldsymbol{x}}), \boldsymbol{h}(\bar{\boldsymbol{x}}))$ tenemos que el problema perturbado con $(\bar{\boldsymbol{y}}, \bar{\boldsymbol{q}})$ es más restringido que el problema P. Pero, como $\bar{\boldsymbol{x}}$ también es factible para dicho problema, $\pi(\bar{\boldsymbol{y}}, \bar{\boldsymbol{q}}) = \pi(\mathbf{0})$. Por la Ecuación (5.6), esto implica que $\bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) \geq \mathbf{0}$ pero, dado que $\boldsymbol{g}(\bar{\boldsymbol{x}}) \leq \mathbf{0}$ y $\bar{\boldsymbol{u}} \geq \mathbf{0}$, necesariamente $\bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) = 0$, con lo que se cumple iii).

Nos queda por probar que se cumple i), que dice que $L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = \min_{\boldsymbol{x} \in S} L(\boldsymbol{x}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$. Tenemos que, para todo $(\boldsymbol{y}, \boldsymbol{q}) \in \mathbb{R}^{m+l}$,

$$L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = f(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{v}}^\intercal \boldsymbol{h}(\bar{\boldsymbol{x}}) = f(\bar{\boldsymbol{x}}) = \pi(\mathbf{0}) \leq \pi(\boldsymbol{y}, \boldsymbol{q}) + (\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})^\intercal (\boldsymbol{y}, \boldsymbol{q}).$$

Dado $\hat{\boldsymbol{x}} \in S$ definimos $(\hat{\boldsymbol{y}}, \hat{\boldsymbol{q}}) = (\boldsymbol{g}(\hat{\boldsymbol{x}}), \boldsymbol{h}(\hat{\boldsymbol{x}}))$. Claramente, $\pi(\hat{\boldsymbol{y}}, \hat{\boldsymbol{q}}) \leq f(\hat{\boldsymbol{x}})$, pues $\hat{\boldsymbol{x}}$ es factible para el problema perturbado con $(\hat{\boldsymbol{y}}, \hat{\boldsymbol{q}})$. Usando esto en la anterior ecuación, obtenemos que, para todo $\hat{\boldsymbol{x}} \in S$, $L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \leq f(\hat{\boldsymbol{x}}) + \bar{\boldsymbol{u}}^\intercal \boldsymbol{g}(\hat{\boldsymbol{x}}) + \bar{\boldsymbol{v}}^\intercal \boldsymbol{h}(\hat{\boldsymbol{x}})$, como queríamos probar. □

El siguiente resultado relaciona la convexidad del problema primal con la convexidad de la función perturbación.

Proposición 5.25. *Dado el problema primal de la Ecuación (5.2) la función f y las funciones g_i son convexas, las funciones h_j afines y el conjunto S es convexo, entonces la función perturbación es convexa.*

••Ejercicio 5.9. Demuestra la Proposición 5.25. ◁

Una implicación de la combinación del Teorema 5.24 y la Proposición 5.25 es que, en problemas de optimización convexa, el vector $-(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ es un subgradiente de π en $\mathbf{0}$.

A continuación presentamos un último resultado, que contiene la formalización de la interpretación de los multiplicadores de Lagrange como tasas de cambio marginales de la función objetivo del problema P ante perturbaciones en los lados derechos de las restricciones.

Proposición 5.26. *Si $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ es un punto de silla asociado a P y D y la función perturbación π es continuamente diferenciable en $\mathbf{0}$, entonces*

$$\nabla \pi(\mathbf{0}) = -(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}).$$

Demostración. Como π es continuamente diferenciable en $\mathbf{0}$ podemos tomar la aproximación de Taylor de primer orden de π en ese punto, obteniendo que, para todo $(\boldsymbol{y}, \boldsymbol{q}) \in \mathbb{R}^{m+l}$

$$\pi(\boldsymbol{y}, \boldsymbol{q}) = \pi(\mathbf{0}) + \nabla \pi(\mathbf{0})^\intercal (\boldsymbol{y}, \boldsymbol{q}) + \varphi(\boldsymbol{y}, \boldsymbol{q}) \|(\boldsymbol{y}, \boldsymbol{q})\|, \quad \text{donde } \lim_{(\boldsymbol{y},\boldsymbol{q}) \to \mathbf{0}} \varphi(\boldsymbol{y}, \boldsymbol{q}) = 0.$$

Por el Teorema 5.24, $\pi(\boldsymbol{y}, \boldsymbol{q}) \geq \pi(\mathbf{0}) - (\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})^\intercal (\boldsymbol{y}, \boldsymbol{q})$. Restando las dos últimas ecuaciones tenemos que

$$-\Big(\nabla \pi(\mathbf{0}) + (\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \Big)^\intercal (\boldsymbol{y}, \boldsymbol{q}) \geq \varphi(\boldsymbol{y}, \boldsymbol{q}) \|(\boldsymbol{y}, \boldsymbol{q})\| \quad \text{para todo } (\boldsymbol{y}, \boldsymbol{q}) \in \mathbb{R}^{m+l}.$$

Tomemos ahora la sucesión $\{(\boldsymbol{y}^t, \boldsymbol{q}^t)\}_{t \in \mathbb{N}}$, donde $(\boldsymbol{y}^t, \boldsymbol{q}^t) = \frac{1}{t}\big(\nabla \pi(\mathbf{0}) + (\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \big)$. Entonces,

$$-\frac{1}{t} \|\nabla \pi(\mathbf{0}) + (\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})\|^2 \geq \varphi(\boldsymbol{y}^t, \boldsymbol{q}^t) \frac{1}{t} \|\nabla \pi(\mathbf{0}) + (\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})\|,$$

de donde obtenemos que $-\|\nabla \pi(\mathbf{0}) + (\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})\| \geq \varphi(\boldsymbol{y}^t, \boldsymbol{q}^t)$. Pero, dado que $\lim_{t \to \infty} \varphi(\boldsymbol{y}^t, \boldsymbol{q}^t) = 0$, la anterior desigualdad sólo es posible si $\nabla \pi(\mathbf{0}) = -(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$. □

La Proposición 5.26 asume la existencia de un punto de silla $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$. Dicho supuesto, combinado con los teoremas 5.21 y 5.23, nos asegura que $\bar{\boldsymbol{x}}$ y $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ son soluciones óptimas de primal y dual respectivamente y que $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ se corresponde los multiplicadores de Lagrange asociados a $\bar{\boldsymbol{x}}$. Por tanto, la ecuación $\nabla \pi(\mathbf{0}) = -(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ es la justificación matemática de la interpretación económica de los multiplicadores de Lagrange comentada en la Sección 5.3.5 y durante la interpretación geométrica de la dualidad lagrangiana en la Sección 5.4.3.

Recapitulando

Concluimos esta sección dedicada a los resultados matemáticos asociados a los puntos de silla recapitulando un poco lo que hemos aprendido con ellos. Los teoremas vistos en esta sección nos muestran que, bajo supuestos de convexidad, no hay diferencia entre hablar de los multiplicadores de Lagrange, de soluciones óptimas de primal y dual, de puntos de silla de la función lagrangiana o de subgradientes de la función perturbación en el origen.

Si no tenemos convexidad y no tenemos garantizada la existencia de puntos de silla, todas las conexiones aquí desarrolladas pierden mucha fuerza. Sin embargo, es importante comentar que, a nivel local, muchas propiedades siguen siendo ciertas. Por ejemplo, la interpretación económica de los multiplicadores de Lagrange, que es una propiedad local, se mantiene. A continuación presentamos informalmente la intuición detrás de esta afirmación. Sea $\bar{\boldsymbol{x}}$ un punto KKT del problema P, podemos trabajar con $LP(\bar{\boldsymbol{x}})$, la aproximación lineal de primer orden de P en $\bar{\boldsymbol{x}}$, estudiada brevemente en la Sección 5.3.8. El Teorema 5.12 nos asegura que $\bar{\boldsymbol{x}}$ es un óptimo global de $LP(\bar{\boldsymbol{x}})$, y de la demostración se sigue que los multiplicadores de Lagrange son los mismos que para el problema P. Como el problema linealizado es un problema convexo, los multiplicadores de Lagrange admiten la interpretación económica arriba descrita. Ahora bien, bajo las condiciones de regularidad adecuadas, cerca de $\bar{\boldsymbol{x}}$ el problema $LP(\bar{\boldsymbol{x}})$ es una buena aproximación de P, con lo que la interpretación económica también es válida para el problema original.

5.4.6 Resolución de dual y primal

Hasta ahora hemos presentado un gran número de resultados relativos a la dualidad lagrangiana, que deberían ser suficientes para comenzar a ver la gran importancia de la misma a nivel conceptual y también a nivel práctico. En este último punto, ya comentamos que muchos de los algoritmos más eficientes computacionalmente para distintas clases de problemas se apoyan de una u otra manera en la formulación dual de los problemas en estudio. En particular, los algoritmos iterativos que van calculando simultáneamente soluciones de primal y dual pueden ser muy efectivos gracias a la información proporcionada por las cotas. Recordemos que, dado un problema P de minimización, cada solución factible del primal nos da una cota superior del valor óptimo de la función objetivo y cada solución factible del dual nos da una cota inferior. En problemas de programación convexa la diferencia entre estas cotas convergerá a cero, proporcionando no sólo criterios de parada muy fiables, sino también información de la evolución del algoritmo con las iteraciones.

También desde un punto de vista práctico surgen múltiples cuestiones alrededor de la resolución del problema D ¿cómo de fácil es resolverlo? ¿es más fácil o más difícil que resolver el problema P? Una vez resuelto D, ¿podemos usarlo para recuperar una solución óptima de P? El resto de esta sección está dedicada a responder a estas y otras preguntas, aunque omitiremos buena parte de los desarrollos formales, que pueden consultarse en el libro Bazaraa y otros (2006).

Propiedades de la función dual

Para poder afrontar la resolución del problema dual es importante conocer primero sus propiedades, a lo que dedicamos este apartado. El primer resultado nos dice que la función dual es cóncava y, dado que estamos ante un problema de maximización, tendremos todas las buenas propiedades de la optimización convexa. En particular, todo óptimo local será un óptimo global.

Durante este apartado asumiremos que el conjunto S es compacto y, para facilitar la exposición, usaremos la notación $\boldsymbol{w} = (\boldsymbol{u}, \boldsymbol{v})$ y $\boldsymbol{\sigma} = (\boldsymbol{g}, \boldsymbol{h})$.

Proposición 5.27. *Dado el problema P de la Ecuación (5.2), si S es no vacío y compacto y las funciones f y $\boldsymbol{\sigma}$ son continuas, entonces la función dual $\mathcal{L}^D(\boldsymbol{w}) = \inf_{\boldsymbol{x}\in S}\{f(\boldsymbol{x}) + \boldsymbol{w}^\intercal\boldsymbol{\sigma}(\boldsymbol{x})\}$ es cóncava.*

Demostración. Como f y $\boldsymbol{\sigma}$ son continuas y S es compacto, el ínfimo en $\mathcal{L}^D(\boldsymbol{w})$ será siempre finito, con lo que $\mathcal{L}^D(\boldsymbol{w}) \in \mathbb{R}$ para todo $\boldsymbol{w} \in \mathbb{R}^{m+l}$. Dados $\boldsymbol{w}$ y $\hat{\boldsymbol{w}} \in \mathbb{R}^{m+l}$ y $\lambda \in (0,1)$, tenemos

$$\begin{aligned}
\mathcal{L}^D(\lambda\boldsymbol{w} + (1-\lambda)\hat{\boldsymbol{w}}) &= \inf_{\boldsymbol{x}\in S}\{f(\boldsymbol{x}) + (\lambda\boldsymbol{w} + (1-\lambda)\hat{\boldsymbol{w}})^\intercal\boldsymbol{\sigma}(\boldsymbol{x})\} \\
&= \inf_{\boldsymbol{x}\in S}\{\lambda\Big(f(\boldsymbol{x}) + \boldsymbol{w}^\intercal\boldsymbol{\sigma}(\boldsymbol{x})\Big) + (1-\lambda)\Big(f(\boldsymbol{x}) + \hat{\boldsymbol{w}}^\intercal\boldsymbol{\sigma}(\boldsymbol{x})\Big)\} \\
&\geq \lambda \inf_{\boldsymbol{x}\in S}\{f(\boldsymbol{x}) + \boldsymbol{w}^\intercal\boldsymbol{\sigma}(\boldsymbol{x})\} + (1-\lambda)\inf_{\boldsymbol{x}\in S}\{f(\boldsymbol{x}) + \hat{\boldsymbol{w}}^\intercal\boldsymbol{\sigma}(\boldsymbol{x})\} \\
&= \lambda\mathcal{L}^D(\boldsymbol{w}) + (1-\lambda)\mathcal{L}^D(\hat{\boldsymbol{w}}).
\end{aligned}$$

□

Aunque la demostración anterior es bastante sencilla, se podría obtener también a partir de las propiedades de las funciones convexas, apoyándonos en el hecho de que, para cada $\boldsymbol{x} \in S$, la función lagrangiana $L(\boldsymbol{x}, \boldsymbol{w}) = f(\boldsymbol{x}) + \boldsymbol{w}\boldsymbol{\sigma}(\boldsymbol{x})$ es una función afín en $\boldsymbol{w}$; y recordemos que toda función afín es cóncava y convexa. Por tanto, la función dual $\mathcal{L}^D(\boldsymbol{w}) = \inf_{\boldsymbol{x}\in S}\{f(\boldsymbol{x}) + \boldsymbol{w}^\intercal\boldsymbol{\sigma}(\boldsymbol{x})\} = -\sup_{\boldsymbol{x}\in S}\{-f(\boldsymbol{x}) - \boldsymbol{w}^\intercal\boldsymbol{\sigma}(\boldsymbol{x})\}$. Entonces, $-\mathcal{L}^D(\boldsymbol{w}) = \sup_{\boldsymbol{x}\in S}\{-f(\boldsymbol{x}) - \boldsymbol{w}^\intercal\boldsymbol{\sigma}(\boldsymbol{x})\}$ es convexa por ser supremo de convexas,[7] lo que implica que $\mathcal{L}^D(\boldsymbol{w})$ es cóncava.

Aunque la concavidad del problema dual es una buena noticia para su maximización, el principal problema sigue siendo que la evaluación de la función dual $\mathcal{L}^D(\boldsymbol{w})$ pasa por resolver un problema de minimización. A continuación vamos a ver que al menos el cálculo de los gradientes/subgradientes de dicha $\mathcal{L}^D$ en un punto $\boldsymbol{w}$ es relativamente sencillo. Para ello es necesario estudiar la diferenciabilidad de la función dual y los resultados que presentaremos a continuación será de utilidad el conjunto de soluciones óptimas de $\mathcal{L}^D(\boldsymbol{w})$:

$$S(\boldsymbol{w}) = \Big\{\bar{\boldsymbol{x}} : \bar{\boldsymbol{x}} \text{ es un óptimo global de } \inf_{\boldsymbol{x}\in S}\{f(\boldsymbol{x}) + \boldsymbol{w}^\intercal\boldsymbol{\sigma}(\boldsymbol{x})\}\Big\}.$$

Proposición 5.28. *Supongamos que S es no vacío y compacto y que las funciones f y $\boldsymbol{\sigma}$ son continuas. Dado, $\boldsymbol{w} \in \mathbb{R}^{m+l}$, si $S(\boldsymbol{w}) = \{\bar{\boldsymbol{x}}\}$, entonces $\mathcal{L}^D$ es diferenciable en $\boldsymbol{w}$ con gradiente*

$$\nabla\mathcal{L}^D(\boldsymbol{w}) = \boldsymbol{\sigma}(\bar{\boldsymbol{x}}).$$

[7] Por la Proposición 1.10, el máximo de una cantidad finita de funciones convexas es una función convexa, pero aquí tenemos una cantidad infinita de funciones. La generalización para el caso $f(\boldsymbol{x}) = \sup_{i\in I} f_i(\boldsymbol{x})$ donde las f_i son convexas es sencilla a partir del Teorema 1.14, que nos dice que una función es convexa si y sólo si su epigrafo es un conjunto convexo. Entonces, como $\operatorname{epi}(f) = \bigcap_{i\in I}\operatorname{epi} f_i$, tenemos que $\operatorname{epi}(f)$ es una intersección conjuntos convexos, con lo que claramente es convexo.

Este resultado nos dice que una condición suficiente para que la función dual sea diferenciable en un punto $\boldsymbol{w}$ es que el problema de optimización $\inf_{\boldsymbol{x}\in S}\{f(\boldsymbol{x})+\boldsymbol{w}^\intercal\boldsymbol{\sigma}(\boldsymbol{x})\}$ tenga un único óptimo $\bar{\boldsymbol{x}}$. Además, el gradiente se obtiene sin más que evaluar las restricciones de P en $\bar{\boldsymbol{x}}$. Esto no debería se sorprendente, pues si nos abstraemos del ínfimo, el gradiente de $f(\boldsymbol{x})+\boldsymbol{w}^\intercal\boldsymbol{\sigma}(\boldsymbol{x})$ en $\boldsymbol{w}$ es justamente $\boldsymbol{\sigma}(\boldsymbol{x})$. Intuitivamente, si para un $\boldsymbol{w}\in\mathbb{R}^{m+l}$ el óptimo $\bar{\boldsymbol{x}}$ es tal que una cierta restricción $g_i(\bar{\boldsymbol{x}})$ toma un valor positivo, entonces me interesará que el peso asociado $w_i=u_i$ crezca, ya que queremos maximizar $\mathcal{L}^D$.

A continuación presentamos el resultado relativo a los subgradientes de $\mathcal{L}^D$, incluyendo su demostración dada la sencillez de la misma.

Proposición 5.29. *Supongamos que S es no vacío y compacto y que las funciones f y $\boldsymbol{\sigma}$ son continuas. Dados $\boldsymbol{w}\in\mathbb{R}^{m+l}$ y $\bar{\boldsymbol{x}}\in S(\boldsymbol{w})$, entonces $\boldsymbol{\sigma}(\bar{\boldsymbol{x}})$ es un subgradiente de $\mathcal{L}^D$ en $\boldsymbol{w}$.*

Demostración. Dado $\boldsymbol{w}\in\mathbb{R}^{m+l}$, los supuestos sobre S y las funciones f y $\boldsymbol{\sigma}$ nos garantizan que $S(\boldsymbol{w})\neq\emptyset$. Entonces, dado $\bar{\boldsymbol{x}}\in S(\boldsymbol{w})$ y dado $\hat{\boldsymbol{w}}\in\mathbb{R}^{m+l}$

$$\begin{aligned}\mathcal{L}^D(\hat{\boldsymbol{w}}) &= \inf_{\boldsymbol{x}\in S}\{f(\boldsymbol{x})+\hat{\boldsymbol{w}}^\intercal\boldsymbol{\sigma}(\boldsymbol{x})\}\\ &\leq f(\bar{\boldsymbol{x}})+\hat{\boldsymbol{w}}^\intercal\boldsymbol{\sigma}(\bar{\boldsymbol{x}})\\ &= f(\bar{\boldsymbol{x}})+(\hat{\boldsymbol{w}}-\boldsymbol{w})^\intercal\boldsymbol{\sigma}(\bar{\boldsymbol{x}})+\boldsymbol{w}^\intercal\boldsymbol{\sigma}(\bar{\boldsymbol{x}})\\ &= \mathcal{L}^D(\boldsymbol{w})+(\hat{\boldsymbol{w}}-\boldsymbol{w})^\intercal\boldsymbol{\sigma}(\bar{\boldsymbol{x}}),\end{aligned}$$

con lo que $\boldsymbol{\sigma}(\bar{\boldsymbol{x}})$ es un subgradiente de $\mathcal{L}^D$ en $\boldsymbol{w}$.[8] □

Proposición 5.30. *Supongamos que S es no vacío y compacto y que las funciones f y $\boldsymbol{\sigma}$ son continuas. Dado $\boldsymbol{w}\in\mathbb{R}^{m+l}$, el subdiferencial de $\mathcal{L}^D$ en $\boldsymbol{w}$ viene dado por*

$$\operatorname{conv}\Big(\{\boldsymbol{\sigma}(\bar{\boldsymbol{x}}):\bar{\boldsymbol{x}}\in S(\boldsymbol{w})\}\Big).$$

Resolución del problema dual

Recordemos que, a la hora de definir la función dual lagrangiana, hay cierta libertad para elegir qué restricciones se llevan a la función objetivo y cuáles se dejan en el conjunto S. Esta elección debe tener en cuenta la *tensión* entre:

- Dificultad para evaluar $\mathcal{L}^D(\boldsymbol{u},\boldsymbol{v})$ para cada par $(\boldsymbol{u},\boldsymbol{v})$.
- Duality gap entre P y D.

Por ejemplo, si no subimos ninguna restricción, tenemos que P=D, no hay duality gap, pero evaluar $\mathcal{L}^D$ es equivalente a resolver el problema P. Es importante elegir bien la formulación del dual que resulta más adecuada para cada problema.

A la vista de los resultados relativos a los gradientes y subgradientes de $\mathcal{L}^D$, parece natural resolver el problema dual aplicando algún método de optimización sin restricciones visto en el Tema 4, como el método de máximo descenso (Sección 4.8.1) o el método de subgradiente

[8] Nótese que como estamos trabajando con una función cóncava, la desigualdad en la definición de subgradiente se invierte con respecto al caso de funciones convexas.

(Sección 4.9.1). Dada la concavidad de $\mathcal{L}^D$, estos métodos deberían converger a un óptimo global de $\mathcal{L}^D$. Supongamos que estamos usando un algoritmo iterativo que en cada paso determina una dirección de ascenso de $\mathcal{L}^D$ y realiza una búsqueda de línea en esa dirección para determinar la nueva solución. Entonces uno tiene que tener cuidado con las siguientes dificultades técnicas:

Pérdida de Factibilidad. El problema D no es puramente un problema sin restricciones, debido a la condición $\boldsymbol{u} \geq \mathbf{0}$. Si trabajamos con el método de gradiente y en una iteración concreta estamos en el punto $(\boldsymbol{u}, \boldsymbol{v})$ con $S(\boldsymbol{u}, \boldsymbol{v}) = \{\bar{\boldsymbol{x}}\}$, tenemos que $\nabla \mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{g}(\bar{\boldsymbol{x}}), \boldsymbol{h}(\bar{\boldsymbol{x}}))$. Si para algún $i \in \{1, \ldots, m\}$ tenemos $u_i = 0$ y $g_i(\bar{\boldsymbol{x}}) < 0$, entonces $u_i + \lambda g_i(\bar{\boldsymbol{x}}) < 0$ para todo $\lambda > 0$, violando la restricción de no negatividad de u_i.

Una posible solución para este problema pasa por usar direcciones proyectadas, de la forma $(\hat{\boldsymbol{g}}(\bar{\boldsymbol{x}}), \boldsymbol{h}(\bar{\boldsymbol{x}}))$, donde

$$\hat{g}_i(\boldsymbol{x}) = \begin{cases} g_i(\boldsymbol{x}) & \text{si } u_i > 0 \\ \max\{0, g_i(\boldsymbol{x})\} & \text{si } u_i = 0. \end{cases}$$

No diferenciabilidad de $\mathcal{L}^D$. Obliga a trabajar con subgradientes. El problema es que en la práctica no resulta fácil encontrar subgradientes que sean direcciones de ascenso.

Debido a estas y otras dificultades, es habitual que se recurra a versiones especializadas de los métodos de gradiente/subgradiente y de otras técnicas de optimización no diferenciable como los métodos de planos de corte. En caso de que la estructura del problema lo permita, también es habitual usar técnicas de descomposición para resolver el dual.

Resolución del problema primal

Una vez discutidos los aspectos más relevantes de la resolución del problema dual, cabe preguntarse hasta qué punto esto resulta de utilidad para resolver el primal. Si P es un problema de optimización convexa, entonces el Teorema de dualidad fuerte (Teorema 5.19) nos asegura que no hay duality gap. Además, en este caso se puede obtener un óptimo (global) de P a partir de un óptimo de D sin más que resolver un cierto problema de programación lineal.

Por otro lado, cuando el problema P no es un problema de optimización convexa, que es cuando resulta más difícil resolverlo, ya resulta mucho más difícil obtener un óptimo (local) de P a partir de un óptimo de D.

Como conclusión podemos decir que, aunque en muchas clases de problemas el dual resulta de gran utilidad, no hace "milagros".

5.5 Aplicaciones de la dualidad y de las condiciones de KKT

5.5.1 Descomposición mediante dualidad

A la hora de resolver problemas de optimización en la práctica, es habitual que estos tengan una estructura que, manipulándolos adecuadamente, permita descomponerlos en problemas más sencillos. Es por esto que el estudio de técnicas de descomposición, que son el centro del Tema 6, es un campo de investigación de gran actividad y de mucha relevancia en aplicaciones reales,

donde cada vez es más habitual encontrarse problemas de gran tamaño y cuya descomposición en problemas más pequeños puede acelerar drásticamente los tiempos de resolución.

Relajación lagrangiana: Descomposición mediante dualidad lagrangiana

La relajación lagrangiana es una de las técnicas más habituales para descomponer problemas de programación no lineal. A continuación ilustramos su idea en una clase de problemas, relativamente frecuentes en la práctica, para la cual la resolución del dual puede ser de gran utilidad.[9] Como es habitual, partimos de un problema primal de la forma

$$\begin{array}{ll} \textbf{Problema primal P} & \\ \text{minimizar} & f(\boldsymbol{x}) \\ \quad \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 \qquad i = 1, \ldots, m \\ & h_j(\boldsymbol{x}) = 0 \qquad j = 1, \ldots, l \\ & \boldsymbol{x} \in S. \end{array}$$

Supongamos ahora que el conjunto S y las funciones que definen objetivo y restricciones son separables de alguna manera. En el caso más sencillo tendríamos que existen $\boldsymbol{a} \in \mathbb{R}^n$ y $\boldsymbol{b} \in \mathbb{R}^n$ tales que $S = \prod_{k=1}^n [a_k, b_k]$ y, además,

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{k=1}^n f^k(x_k) \\ g_i(\boldsymbol{x}) &= \sum_{k=1}^n g_i^k(x_k), \quad i = 1, \ldots, m \\ h_j(\boldsymbol{x}) &= \sum_{k=1}^n h_j^k(x_k), \quad j = 1, \ldots, l. \end{aligned}$$

En este caso, la función lagrangiana también se puede descomponer como

$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{x}) + \sum_{i=1}^m u_i g_i(\boldsymbol{x}) + \sum_{j=1}^l v_j h_j(\boldsymbol{x}) = \sum_{k=1}^n \left(f^k(x_k) + \sum_{i=1}^m u_i g_i^k(x_k) + \sum_{j=1}^l v_j h_j^k(x_k) \right).$$

Esta separabilidad en las variables $\boldsymbol{x}$ puede ser aprovechada para facilitar la evaluación de la función dual:

$$\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v}) = \inf_{\boldsymbol{x} \in S} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = \sum_{k=1}^n \inf_{x_k \in [a_k, b_k]} \left(f^k(x_k) + \sum_{i=1}^m u_i g_i^k(x_k) + \sum_{j=1}^l v_j h_j^k(x_k) \right),$$

con lo que el cálculo de la misma se descompone en el cálculo de n subproblemas. A veces estos subproblemas podrán ser resueltos de forma explícita y, en otros casos, el simple hecho de poder resolver los subproblemas de forma independiente puede dar lugar a un dual más fácil de resolver que el primal. Si además tenemos que estamos ante un problema de optimización convexa, entonces, como sabemos por el Teorema de dualidad fuerte (Teorema 5.19) que no hay

[9] El contenido de este apartado está basado en la Sección 4.4 del libro Ruszczyński (2006).

duality gap, resolver el dual puede ser manera más eficiente de encontrar la solución óptima del primal. En particular, en la situación que acabamos de describir, la clave radica en que evaluar $\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v})$ pasa por resolver n problemas de optimización unidimensionales. Un aspecto de gran importancia a la hora de aplicar este tipo de ideas consiste en determinar bien qué restricciones se pasan a la función objetivo, se "dualizan", y cuáles se dejan en el conjunto S.

En el Tema 6 presentaremos una introducción a las técnicas de descomposición aunque, para hacer la exposición más accesible, estará centrada principalmente en problemas de programación lineal.

Ejemplo de descomposición: Un problema de planificación energética

Supongamos que tenemos n plantas energéticas que, conjuntamente, tienen que satisfacer una demanda de $d > 0$ unidades. Denotamos por x_k a la cantidad de energía a generar por la planta k, que suponemos puede generar entre 0 y $b_k \geq 0$ unidades de energía con un coste dado por

$$f^k(x_k) = c_k x_k + \frac{q_k}{2} x_k^2.$$

Asumimos que, para todo $k \in \{1, \ldots, n\}$, $c_k > 0$ y $q_k > 0$, con lo que las funciones de coste son estrictamente convexas. Suponemos ahora que $S = \prod_{k=1}^n [0, b_k]$, con $\sum_{k=1}^n b_k \geq d$ para asegurar que el problema tiene soluciones factibles. Entonces, el problema de optimización de satisfacer toda la demanda a mínimo coste puede escribirse como

$$\begin{array}{rl} & \textbf{Problema primal P} \\ \text{minimizar} & \displaystyle\sum_{k=1}^n f^k(x_k) \\ \text{sujeto a} & \displaystyle\sum_{k=1}^n x_k \geq d \\ & \boldsymbol{x} \in S. \end{array}$$

Como S es compacto, el problema tiene alguna solución óptima. Además, al tratarse de un problema de optimización convexa, el Teorema de dualidad fuerte (Teorema 5.19) asegura que no hay duality gap. Por tanto, podemos resolver el dual para encontrar el óptimo del primal. La función lagrangiana asociada a este problema viene dada por

$$L(\boldsymbol{x}, u) = \sum_{k=1}^n f^k(x_k) + u(d - \sum_{k=1}^n x_k),$$

con lo que la función dual es

$$\mathcal{L}^D(u) = ud + \min_{\boldsymbol{x} \in S} \sum_{k=1}^n (f^k(x_k) - u x_k) = ud + \sum_{k=1}^n \min_{\boldsymbol{x}_k \in [0, b_k]} (f^k(x_k) - u x_k).$$

Para cada $k \in \{1, \ldots, n\}$ podemos estudiar el subproblema de minimización

$$\mathcal{L}_k^D(u) = \min_{\boldsymbol{x}_k \in [0, b_k]} (f^k(x_k) - u x_k) = \min_{\boldsymbol{x}_k \in [0, b_k]} (c_k x_k + \frac{q_k}{2} x_k^2 - u x_k).$$

Cada supbroblema $\mathcal{L}_k^D(u)$ admite la siguiente interpretación: dado un precio u recibido por unidad de energía producida, encontrar la producción que minimiza las pérdidas (maximiza los beneficios) de la planta k. Como $\mathcal{L}_k^D(u)$ es un problema de optimización cuadrático en una variable, puede ser resuelto mediante cálculo elemental, obteniendo la solución óptima

$$\bar{x}_k(u) = \begin{cases} 0 & \text{si } 0 \leq u \leq c_k \\ \frac{u-c_k}{q_k} & \text{si } c_k \leq u \leq c_k + q_k b_k \\ b_k & \text{si } u \geq c_k + q_k b_k. \end{cases}$$

Por tanto, al poder resolver de forma explícita todos los subproblemas, también podemos evaluar de forma explícita la función dual $\mathcal{L}_k^D(u)$. En particular, vemos que la energía producida por una planta es una función no decreciente del multiplicador u: a mayor precio mayor producción. Como el primal tiene una solución óptima y se cumplen las condiciones del Teorema de dualidad fuerte (Teorema 5.19), también existe también una solución óptima del dual, $\bar{u}$. Además, el vector $\bar{\boldsymbol{x}}(\bar{u}) = (\bar{x}_1(\bar{u}), \ldots, \bar{x}_n(\bar{u}))$ será una solución óptima del problema P. Esto implica que $\bar{\boldsymbol{x}}$ tiene que cumplir la restricción de abastecimiento $\sum_{k=1}^n x_k \geq d$. Además, nunca será óptimo producir por encima del nivel mínimo, con lo que la restricción se cumplirá con igualdad. Esto asegura que se cumpla la condición de holguras complementarias, $\bar{u}(\sum_{k=1}^n \bar{x}_k(\bar{u}) - d) = 0$. Entonces, en el óptimo tendremos que

$$\sum_{k=1}^n \bar{x}_k(\bar{u}) = d.$$

Como tenemos una fórmula explícita para $\bar{x}_k(\bar{u})$ que únicamente depende de la variable $\bar{u}$, despejando en la anterior ecuación obtendremos el óptimo del dual, $\bar{u}$, a partir del que después calcularemos el óptimo del problema original, $\bar{\boldsymbol{x}}(\bar{u})$. Intuitivamente, dado que las funciones $\bar{x}_k(u)$ son no decrecientes, el cálculo de $\bar{u}$ se puede hacer sin más que determinar el precio al cual se producen exactamente las d unidades de energía que se demandan. En este ejemplo se ve muy bien el papel de la penalización impuesta por el multiplicador de Lagrange, equilibrando la tensión entre optimalidad y factibilidad.

Apoyándonos en la expresión explícita de los $\bar{x}_k(u)$ podemos obtener, para cada $k \in \{1, \ldots, n\}$, la expresión explícita de $\mathcal{L}_k^D(u)$:

$$\mathcal{L}_k^D(u) = \begin{cases} 0 & \text{si } 0 \leq u \leq c_k \\ -\frac{(u-c_k)^2}{2q_k} & \text{si } c_k \leq u \leq c_k + q_k b_k \\ (c_k - u)b_k + \frac{q_k b_k^2}{2} & \text{si } u \geq c_k + q_k b_k. \end{cases}$$

Dado que, para todo $u \geq 0$, tenemos que $\mathcal{L}_k^D(u) \leq 0$, vemos que ninguna planta sufrirá pérdidas. Esto es natural, pues en los subproblemas siempre se puede elegir no producir. Además, siempre que una planta produce obtiene beneficios.

En mercados energéticos regulados, donde hay multitud de productores de energía conectados a una misma red, surgen problemas similares al que acabamos de describir, aunque de una mayor complejidad. En estos problemas, el regulador del mercado (la CNMC en España) puede resolver el problema completo, que involucra características operativas propias de cada

productor (y que determinan su función de coste) sin necesidad de acceder a esa información privada. Para ello, el regulador puede solicitar a las distintas empresas información acerca del nivel de producción que alcanzarían para distintos valores del precio sombra u y así resolver el problema original. En la práctica es habitual que, una vez resuelto el problema, el precio sombra óptimo obtenido se utilice como precio de referencia en subastas de compra-venta de la energía en cuestión.

5.5.2 Resolución de problemas multinivel

Es relativamente frecuente encontrarse problemas reales cuya modelización pasa por la formulación de problemas multinivel, en los que aparecen múltiples problemas de optimización anidados y cuya resolución puede requerir el uso de técnicas específicas.

En este apartado pretendemos mostrar cómo la dualidad o, en este caso, directamente las condiciones de KKT, pueden ser de gran utilidad a la hora de transformar un problema multinivel en un problema clásico de un único nivel, como los que hemos estudiado hasta ahora. Para ello nos centraremos en problemas binivel y, más concretamente, en un ejemplo muy próximo a la teoría de juegos, donde estos problemas binivel aparecen frecuentemente.

Problemas binivel

La formulación clásica de un problema de optimización binivel viene dada por

$$\begin{array}{rlll} \text{minimizar}_{\boldsymbol{x},\boldsymbol{y}} & f^{\mathrm{S}}(\boldsymbol{x},\boldsymbol{y}) & & \\ \text{sujeto a} & g_i^{\mathrm{S}}(\boldsymbol{x},\boldsymbol{y}) \leq 0 & & i=1,\dots,m^{\mathrm{S}} \\ & h_j^{\mathrm{S}}(\boldsymbol{x},\boldsymbol{y}) = 0 & & j=1,\dots,l^{\mathrm{S}} \\ & \boldsymbol{y} \in \operatorname{argmin}_{\boldsymbol{y}} & f^{\mathrm{I}}(\boldsymbol{x},\boldsymbol{y}) & \\ & \text{sujeto a} & g^{\mathrm{I}}(\boldsymbol{x},\boldsymbol{y}) \leq 0 & i=1,\dots,m^{\mathrm{I}} \\ & & h^{\mathrm{I}}(\boldsymbol{x},\boldsymbol{y}) = 0 & j=1,\dots,l^{\mathrm{I}}. \end{array}$$

En esta formulación tenemos dos problemas anidados:

Nivel superior. Tenemos un problema con función objetivo f^{S} y restricciones $\boldsymbol{g}^{\mathrm{S}}$ y $\boldsymbol{h}^{\mathrm{S}}$. Además, en el problema superior tenemos una restricción adicional, que nos dice que, fijadas las variables $\boldsymbol{x}$, las variables $\boldsymbol{y}$ deben ser un óptimo del problema inferior.

Nivel inferior. Tenemos un problema con función objetivo f^{I} y restricciones $\boldsymbol{g}^{\mathrm{I}}$ y $\boldsymbol{h}^{\mathrm{I}}$.

El problema que surge entonces es el de cómo resolver este problema de optimización binivel en el que la región factible del problema superior depende de la resolución del problema de optimización del nivel inferior. Una de las situaciones más sencillas que se pueden plantear es aquella en la que este último problema sea un problema de optimización convexa en $\boldsymbol{y}$ (para todo valor de $\boldsymbol{x}$). En este caso, las condiciones de optimalidad de KKT serán condiciones necesarias y suficientes de optimalidad global, con lo que podremos reemplazar el problema de nivel inferior por sus condiciones de KKT. En el siguiente apartado presentamos una aplicación en la que este enfoque puede resultar de gran utilidad.

También conviene mencionar que, en el caso de que el problema del nivel inferior sea además lineal, se puede reemplazar dicho problema por las restricciones de factibilidad de primal y

dual, además de una restricción que exija que la función objetivo tome el mismo valor en los problemas primal y dual. Esto, unido al Teorema de dualidad fuerte, asegurará la optimalidad de las variables primales y duales.

Ejemplo de problema binivel: Duopolio de Stackelberg

En este apartado vamos a hacer pequeño análisis de competencia entre empresas clásico en teoría de juegos, adaptado de distintos ejemplos discutidos en González-Díaz y otros (2010). Presentaremos un modelo sencillo, pues el objetivo es el de presentar una herramienta metodológica para problemas binivel, que por supuesto podría aplicarse en situaciones mucho más complejas y realistas que la aquí estudiada. Más concretamente, tendremos una situación en la que las distintas empresas tienen que determinar la cantidad a producir de un cierto producto, con el objetivo de maximizar los beneficios finales.

Supongamos que la demanda máxima del mercado viene dada por $d > 0$ unidades del producto en cuestión y que su precio, como función de dicha demanda máxima y de la producción total, viene dado por $\pi_d(p) = \text{máx}\{0, d - p\}$ euros. Supongamos además que producir cada unidad del producto tiene un coste de c euros, con $0 < c < d$. Lo que haremos a continuación será estudiar y comparar las soluciones obtenidas en los siguientes escenarios:

ESCENARIO I: Monopolio. Hay una única empresa en el mercado.

ESCENARIO II: Duopolio de Cournot. Dos empresas compiten eligiendo de manera simultánea e independiente la cantidad a producir. Este modelo es considerado uno de los precursores de la teoría de juegos moderna, publicado por Cournot (1838) en el siglo XIX.

ESCENARIO III: Duopolio de Stackelberg. Similar al anterior, pero ahora una de las empresas actúa como líder, eligiendo primero la cantidad a producir. A la vista de la decisión de la empresa líder, la segunda empresa (seguidora) decide su propia producción. ¿Qué empresa saldrá favorecida? ¿líder o seguidora? Estas situaciones fueron planteadas por primera vez por von Stackelberg (1934). Como veremos, es en este modelo en el que aparecen los problemas binivel.

ESCENARIO I: Monopolio.

Supongamos que tenemos un monopolio con una demanda máxima dada por $d > 0$. Claramente, nunca será óptimo para el monopolista producir $x > d$ unidades, pues incurriría en un coste de producción $cx > 0$ y el precio de venta sería 0 euros por unidad. En otro caso, el beneficio del monopolista con una producción $x \leq d$ vendrá dado por $f_d(x) = \pi_d(x)x - cx = (d - x)x - cx = (d - c)x - x^2$. Por tanto, la estrategia óptima del monopolista pasará por maximizar $f_d(x)$ con $x \in [0, d]$. Ahora bien, como $f_d''(x) = -2$ para todo x, tenemos una función cóncava y, por tanto, la condición $f_d'(x) = 0$ es una condición suficiente de optimalidad global. Como $f_d'(x) = (d - c) - 2x$ tenemos que $f_d'(x) = 0$ si y sólo si $x = (d - c)/2$.

Por tanto, en el caso de un monopolista tenemos que la producción óptima es $\bar{x} = (d-c)/2$, con un beneficio asociado para el monopolista dado por $f_d(\bar{x}) = (d - c)^2/4$. Por otro lado, el precio de venta por unidad de producto sería $(d + c)/2$.

ESCENARIO II: Duopolio de Cournot.

Procedemos ahora a estudiar el escenario en el que ambas empresas eligen sus niveles de producción simultáneamente. Tenemos que el beneficio de una empresa depende tanto del nivel de producción elegido por ella como del nivel de producción elegido por la otra empresa. Esto se corresponde con un juego no cooperativo con dos jugadores. Tanto el jugador 1 como el jugador 2 han de elegir sus estrategias, $x_1 \in [0, d]$ y $x_2 \in [0, d]$, y tienen funciones de beneficio asociadas dadas por

$$\begin{aligned} f_d^1(x_1, x_2) &= \pi_d(x_1 + x_2)x_1 - cx_1 = \text{máx}\{0, (d - x_1 - x_2)\}x_1 - cx_1 \quad \text{y} \\ f_d^2(x_1, x_2) &= \pi_d(x_1 + x_2)x_2 - cx_2 = \text{máx}\{0, (d - x_1 - x_2)\}x_2 - cx_2. \end{aligned}$$

En este juego estaremos interesados en buscar *equilibrios de Nash* (Nash, 1950, 1951), es decir, estrategias $\bar{x}_1$ y $\bar{x}_2$ tal que ningún jugador pueda ganar desviándose unilateralmente de las mismas. Formalmente, queremos encontrar valores $\bar{x}_1$ y $\bar{x}_2$ tales que

$$\begin{aligned} f_d^1(\bar{x}_1, \bar{x}_2) &\geq f_d^1(x_1, \bar{x}_2) \text{ para todo } x_1 \in [0, d] \quad \text{y} \\ f_d^2(\bar{x}_1, \bar{x}_2) &\geq f_d^2(\bar{x}_1, x_2) \text{ para todo } x_2 \in [0, d]. \end{aligned}$$

Por tanto, en un equilibrio de Nash ambos jugadores están maximizando dada la estrategia del rival. Llamemos $\text{BR}_i(x_j)$ a la mejor respuesta (*Best Reply*) del jugador i ante una producción x_j de su rival. Si el jugador 1 toma como dada una estrategia $x_2 \in [0, d]$ del jugador 2, es fácil ver que su problema de maximización es equivalente al problema de un monopolista en un mercado con demanda máxima dada por $d - x_2$. Por tanto, por lo visto en el estudio de la situación de monopolio, su respuesta óptima sería $\text{BR}_1(x_2) = (d - x_2 - c)/2$. Del mismo modo, $\text{BR}_2(x_1) = (d - x_1 - c)/2$. Por tanto, un equilibrio de Nash debe resolver el sistema

$$x_1 = \frac{d - x_2 - c}{2} \quad \text{y} \quad x_2 = \frac{d - x_1 - c}{2},$$

que tiene por única solución $\bar{x}_1 = \bar{x}_2 = (d-c)/3$. La producción total viene dada por $2(d-c)/3$, que es mayor que la obtenida en el caso del monopolio, $(d-c)/2$. El beneficio de cada empresa viene dado por $(d-c)^2/9$, con lo que la suma de los beneficios de las empresas, $2(d-c)^2/9$, también es menor que el beneficio del monopolista en el caso anterior, $(d-c)^2/4$. Por último, el precio del producto pasa a ser $(d+2c)/3$, menor que en el caso de un monopolista, $(d+c)/2$. Esto es consistente con el hecho de que, a mayor competencia, menores precios para los consumidores y menores beneficios para las empresas.

ESCENARIO III: Duopolio de Stackelberg.

Para concluir, discutimos el escenario en el que la elección de producción es secuencial. Supondremos que la empresa 1 actúa como líder, eligiendo la producción en primer lugar, y que la empresa 2 es la seguidora, eligiendo su producción conociendo ya la producción de la primera empresa. Desde el punto de vista de la teoría de juegos estamos ante un juego en forma extensiva, en el que el concepto de solución apropiado sería el llamado *equilibrio perfecto en subjuegos* (Selten, 1975). La estrategia del jugador 1 consiste nuevamente en elegir $x_1 \in [0, d]$ y la estrategia del jugador 2 será una función $g_d^2 : [0, d] \to [0, d]$ que, para cada $x_1 \in [0, d]$, elige un nivel de producción $x_2 \in [0, d]$. Sus funciones de beneficio serán las mismas del apartado

anterior, f_d^1 y f_d^2. De modo similar al equilibrio de Nash, un equilibrio perfecto en subjuegos exige que ambos jugadores estén maximizando dada la estrategia del rival. Formalmente,

$$\begin{array}{rcl} f_d^1(\bar{x}_1, g_d^2(\bar{x}_1)) & \geq & f_d^1(x_1, g_d^2(x_1)) \text{ para todo } x_1 \in [0,d] \quad \text{y} \\ f_d^2(\bar{x}_1, g_d^2(\bar{x}_1)) & \geq & f_d^2(\bar{x}_1, x_2) \text{ para todo } x_2 \in [0,d]. \end{array}$$

En este sencillo ejemplo podemos caracterizar explícitamente la función g_d^2. En equilibrio, $\bar{g}_d^2(x_1)$ será lo mejor que el jugador 2 puede hacer tomando como dada la producción del jugador 1. Por tanto, podemos usar la función mejor respuesta del apartado anterior: $\bar{g}_d^2(x_1) = \text{BR}_2(x_1) = (d - x_1 - c)/2$. Por tanto, el problema al que se enfrenta el jugador 1 es el de maximizar, con $x_1 \in [0,d]$, la función

$$\bar{f}_d(x_1) = f_d^1(x_1, (d - x_1 - c)/2) = (d - x_1 - (d - x_1 - c)/2)x_1 - cx_1 = (\frac{d - x_1 - c}{2})x_1.$$

Como $\bar{f}_d''(x_1) = -1$ para todo $x_1 \in [0,d]$, tenemos nuevamente una función cóncava y $\bar{f}_d'(x_1) = 0$ es una condición suficiente de optimalidad global. En este caso tenemos

$$\bar{f}_d'(x_1) = \frac{d - x_1 - c}{2} - \frac{x_1}{2} = \frac{d - c}{2} - x_1,$$

con lo que la condición $\bar{f}_d'(x_1) = 0$ implica que el óptimo es $\bar{x}_1 = (d - c)/2$. Por tanto, la producción óptima de la empresa líder en el duopolio de Stackelberg coincide con la producción en caso de monopolio. Sin embargo, aquí tenemos la producción adicional de la empresa seguidora, $\bar{x}_2 = \bar{g}_d^2(\bar{x}_1) = (d - c)/4$. La producción total es $3(d - c)/4$. El beneficio de la empresa líder es $(d - c)^2/8$, el doble que el beneficio de la seguidora, $(d - c)^2/16$, para un beneficio total de $3(d-c)^2/16$. El precio en este caso es $(d+3c)/4$. Nuevamente, podemos ver que la competencia aumenta la producción, reduce el beneficio de las empresas y también el precio a pagar por el consumidor.

En la Tabla 5.2 presentamos el resumen de los resultados en los tres escenarios discutidos, incluyendo una ilustración numérica para el caso en el que $d = 15$ y $c = 3$.

El poder obtener una expresión explícita para la función g_d^2 ha sido fundamental para poder obtener la caracterización analítica de la solución del duopolio de Stackelberg. En general, cuando nos enfrentamos a un problema binivel, no es posible tener dicha expresión explícita de la solución del problema del nivel inferior como función de la elección del nivel superior, lo que complica sustancialmente la resolución de estos problemas. A continuación vamos a presentar un procedimiento alternativo para resolver el duopolio de Stackelberg que tiene la gran ventaja de servir de modo mucho más general, sin requerir una expresión explícita para las soluciones del nivel inferior. A continuación presentamos la formulación matemática del problema binivel asociado al duopolio de Stackelberg, en la que usamos el hecho de que ningún jugador elegirá un nivel de producción que lleve el nivel total por encima de la demanda máxima d:

$$\begin{array}{rl} \text{maximizar}_{x_1,x_2} & (d - x_1 - x_2)x_1 - cx_1 \\ \text{sujeto a} & x_1 \leq d \\ & x_1 \geq 0 \\ & x_2 \in \text{argmax}_{x_2}(d - x_1 - x_2)x_2 - cx_2 \\ & \quad \text{sujeto a} \quad x_1 + x_2 \leq d \\ & \qquad\qquad\quad x_2 \geq 0. \end{array}$$

Resumen general de resultados							
	Prod. $\bar{x}_1$	Prod. $\bar{x}_2$	Prod. Total	Benef. 1	Benef. 2	Benef. Total	Precio
Monopolio			$\frac{d-c}{2}$			$\frac{(d-c)^2}{4}$	$\frac{d+c}{2}$
Duop. Cournot	$\frac{d-c}{3}$	$\frac{d-c}{3}$	$\frac{2(d-c)}{3}$	$\frac{(d-c)^2}{9}$	$\frac{(d-c)^2}{9}$	$\frac{2(d-c)^2}{9}$	$\frac{d+2c}{3}$
Duop. Stackelberg	$\frac{d-c}{2}$	$\frac{d-c}{4}$	$\frac{3(d-c)}{4}$	$\frac{(d-c)^2}{8}$	$\frac{(d-c)^2}{16}$	$\frac{3(d-c)^2}{16}$	$\frac{d+3c}{4}$

Ilustración para el caso $d = 15$ y $c = 3$							
	Prod. $\bar{x}_1$	Prod. $\bar{x}_2$	Prod. Total	Benef. 1	Benef. 2	Benef. Total	Precio
Monopolio			6			36	9
Duop. Cournot	4	4	8	16	16	32	7
Duop. Stackelberg	6	3	9	18	9	27	6

Tabla 5.2: Resumen de los resultados de los distintos modelos de competencia.

Resolver este problema es equivalente a encontrar la producción asociada al equilibrio perfecto en subjuegos que acabamos de caracterizar. El jugador 1 quiere maximizar sus beneficios, sujeto a que el jugador 2, a la vista de la producción elegida por el jugador 1, también lo esté haciendo. Para mostrar el planteamiento general para resolver este tipo de problemas binivel, presentamos una reformulación equivalente como problema de minimización y con todas las restricciones del tipo "menor o igual", para tener la formulación habitual con la que hemos trabajado en este tema:

$$\begin{array}{rl} \text{minimizar}_{x_1,x_2} & x_1^2 - (d-c-x_2)x_1 \\ \text{sujeto a} & x_1 \leq d \\ & -x_1 \leq 0 \\ & x_2 \in \text{argmin}_{x_2}\, x_2^2 - (d-c-x_1)x_2 \\ & \quad \text{sujeto a} \quad x_1 + x_2 \leq d \\ & \qquad\qquad\quad\; - x_2 \leq 0. \end{array} \tag{5.7}$$

Si aislamos el problema del nivel inferior tenemos

$$\begin{array}{rl} \text{minimizar}_{x_2} & f(x_2) = x_2^2 - (d-c-x_1)x_2 \\ \text{sujeto a} & g_1(x_2) = x_1 + x_2 - d \leq 0 \\ & g_2(x_2) = -x_2 \leq 0, \end{array}$$

donde d, c y también x_1 son parámetros fijos. La función objetivo es convexa y la región factible también. Por tanto, al tener un problema de programación convexa, las condiciones de KKT (Teorema 5.6) son necesarias y suficientes para garantizar la optimalidad. Para presentar dichas condiciones necesitamos conocer primero los gradientes de función objetivo y restricciones:

$$\nabla f(x_2) = 2x_2 - d + c + x_1, \quad \nabla g_1(x_2) = 1 \quad \text{y} \quad \nabla g_1(x_2) = -1.$$

Ahora las condiciones de KKT, Teorema 5.6, unidas a la convexidad del problema en estudio, nos dicen que una condición necesaria y suficiente para que un punto factible $\bar{x}_2$ en el que los gradientes de las restricciones activas son independientes sea un mínimo es que existan multiplicadores u_1 y u_2 tales que

$$\begin{aligned}&\nabla f(x_2) + u_1 \nabla g_1(x_2) + u_2 \nabla g_1(x_2) = 0\\ &u_1 g_1(x_2) = 0, \quad u_2 g_2(x_2) = 0\\ &u_1 \geq 0, \quad u_2 \geq 0,\end{aligned}$$

y, sustituyendo, tenemos

$$\begin{aligned}&2x_2 - d + c + x_1 + u_1 - u_2 = 0,\\ &u_1(x_1 + x_2 - d) = 0, \quad -u_2 x_2 = 0\\ &u_1 \geq 0, \quad u_2 \geq 0.\end{aligned} \tag{5.8}$$

Es importante observar que los gradientes de las restricciones son paralelos entre sí, lo que podría ser un problema para cumplir la condición de regularidad del Teorema 5.6. Sin embargo, esto únicamente sería un problema si las dos restricciones estuviesen simultáneamente activas en el punto en estudio, cosa que únicamente podría suceder si $x_1 = d$ y $x_2 = 0$.

Ahora podemos reemplazar el problema inferior en el problema binivel (5.7) por sus restricciones de factibilidad más las restricciones de optimalidad de la Ecuación (5.8), obteniendo la siguiente formulación en un único nivel:

$$\begin{aligned}\text{minimizar}_{x_1,x_2,u_1,u_2} \quad & x_1^2 - (d - c - x_2)x_1\\ \text{sujeto a} \quad & x_1 \leq d\\ & -x_1 \leq 0\\ & x_1 + x_2 - d \leq 0\\ & -x_2 \leq 0\\ & 2x_2 - d + c + x_1 + u_1 - u_2 = 0\\ & u_1(x_1 + x_2 - d) = 0\\ & -u_2 x_2 = 0\\ & u_1 \geq 0\\ & u_2 \geq 0.\end{aligned} \tag{5.9}$$

Este problema captura, mediante las últimas cinco restricciones, la condición de optimalidad del problema del nivel inferior del problema binivel original y, mediante las dos anteriores, las condiciones de factibilidad de dicho problema, con lo que su resolución es equivalente a la resolución del problema original. Por tanto, la resolución de este problema en cualquier optimizador global para distintos valores de d y c, a través de AMPL por ejemplo, nos permitiría llegar a los mismos óptimos que con las soluciones analíticas obtenidas anteriormente. Como ya hemos comentado, la ventaja de este método radica en que es un enfoque mucho más general, aplicable más allá de pequeños problemas académicos como el que hemos estado discutiendo en este apartado.

5.5.3 Clasificación mediante *support vector machines*

En este apartado vamos a ver un pequeño ejemplo de cómo la optimización matemática y, en particular, la dualidad lagrangiana, aparece de modo natural en un campo tan activo en

la actualidad como el del aprendizaje automático (*machine learning*). Dado que el objetivo es únicamente mostrar una aplicación más de los conceptos teóricos desarrollados en este tema, solamente introduciremos aquellos conceptos propios del aprendizaje automático que sean necesarios para hacer la exposición autocontenida.[10]

Uno de los problemas más habituales en el campo del aprendizaje automático es el problema de *clasificación*. Se trata de un problema en el que se dispone de un conjunto predeterminado de categorías o clases y de un conjunto de observaciones de entrenamiento para cuyos elementos se conoce la clase a la que pertenecen. El objetivo es ser capaces de clasificar nuevas observaciones, es decir, determinar a qué clase pertenecen a partir del conocimiento aprendido con la muestra de entrenamiento. La clasificación se considera un tipo de aprendizaje *supervisado*, pues conocemos de antemano las clases disponibles y, para cada observación del conjunto de entrenamiento, la clase a la que pertenece. Por contraposición, la versión no supervisada es lo que se conoce habitualmente como agrupamiento (*clustering*).

A la hora de ver cómo resolver un problema de clasificación, vamos a centrarnos en una de las técnicas de resolución más exitosas: las *máquinas de vectores de soporte*, conocidas habitualmente por sus siglas en inglés, SVM (*support vector machines*). Los trabajos pioneros en esta técnica son Vapnik y Lerner (1963), Vapnik y Chervonenkis (1964, 1974) y, algo más recientemente, destacan los avances introducidos en Cortes y Vapnik (1995). De una u otra manera, la investigación actual en SVMs se apoya en los resultados e intuiciones desarrolladas en estos primeros trabajos.

Problemas de clasificación binaria

Durante toda la exposición nos centraremos en problemas de clasificación binaria, en los que tenemos únicamente dos categorías en las que clasificar nuestras observaciones. Ejemplos de estos problemas pueden ser la clasificación de pacientes en enfermos y sanos, de correos electrónicos en basura y no basura, clasificación en válido o inválido en pruebas de calidad.

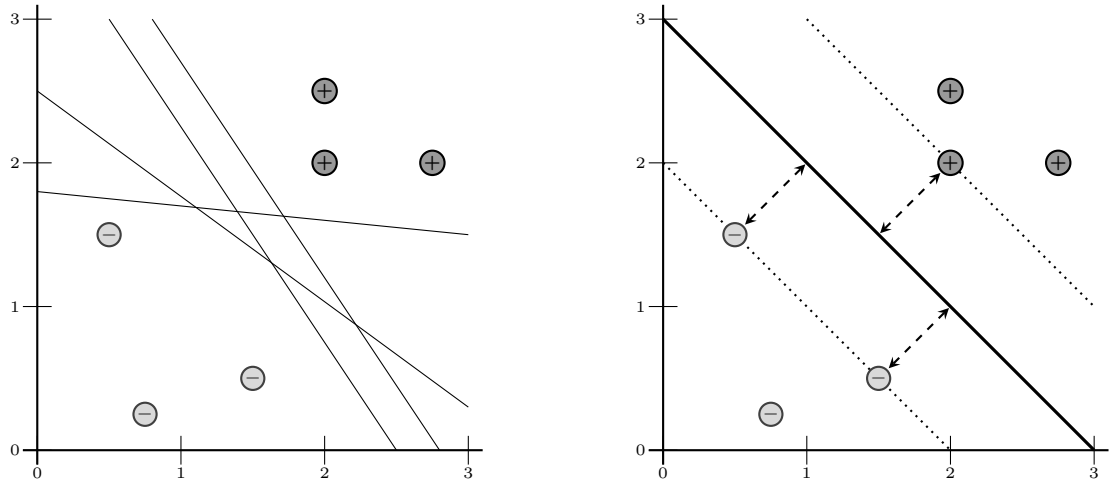
Partimos de un conjunto de l observaciones, para cada una de las cuales tenemos los valores de n variables explicativas. Por tanto, las observaciones vienen dadas por $\{\boldsymbol{x}^1, \dots, \boldsymbol{x}^l\} \subset \mathbb{R}^n$. Además, para cada observación $i \in \{1, \dots, l\}$ tenemos su clasificación $y_i \in \{-1, 1\}$. El objetivo es usar la información disponible en este conjunto de datos de entrenamiento para clasificar una nueva observación $\boldsymbol{x}^0 \in \mathbb{R}^n$.

Resolución mediante *support vector machines*

En este apartado trabajaremos bajo el supuesto de que los datos son linealmente separables, es decir, que existe un hiperplano en $\mathbb{R}^n$ capaz de dejar a un lado del mismo las observaciones con etiqueta positiva y al otro las observaciones con etiqueta negativa. La Figura 5.12(a) muestra un ejemplo de conjunto de entrenamiento separable y varios hiperplanos separadores. Aunque el supuesto de que los datos son linealmente separables es muy restrictivo, da lugar a problemas en los que se puede ilustrar más limpiamente los fundamentos de los SVMs. Al final de este

[10] La inclusión de este apartado ha sido motivada por el Trabajo de Fin de Grado Dono-Caramés (2019), y la exposición se apoya parcialmente en el mismo.

apartado comentaremos brevemente cómo las técnicas aquí presentadas se pueden usar con datos no separables.



(a) Conjunto de datos linealmente separables.

(b) Hiperplano separador de máximo margen.

Figura 5.12: Ejemplo de clasificación con SVMs.

Recordemos que, en general, un *hiperplano* en $\mathbb{R}^n$ se define mediante un vector no nulo $\boldsymbol{w} \in \mathbb{R}^n$ y un escalar $b \in \mathbb{R}$ como el conjunto $H(\boldsymbol{w}, r) = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{w}^\intercal \boldsymbol{x} = r\} = \{\boldsymbol{x} \in \mathbb{R}^n : \sum_{i=1}^n w_i x_i = r\}$. Por tanto, en el caso de datos linealmente separables podremos elegir un hiperplano $H(\boldsymbol{w}, r)$ tal que, dado $i \in \{1, \ldots, l\}$:

- Si $y_i = -1$, $\boldsymbol{w}^\intercal \boldsymbol{x}^i - r < 0$.
- Si $y_i = 1$, $\boldsymbol{w}^\intercal \boldsymbol{x}^i - r > 0$.

Además, dada una nueva observación $\boldsymbol{x}^0$, esta será clasificada como $y_0 = -1$ si $\boldsymbol{w}^\intercal \boldsymbol{x}^0 - r < 0$ y como $y_0 = 1$ si $\boldsymbol{w}^\intercal \boldsymbol{x}^0 - r > 0$. En el caso de que $\boldsymbol{w}^\intercal \boldsymbol{x}^0 = r$ diremos que esta nueva observación no es clasificable por $H(\boldsymbol{w}, r)$. Nótese que esta técnica de clasificación no permite únicamente asignar nuevas observaciones a clases según el signo de $\boldsymbol{w}^\intercal \boldsymbol{x}^0 - r$, sino que la magnitud de este valor también nos da una forma de cuantificar la confianza que tenemos en una clasificación dada. De hecho, el error de un clasificador lineal a la hora de clasificar nuevos datos puede ser acotado en función del margen que deja el hiperplano con respecto a los datos de entrenamiento.[11] Debido a esto, es natural tratar de encontrar el hiperplano separador de máximo margen, ilustrado en la Figura 5.12(b) y en cuyo cálculo nos centraremos a continuación.

Por comodidad, dado un hiperplano $H(\boldsymbol{w}, r)$ trabajaremos con $b = -r$, con lo que la regla de clasificación nos queda como $y_0 = -1$ si $\boldsymbol{w}^\intercal \boldsymbol{x}^0 + b < 0$ y como $y_0 = 1$ si $\boldsymbol{w}^\intercal \boldsymbol{x}^0 + b > 0$. Además, como estamos suponiendo que nuestros datos de entrenamiento son linealmente separables tendremos que, para todo $i \in \{1, \ldots, l\}$, $\boldsymbol{w}^\intercal \boldsymbol{x}^i + b < 0$ si $y_i = -1$ y $\boldsymbol{w}^\intercal \boldsymbol{x}^i + b > 0$ si $y_i = 1$.

[11] Un análisis riguroso al respecto puede verse en Cristianini y Shawe-Taylor (2000).

Como tenemos una cantidad finita de observaciones, podemos definir

$$\beta = \min_{i\in\{1,\dots,l\}} |\boldsymbol{w}^\intercal \boldsymbol{x}^i + b|,$$

que en cierta manera representa la distancia al hiperplano de la observación más cercana al mismo. El supuesto de separabilidad de los datos de entrenamiento nos asegura que $\beta \neq 0$. Ahora podemos "normalizar" nuestro hiperplano separador dividiendo por β, pues $H(\boldsymbol{w}, r)$ y $H(\boldsymbol{w}/\beta, r/\beta)$ representan al mismo hiperplano de $\mathbb{R}^n$. Supongamos entonces, sin pérdida de generalidad, que estamos trabajando ya con el hiperplano normalizado, con lo que sabemos que

$$\begin{cases} \boldsymbol{w}^\intercal \boldsymbol{x}^i + b \leq -1 & \text{si } y_i = -1 \\ \boldsymbol{w}^\intercal \boldsymbol{x}^i + b \geq 1 & \text{si } y_i = 1. \end{cases}$$

Estas condiciones se pueden expresar a través de las ecuaciones

$$y_i(\boldsymbol{w}^\intercal \boldsymbol{x}^i + b) - 1 \geq 0, \quad \text{para todo } i \in \{1,\dots,l\}, \tag{5.10}$$

que en nuestro problema de optimización serán las restricciones que aseguren que elegimos un hiperplano que clasifica correctamente (separa estrictamente) las observaciones del conjunto de entrenamiento. El siguiente paso es determinar la función objetivo que, como hemos dicho, consistirá en elegir el hiperplano de máximo margen. Veamos antes de nada cómo calcular dicho margen. Sea $\boldsymbol{x}^k$ una observación en la que se alcanza $\min_{i\in\{1,\dots,l\}} |\boldsymbol{w}^\intercal \boldsymbol{x}^i + b|$. Supongamos además, sin pérdida de generalidad que $\boldsymbol{x}^k$ es una observación positiva y, por tanto, al estar ya asumiendo que el hiperplano está normalizado tenemos que $\boldsymbol{w}^\intercal \boldsymbol{x}^k + b = 1$. Para calcular la distancia de $\boldsymbol{x}^k$ al hiperplano basta calcular la distancia de $\boldsymbol{x}^k$ a $\boldsymbol{x}^{\mathrm{p}} = \mathrm{proy}(\boldsymbol{x}^k, H(\boldsymbol{w}, -b))$, la proyección de $\boldsymbol{x}^k$ sobre el hiperplano. Ahora bien, calcular $\boldsymbol{x}^{\mathrm{p}}$ es equivalente a encontrar λ tal que $\boldsymbol{w}^\intercal(\boldsymbol{x}^k + \lambda \boldsymbol{w}) + b = 0$:

$$\boldsymbol{w}^\intercal(\boldsymbol{x}^k + \lambda \boldsymbol{w}) + b = 0 \;\Leftrightarrow\; \boldsymbol{w}^\intercal \boldsymbol{x}^k + b + \lambda \|\boldsymbol{w}\|^2 = 0 \;\Leftrightarrow\; 1 + \lambda \|\boldsymbol{w}\|^2 = 0 \;\Leftrightarrow\; \lambda = \frac{-1}{\|\boldsymbol{w}\|^2}.$$

Entonces, $\boldsymbol{x}^{\mathrm{p}} = x^k - \boldsymbol{w}/\|\boldsymbol{w}\|^2$ y la distancia entre $\boldsymbol{x}^k$ y el hiperplano se calcula como

$$\|\boldsymbol{x}^k - \boldsymbol{x}^{\mathrm{p}}\| = \|-\boldsymbol{w}/\|\boldsymbol{w}\|^2\| = 1/\|\boldsymbol{w}\|.$$

Entonces $1/\|\boldsymbol{w}\|$ se corresponde con la distancia entre el hiperplano y la observación más próxima al mismo, que es lo que querremos maximizar. Tenemos ya todos los ingredientes necesarios para definir el hiperplano separador de máximo margen. Debemos encontrar valores para $\boldsymbol{w}$ y b de tal manera que se cumplan las restricciones dadas por (5.10) y se maximice $1/\|\boldsymbol{w}\|$. Equivalentemente, podremos minimizar $\|\boldsymbol{w}\|$ o, incluso más fácil, $\|\boldsymbol{w}\|^2 = \boldsymbol{w}^\intercal \boldsymbol{w} = \sum_{i=1}^n w_i^2$. Para facilitar desarrollos posteriores será conveniente dividir por dos esta función objetivo, resultando en el problema de optimización

Clasificación con SVMs. Problema primal.

$$\begin{aligned} &\text{minimizar}_{\boldsymbol{w},b} && \tfrac{1}{2}\boldsymbol{w}^\intercal \boldsymbol{w} \\ &\quad \text{sujeto a} && y_i(\boldsymbol{w}^\intercal \boldsymbol{x}^i + b) - 1 \geq 0 \qquad i = 1,\dots,l. \end{aligned} \tag{5.11}$$

La función objetivo se puede como la función cuadrática $\frac{1}{2}\boldsymbol{w}^\top I_{n\times n}\boldsymbol{w}$, donde la matriz $I_{n\times n}$ es definida positiva. Dado que las restricciones de este problema son lineales, estamos ante un problema de optimización convexa, lo que sabemos facilitará su resolución. Además, como la función objetivo es estrictamente convexa, el Teorema 2.2 nos asegura que el óptimo global es único. Sin embargo, estos problemas de clasificación surgen habitualmente en contextos *big data* con un número muy elevado de observaciones en el conjunto de entrenamiento, lo que se traducirá en un número muy elevado de restricciones. Esto puede dar lugar a problemas difíciles de resolver, a pesar de la convexidad.

A continuación vamos a presentar el dual del problema de optimización (5.11). La convexidad de (5.11) nos asegura que se cumplirá el Teorema de dualidad fuerte (Teorema 5.19), con lo que será equivalente resolver los problemas primal y dual. En este caso, la utilidad del problema dual no es tanto que ayude a resolver el problema más rápido, sino que permite comprender mejor ciertas propiedades del hiperplano de máximo margen. Además, la mayoría de las extensiones de la clasificación con SVMs se definen de modo más natural sobre el problema dual.

La función lagrangiana, cambiando previamente las restricciones a $-y_i(\boldsymbol{w}^\top\boldsymbol{x}^i + b) + 1 \leq 0$, viene dada por

$$L((\boldsymbol{w}, b), \boldsymbol{u}) = \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} - \sum_{i=1}^{l} u_i\big(y_i(\boldsymbol{w}^\top\boldsymbol{x}^i + b) + 1\big)$$

y la función dual asociada es $\mathcal{L}^D(\boldsymbol{u}) = \text{ínf}_{\boldsymbol{w},b}\, L((\boldsymbol{w}, b), \boldsymbol{u})$. Entonces, el problema dual de (5.11) se corresponde con $\text{maximizar}_{\boldsymbol{u}\geq\boldsymbol{0}}\, \mathcal{L}^D(\boldsymbol{u})$. A continuación mostraremos que las soluciones del problema $\mathcal{L}^D(\boldsymbol{u})$ se pueden caracterizar analíticamente, lo que facilitará la formulación y resolución del dual. Para cada vector $\boldsymbol{u}$ de multiplicadores, $\mathcal{L}^D(\boldsymbol{u})$ también es un problema convexo y sus condiciones de optimalidad son,

$$\frac{\partial\mathcal{L}^D(\boldsymbol{u})}{\partial b} = 0 \;\Leftrightarrow \sum_{i=1}^{l} u_i y_i = 0 \tag{5.12}$$

y, para cada $j \in \{1, \ldots, n\}$,

$$\frac{\partial\mathcal{L}^D(\boldsymbol{u})}{\partial w_j} = 0 \;\Leftrightarrow w_j - \sum_{i=1}^{l} u_i y_i x_j^i = 0, \quad \text{con lo que} \quad \boldsymbol{w} = \sum_{i=1}^{l} u_i y_i \boldsymbol{x}^i. \tag{5.13}$$

Las ecuaciones (5.12) y (5.13), resultantes de las condiciones de optimalidad de la función dual, nos ayudan a entender mejor el problema que estamos resolviendo y ciertas propiedades de sus soluciones:

- La expresión $\boldsymbol{w} = \sum_{i=1}^{l} u_i y_i \boldsymbol{x}^i$ nos dice que el vector normal al hiperplano de máximo margen se corresponde con una combinación lineal de los datos de entrenamiento.
- También sabemos que los únicos coeficientes u_i que serán distintos de cero serán aquellos cuyas restricciones se saturen en el óptimo. Estas observaciones son las que se llaman "vectores de soporte" y dan nombre al método. En el problema primal (5.11) las restricciones saturadas son justamente aquellas que cuyas observaciones asociadas están a

distancia mínima del hiperplano. Por tanto, el hiperplano de máximo margen no se verá modificado por pequeños cambios en observaciones que no estén a distancia mínima de él (cambios en vectores que no son de soporte).

- A la vista de $\boldsymbol{w} = \sum_{i=1}^{l} u_i y_i \boldsymbol{x}^i$, si entendemos cada u_i como una medida de la importancia de la observación $\boldsymbol{x}^i$ en determinar $\boldsymbol{w}$, entonces la condición $\sum_{i=1}^{l} u_i y_i = 0$ nos dice que el peso total de las observaciones positivas y negativas es el mismo.
- Estamos también en condiciones de entender la razón del nombre "vectores de soporte". Si miramos la Figura 5.12(b), vemos que hay tres vectores de soporte, dos correspondientes a observaciones negativas y uno correspondiente a una observación positiva. Supongamos ahora que desde cada una de ellas aplicamos una fuerza de intensidad u_i y dirección $y_i \boldsymbol{w}/\|\boldsymbol{w}\|$. La condición $\sum_{i=1}^{l} u_i y_i = 0$ nos dice entonces que las fuerzas aplicadas sobre el hiperplano se compensan, este no se desplazaría en la dirección $\boldsymbol{w}/\|\boldsymbol{w}\|$ ni hacia las observaciones positivas ni hacia las negativas. En cierta manera, estas tres observaciones están "soportando" al hiperplano en equilibrio. De hecho, se puede demostrar además que el momento de las fuerzas aplicadas es cero, con lo que no habría ningún efecto de rotación en el hiperplano.
- La clasificación de una nueva observación x^0 depende del signo de $\boldsymbol{w}^\intercal \boldsymbol{x}^0 + b$, que es lo mismo que $\sum_{i=1}^{l} (u_i y_i (\boldsymbol{x}^i)^\intercal \boldsymbol{x}^0) + b$. Si pensamos en $(\boldsymbol{x}^i)^\intercal \boldsymbol{x}^0$ como una medida de la "semejanza" o "afinidad de direcciones" entre $\boldsymbol{x}^i$ y $\boldsymbol{x}^0$ (cuanto más ortogonales más próximo a cero estará este valor), vemos que a la hora de clasificar $\boldsymbol{x}^0$ tendrán más peso las observaciones más "afines".

Reescribamos ahora la función $\mathcal{L}^D(\boldsymbol{u})$ usando las ecuaciones (5.12) y (5.13):

$$\begin{aligned}
\mathcal{L}^D(\boldsymbol{u}) &= \inf_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^\intercal \boldsymbol{w} - \sum_{i=1}^{l} u_i (y_i(\boldsymbol{w}^\intercal \boldsymbol{x}^i + b) + 1) \\
&= \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} u_i u_j y_i y_j (\boldsymbol{x}^i)^\intercal \boldsymbol{x}^j - \sum_{i=1}^{l}\sum_{j=1}^{l} u_i u_j y_i y_j (\boldsymbol{x}^i)^\intercal \boldsymbol{x}^j - \sum_{i=1}^{l} u_i y_i b + \sum_{i=1}^{l} u_i \\
&= \sum_{i=1}^{l} u_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} u_i u_j y_i y_j (\boldsymbol{x}^i)^\intercal \boldsymbol{x}^j.
\end{aligned}$$

Por último, apoyándonos en esta formulación explícita de las soluciones del subproblema dual y sin olvidarnos de la Ecuación (5.12), tenemos que el problema dual del problema de clasificación con SVMs se puede formular como

Clasificación con SVMs. Problema dual.

$$\begin{aligned}
\text{maximizar}_{\boldsymbol{u}} \quad & \mathcal{L}^D(\boldsymbol{u}) = \sum_{i=1}^{l} u_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} u_i u_j y_i y_j (\boldsymbol{x}^i)^\intercal \boldsymbol{x}^j \\
\text{sujeto a} \quad & \sum_{i=1}^{l} u_i y_i = 0 \\
& u_i \geq 0 \qquad i = 1, \ldots, l.
\end{aligned} \tag{5.14}$$

Sabemos que el problema de optimización dual siempre es un problema de maximización cóncava, con lo que todo máximo local será también global. Dada una solución óptima $\bar{\boldsymbol{u}}$ del dual, podemos recuperar la solución del primal sin más que calcular $\boldsymbol{w} = \sum_{i=1}^{l} \bar{u}_i y_i \boldsymbol{x}^i$ y obtener b despejando en la restricción asociada a alguna observación con $\bar{u}_i \neq 0$ (pues sabemos que la restricción asociada se satura).

A la hora de elegir entre resolver el primal y el dual que hemos formulado, hay que tener en cuenta que, si bien el dual tiene esencialmente una única restricción, el número de variables es ahora mucho más grande: una por cada observación del conjunto de entrenamiento. Además, la propia formulación del problema se complica, pues definir la función objetivo pasa por hacer un triple bucle que tendrá $l \times l \times n$ sumandos.[12] Además, teniendo en cuenta que el gradiente de $\mathcal{L}^D(\boldsymbol{u})$ tiene también l componentes y que la mayoría de los optimizadores punteros se apoyan de una u otra manera en el gradiente, su cálculo iteración a iteración puede resultar computacionalmente muy costoso. Esta dificultad que acabamos de comentar aparece de modo recurrente en los campos del *big data* y *machine learning*. Es por esto que se han diseñado algoritmos específicos, que intentan aprovechar la estructura de estos problemas. Especialmente populares son los algoritmos de gradiente estocástico, que en cada iteración usan únicamente un subconjunto reducido del conjunto de entrenamiento; en nuestro problema esto supondría calcular únicamente un subconjunto reducido de las componentes del gradiente. Este subconjunto reducido se elige aleatoriamente y cambia iteración a iteración. Esto permite reducir significativamente el coste computacional en cada iteración y se puede demostrar que los métodos resultantes mantienen buenas propiedades de convergencia.

Extensiones y generalizaciones del SVM básico

La clasificación mediante SVMs que acabamos de presentar presenta principalmente dos grandes limitaciones que comentamos a continuación:

Problemas no separables. El método tal y como lo hemos descrito no es aplicable a problemas donde los datos de entrenamiento no sean linealmente separables.

Sensibilidad excesiva. El método es muy sensible a las observaciones dadas por los vectores de soporte. Cambios en una de estas observaciones, incluso manteniendo fijo el resto del conjunto de entrenamiento, pueden dar lugar a hiperplanos totalmente distintos. Dado que las observaciones suelen estar sujetas a ruido estadístico, esta falta de estabilidad puede limitar la fiabilidad del método a la hora de clasificar nuevas observaciones.

Un método habitual para trabajar con problemas que no son linealmente separables es pasar las observaciones de entrenamiento a un espacio de dimensión superior. Ilustraremos la idea con un sencillo ejemplo. Considérese el conjunto de observaciones de entrenamiento de la Figura 5.13, con las observaciones positivas en los cuadrantes primero y tercero y las observaciones negativas en los cuadrantes segundo y cuarto. Claramente, no hay ningún hiperplano que separe las observaciones negativas de las positivas. Sin embargo, es muy fácil dar una regla de clasificación para este conjunto: el signo de una observación coincidirá con el signo del

[12] En el código AMPL complementario a este apartado puede verse que la resolución directa del dual mediante optimizadores generales es mucho más lenta que la del primal.

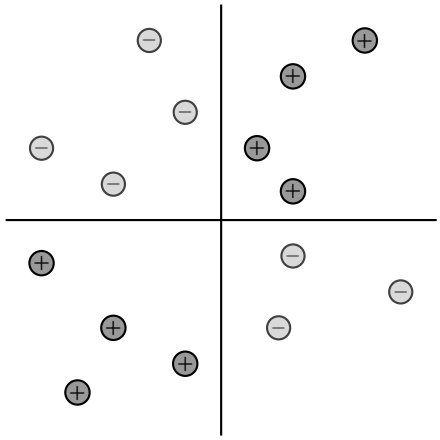


Figura 5.13: Ejemplo de problema no separable linealmente.

producto de sus dos coordenadas. Esta se una regla de clasificación "no lineal", pero que puede ser obtenida también trabajando con SVMs en dimensión superior. En este caso, bastaría con pasar cada observación $\boldsymbol{x}^i = (\boldsymbol{x}^i_1, \boldsymbol{x}^i_2) \in \mathbb{R}^2$ a $\bar{\boldsymbol{x}}^i = (\boldsymbol{x}^i_1, \boldsymbol{x}^i_2, \boldsymbol{x}^i_1 \cdot \boldsymbol{x}^i_2) \in \mathbb{R}^3$. Así definido, con las observaciones en $\mathbb{R}^3$, ya podríamos calcular el hiperplano separador de máximo margen y obtendríamos una regla similar a la regla no lineal anterior.

Es importante destacar que transformaciones como la que acabamos de describir permiten utilizar separación mediante hiperplanos en el espacio transformado para obtener reglas de clasificación no lineales en el espacio original. Para dar más generalidad a los SVMs también se puede reemplazar el producto escalar $(\boldsymbol{x}^i)^\intercal \boldsymbol{x}^j$ que aparece en (5.14) por otro producto escalar convenientemente elegido en el espacio de dimensión superior. En general, esto se lleva a cabo mediante las llamadas *funciones kernel* que "miden" la afinidad de dos observaciones y normalmente se denotan por $k(\boldsymbol{x}^i, \boldsymbol{x}^j)$. Estas funciones se suelen elegir de tal manera que, al reemplazar $(\boldsymbol{x}^i)^\intercal \boldsymbol{x}^j$ por $k(\boldsymbol{x}^i, \boldsymbol{x}^j)$ en el problema (5.14), se siga teniendo un problema de maximización sobre una función cóncava. El enfoque que acabamos de describir se aplicó por primera vez en Boser y otros (1992).

El enfoque que acabamos de comentar es muy general pues, incrementando suficientemente la dimensión del espacio de características de las observaciones de entrenamiento, cualquier problema se puede convertir en un problema linealmente separable en el espacio aumentado. Sin embargo, al hacer esto se corre el riesgo de caer en uno de los problemas más habituales de la estimación estadística: el sobreajuste. Obtendríamos un clasificador muy a medida para las observaciones del conjunto de entrenamiento y que posiblemente no se comportaría de manera adecuada a la hora de clasificar nuevas observaciones. Por tanto, es importante encontrar un buen equilibrio entre la capacidad de separación del SVM resultante y su poder predictivo ante nuevas observaciones.

A continuación presentamos la idea de una de las formas más habituales de conseguir el equilibrio anterior y poder implementar SVMs sobre problemas no separables linealmente que, además, controle los problemas de sensibilidad de los métodos ante cambios en las observaciones asociadas a los vectores de soporte. La idea, discutida originalmente en Cortes y Vapnik (1995) y Vapnik (1995), consiste en calcular hiperplanos de margen "suave" (*soft margin hyperplanes*).

Formalmente, se relajan las restricciones que imponen la correcta clasificación de todas las observaciones del conjunto de entrenamiento añadiendo variables de holgura:

Clasif. "dura". Prob. primal.

$$\begin{array}{ll}\text{minimizar}_{\boldsymbol{w},b} & \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} \\ \text{sujeto a} & y_i(\boldsymbol{w}^\top\boldsymbol{x}^i + b) - 1 \geq 0 \quad \forall i\end{array}$$

Clasif. "suave". Prob. primal.

$$\begin{array}{ll}\text{minimizar}_{\boldsymbol{w},b,\boldsymbol{s}} & \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + \frac{M}{l}\sum_{i=1}^{l} s_i \\ \text{sujeto a} & y_i(\boldsymbol{w}^\top\boldsymbol{x}^i + b) - 1 \geq -s_i \quad \forall i \\ & s_i \geq 0 \quad \forall i.\end{array}$$

De esta manera, cuanto mayor sea M, más se penalizarán las clasificaciones incorrectas, pero el problema tendrá soluciones factibles aunque el conjunto de entrenamiento no sea linealmente separable. Por supuesto, una de las claves del éxito de los clasificadores asociados es la elección del valor M, aunque también hay variantes de la clasificación suave más sofisticadas que la aquí presentada. La clasificación suave se puede llevar a cabo combinada con las funciones kernel descritas anteriormente. Por último, no es difícil probar que el único cambio requerido en el problema dual consiste en añadir cotas superiores $\frac{M}{l}$ a las variables duales, obteniendo la siguiente formulación:[13]

Clasificación "suave" y con kernel. Problema dual.

$$\begin{array}{ll}\text{maximizar}_{\boldsymbol{u}} & \mathcal{L}^D(\boldsymbol{u}) = \displaystyle\sum_{i=1}^{l} u_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} u_i u_j y_i y_j k(\boldsymbol{x}^i, \boldsymbol{x}^j) \\ \text{sujeto a} & \displaystyle\sum_{i=1}^{l} u_i y_i = 0 \\ & u_i \geq 0 \qquad i = 1, \ldots, l \\ & u_i \leq \frac{M}{l} \qquad i = 1, \ldots, l.\end{array}$$

5.6 Ejercicios adicionales

••Ejercicio 5.10. Considera el siguiente problema de programación lineal:

$$\begin{array}{rl}\text{maximizar} & x_1 + 3x_2 \\ \text{sujeto a} & 2x_1 + 3x_2 \leq 6 \\ & -x_1 + 4x_2 \leq 4 \\ & x_1 \geq 0 \\ & x_2 \geq 0.\end{array}$$

- Escribe las condiciones de optimalidad de KKT.
- Verifica, para cada punto extremo de la región factible, las condiciones de KKT, tanto algebraica como geométricamente.

◁

••Ejercicio 5.11. Considera el siguiente problema de optimización:

$$\begin{array}{rl}\text{minimizar} & x_1^3 + 2x_2^2 \\ \text{sujeto a} & x_1 - x_2 - 2 = 0.\end{array}$$

[13] La derivación formal del problema dual puede consultarse en Cortes y Vapnik (1995).

- Encuentra un punto que satisfaga las condiciones de KKT y verifica que, efectivamente, es una solución óptima.
- Responde a la misma pregunta si la función objetivo se reemplaza por $f(\boldsymbol{x}) = x_1^3 + x_2^3$.

◁

••**Ejercicio 5.12.** Considera el siguiente problema de optimización:

$$\begin{array}{rl} \text{minimizar} & x_1^4 + x_2^4 + 12x_1^2 + 6x_2^2 - x_1x_2 - x_1 - x_2 \\ \text{sujeto a} & x_1 + x_2 \geq 6 \\ & 2x_1 - x_2 \geq 3 \\ & x_1 \geq 0 \\ & x_2 \geq 0. \end{array}$$

Escribe las condiciones de KKT y apóyate en ellas para demostrar que el punto $\bar{\boldsymbol{x}} = (3, 3)$ es la única solución óptima. ◁

Tema 6

Optimización con restricciones. Dualidad y técnicas de descomposición

Contenidos

6.1 Introducción a las técnicas de descomposición

La resolución de problemas reales mediante herramientas de optimización matemática requiere modelizar sistemas cada vez más complejos, para los cuales se dispone cada vez de más información. Debido a esto, cada vez más frecuente encontrarse con problemas de optimización de gran tamaño, cuya resolución directa puede resultar computacionalmente prohibitiva. Por otro lado, es también frecuente que estos problemas tengan una estructura subyacente, como la separabilidad del problema de planificación energética de la Sección 5.5.1. Es en este contexto en el que cobran especial importancia las técnicas de descomposición, que buscan explotar estas estructuras para conseguir resolver el problema de partida mediante la resolución, habitualmente iterativa, de una gran cantidad de subproblemas mucho más sencillos. Por ejemplo, en el problema de planificación energética, con la ayuda de la dualidad lagrangiana, se puede resolver el problema original mediante la descomposición del problema dual en sencillos subproblemas unidimensionales que pueden ser resueltos de manera independiente.

Una ventaja importante que ofrecen este tipo de técnicas de descomposición es que la resolución de los subproblemas de cada iteración puede llevarse a cabo en paralelo, siendo idóneos para su despliegue en superordenadores o entornos de computación en la nube.

Las técnicas de descomposición en programación matemática cubren prácticamente todas las áreas de la misma: descomposición en optimización lineal y no lineal, en optimización entera, en optimización estocástica,... Un libro de referencia al respecto es Conejo y otros (2006). Para facilitar la exposición en esta sección, que no pretende ser más que una primera toma de contacto con este tipo de técnicas, nos centraremos en dos técnicas de descomposición que surgieron en el contexto de la optimización lineal y que son de gran aplicabilidad en problemas reales: el algoritmo de Dantzig-Wolfe (Dantzig y Wolfe, 1960) y el algoritmo de Benders (Benders, 1962).

Prácticamente todas las técnicas de descomposición se apoyan de una u otra manera en conceptos de dualidad, lo que hace que algunos de los algoritmos resultantes sean relativamente complejos. Es por esto que, al restringirnos a problemas lineales, las ideas se verán de manera mucho más clara, siendo también estas mismas ideas los pilares en los que se apoyan las técnicas para problemas no lineales, aunque con la dualidad clásica en programación lineal siendo habitualmente reemplazada por la dualidad lagrangiana (en la línea del ejemplo de la Sección 5.5.1).

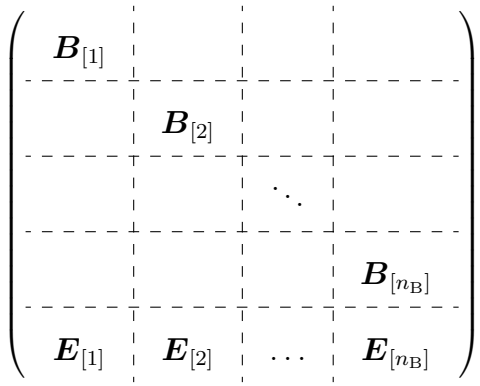
6.1.1 Restricciones y variables "complicantes"

Comenzamos presentando las dos estructuras de descomposición que aparecen más frecuentemente en la práctica:

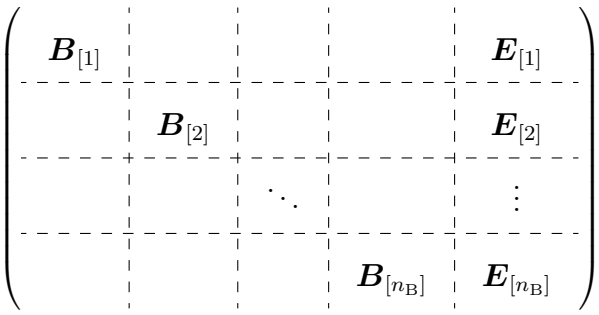
Problemas con restricciones complicantes. Se trata de problemas de programación matemática en los que una parte importante de la complejidad de resolución se debe a un subconjunto, típicamente pequeño, del conjunto de restricciones. Es para este tipo de problemas para los que es adecuado el algoritmo de Dantzig-Wolfe.

En la práctica, las situaciones más frecuentes en las que aparecen restricciones complicantes son aquellas en las que, sin la existencia de dichas restricciones, el problema original podría descomponer en subproblemas independientes. Estas restricciones también suelen

llamarse "linking constraints" o "coupling constraints", pues son restricciones que enganchan entre sí los distintos subproblemas, impidiendo resolverlos de modo independiente. La Figura 6.1(a) muestra la estructura típica de la matriz de restricciones de un problema con restricciones complicantes. En dicha figura podemos ver que si las restricciones asociadas al último bloque de filas de la matriz no estuvieran presentes, entonces, podríamos descomponer el problema original en n_{B} bloques independientes.



(a) Matriz de restricciones de un problema con restricciones complicantes.



(b) Matriz de restricciones de un problema con variables complicantes.

Figura 6.1: Estructuras matriciales susceptibles de ser atacadas con técnicas de descomposición.

Problemas con variables complicantes. Se trata de problemas de programación matemática en los que una parte importante de la complejidad de resolución se debe a un subconjunto, típicamente pequeño, del conjunto de variables. Es para este tipo de problemas para los que es adecuado el algoritmo de Benders.

De modo similar al caso de las restricciones complicantes, unas de las situaciones más frecuentes en las que aparecen variables complicantes son aquellas en las que, si pudiésemos fijar el valor de dichas variables, el problema resultante podría resolverse mucho más eficientemente, ya sea por poder descomponerse en subproblemas independientes o, por ejemplo, porque la variables complicantes son variables enteras y una vez fijadas pasamos a tener un problema de programación lineal. La Figura 6.1(b) muestra la estructura típica de la matriz de restricciones de un problema con variables complicantes que, una vez fijadas, permiten descomponer el problema resultante en subproblemas independientes. En este caso, fijando los valores de las variables asociadas al último bloque de columnas de la matriz, podríamos pasar a los lados derechos las constantes resultantes, obteniendo un problema que se puede descomponer en n_{B} subproblemas independientes.

Es interesante destacar que, a la hora de resolver problemas de programación lineal, los algoritmos de Dantzig-Wolfe y Benders son esencialmente duales entre sí. La razón es que, al ser las matrices de restricciones de un problema y su dual traspuestas entre sí, el dual de un problema con variables complicantes es un problema con restricciones complicantes y viceversa, como puede apreciarse en las dos matrices de la Figura 6.1. En problemas de programación

entera, que son en los que el algoritmo de Benders ha sido particularmente exitoso, esta relación mediante dualidad se desvanece y la comparación entre ambos resulta más compleja.

Los problemas susceptibles de ser resueltos más eficientemente mediante el uso de técnicas de descomposición aparecen en multitud de aplicaciones. El primer capítulo del libro Conejo y otros (2006) está dedicado íntegramente a describir este tipo de aplicaciones. Algunas situaciones en las que aparecen estructuras adecuadas para este tipo de técnicas son las siguientes:

Múltiples plantas/fábricas. Sistemas de producción con multitud de unidades similares entre sí como plantas, fábricas, almacenes, hospitales,... La gestión eficiente de estos sistemas pasa habitualmente por la optimización "independiente" de la gestión de cada una de las unidades que los componen, pero sujetos a algunas restricciones agregadas, como la de satisfacer una cierta demanda de manera conjunta. Un ejemplo de este tipo de problemas lo hemos visto ya en el problema de planificación energética de la Sección 5.5.1.

Optimización dinámica. Cuando tenemos que gestionar un determinado sistema a lo largo de distintos periodos de tiempo, es común que los problemas de optimización de cada etapa sean relativamente independientes entre sí, pero con unas "linking constraints" que enganchen las decisiones de una etapa con las siguientes.

Optimización bajo incertidumbre. En muchos problemas hay incertidumbre sobre el valor que tomarán ciertos parámetros, como podrían ser precios y/o demandas. En estos casos, es natural atribuir probabilidades a los distintos escenarios que pueden producirse. Como la cantidad de escenarios puede ser muy grande, es habitual diseñar técnicas de descomposición que combinen la resolución independiente de los distintos escenarios con la inclusión de "linking constraints" que aseguren que la solución resultante es factible y óptima en el problema completo, no sólo en algunos de los escenarios.

6.1.2 Ejemplo ilustrativo: Problema de transporte (PTM)

Para ilustrar las técnicas de descomposición de este tema vamos a apoyarnos en un problema de flujo en redes. Más concretamente, trabajaremos con un problema de transporte "multicommodity", en el que distintas "mercancías" fluyen por la red. Más concretamente, tendremos los siguientes elementos:

Orígenes. O, conjunto de almacenes de los cuales salen las mercancías.

Destinos. D, conjunto de destinos a los cuales tienen que llegar las mercancías.

Mercancías. M, conjunto de mercancías a transportar entre orígenes y destinos.

Costes. Para cada par origen-destino $(i, j) \in O \times D$ y cada mercancía m, tenemos c_{ijm}, el coste de enviar una unidad de m desde i a j.

Capacidades. Para cada par origen-mercancía $(i, m) \in O \times M$, tenemos $s_{im} \geq 0$, el suministro disponible de la mercancía m en el almacén i.

Demandas. Para cada par destino-mercancía $(j, m) \in D \times M$, tenemos $d_{jm} \geq 0$, la demanda de la mercancía m en el destino j.

Flujos. Las variables del problema de optimización. Para cada par origen-destino (i, j) y cada mercancía m, hay que determinar f_{ijm}, la cantidad de mercancía m a enviar desde i a j.

Problema equilibrado. Para que el problema resultante tenga soluciones factibles es necesario que, para todo $m \in M$, $\sum_{i\in O} s_{im} \geq \sum_{j\in D} d_{jm}$. Se puede asumir, sin pérdida de generalidad, que el problema es equilibrado en todas las mercancías: para todo $m \in M$, $\sum_{i\in O} s_{im} = \sum_{j\in D} d_{jm}$. Basta con añadir un destino ficticio al conjunto D que, para cada $m \in M$, tiene demanda $\sum_{i\in O} s_{im} - \sum_{j\in D} d_{jm}$ y costes de envío asociados igual a cero. Los envíos a este destino ficticio representan las unidades de cada mercancía que se quedan en cada almacén.

Tenemos entonces el problema de elegir los flujos f_{ijm} para minimizar los costes de transporte, al mismo tiempo que se abastecen las demandas y se respetan las capacidades de los almacenes:

Problema de transporte multimercancía

$$\begin{aligned}
\text{minimizar} \quad & \sum_{m\in M}\sum_{i\in O}\sum_{j\in D} c_{ijm} f_{ijm} \\
\text{sujeto a} \quad & \sum_{j\in D} f_{ijm} = s_{im} && i \in O,\ m \in M \\
& \sum_{i\in O} f_{ijm} = d_{jm} && j \in D,\ m \in M \\
& f_{ijm} \geq 0 && i \in O,\ j \in D,\ m \in M.
\end{aligned} \tag{PTM}$$

Claramente, este es un problema que se puede descomponer en $|M|$ problemas independientes, uno por cada mercancía. En este caso, su resolución eficiente no requiere de ninguna técnica específica. A continuación presentamos dos variantes naturales de este problema, una con restricciones complicantes y otra con variables complicantes.

En primer lugar, el problema con restricciones complicantes surge de considerar una situación en la que el transporte entre cada par origen-destino (i, j) se hace mediante vehículos que, en total, admiten un peso máximo $p_{ij}^{\max}$. Entonces, el problema se convierte en:

PTM con restricciones complicantes

$$\begin{aligned}
\text{minimizar} \quad & \sum_{m\in M}\sum_{i\in O}\sum_{j\in D} c_{ijm} f_{ijm} \\
\text{sujeto a} \quad & \sum_{j\in D} f_{ijm} = s_{im} && i \in O,\ m \in M \\
& \sum_{i\in O} f_{ijm} = d_{jm} && j \in D,\ m \in M \\
& \sum_{m\in M} p_m f_{ijm} \leq p_{ij}^{\max} && i \in O,\ j \in D \quad \text{(Restricciones complicantes)} \\
& f_{ijm} \geq 0 && i \in O,\ j \in D,\ m \in M,
\end{aligned} \tag{PTM-RC}$$

donde cada coeficiente p_m representa el peso de la mercancía asociada. Este problema ya no puede ser resuelto tomando el subproblema de cada mercancía de manera independiente, pues no garantizaríamos que se cumpliesen las restricciones de capacidad agregadas. Si pudiésemos prescindir de estas restricciones agregadas recuperaríamos la descomposición del problema en subproblemas independientes, que es lo que explotará el algoritmo de Dantzig-Wolfe.

Para la variante de (PTM) con variables complicantes, supongamos que tenemos que decidir simultáneamente qué almacenes abrir y qué enviar entre cada almacén abierto y cada destino. Tenemos un problema llamado de localización acoplado con nuestro problema de transporte original. Más concretamente, para cada almacén $i \in O$ tenemos que decidir si lo queremos construir o no, y esta decisión la modelizaremos con variables binarias $y_i \in \{0,1\}$. El coste fijo asociado a construir el almacén i viene dado por $c_i^{\mathrm{f}} \geq 0$. En el problema resultante las restricciones deberán contemplar que únicamente se podrán enviar mercancías desde los almacenes que se decida construir:

PTM con variables complicantes

$$
\begin{aligned}
\text{minimizar} \quad & \sum_{i\in M} c_i^{\mathrm{f}} y_i + \sum_{m\in M}\sum_{i\in O}\sum_{j\in D} c_{ijm} f_{ijm} \\
\text{sujeto a} \quad & \sum_{j\in D} f_{ijm} \leq s_{im} y_i && i \in O,\ m \in M \\
& \sum_{i\in O} f_{ijm} = d_{jm} && j \in D,\ m \in M \\
& y_i \in \{0,1\} && i \in O \quad \text{(Variables complicantes)} \\
& f_{ijm} \geq 0 && i \in O,\ j \in D,\ m \in M.
\end{aligned}
\tag{PTM-VC}
$$

La decisión de construir o no un determinado almacén afecta a todas las mercancías, de modo que la presencia de las variables y_i también impide la resolución independiente del problema asociado a cada mercancía. Además, por tratarse de variables binarias, pasamos de tener un problema de optimización continua a tener un problema de optimización discreta. Ahora no tiene sentido asumir que estamos ante un problema equilibrado en el que la suma de suministros coincide con la suma de demandas, pues eso implicaría automáticamente que se tienen que construir todos los almacenes. Una vez fijado el valor de las variables de construcción, el problema resultante vuelve a ser un problema de optimización continua que se puede descomponer en problemas independientes, que es de lo de que se aprovechará el algoritmo de Benders.

6.1.3 Dantzig-Wolfe y Benders: Esquema general

Los algoritmos de Dantzig-Wolfe y Benders proceden de una manera similar, descomponiendo el problema de partida como sigue:

Problema maestro. Se encarga de las restricciones/variables complicantes.

Subproblema(s). Se encargan del resto de restricciones/variables.

Una de las claves del éxito de estos algoritmos radica no tanto en descomponer el problema original en el par maestro-subproblema, sino en el hecho de que el subproblema tenga una

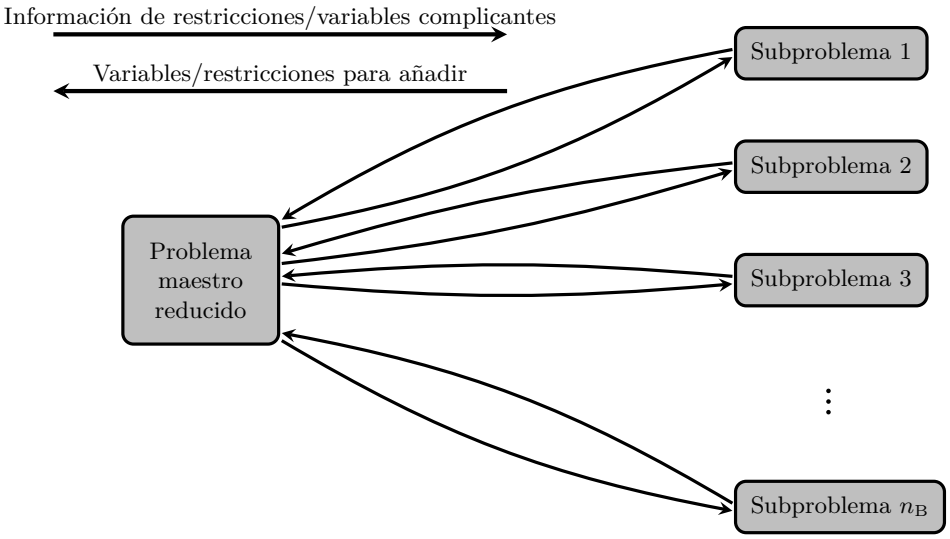


Figura 6.2: Esquema general de los algoritmos de Dantzig-Wolfe y Benders.

estructura adecuada para poder descomponerlo a su vez en una serie de subproblemas que puedan ser resueltos de manera independiente. Una vez se ha obtenido una descomposición como la que acabamos de comentar, ambos algoritmos proceden de acuerdo al siguiente esquema, ilustrado en la Figura 6.2:

(i) Se parte de una versión reducida del problema maestro, que contiene:

En problemas con restricciones complicantes, un subconjunto del conjunto de variables.

En problemas con variables complicantes, un subconjunto del conjunto de restricciones.

(ii) Los algoritmos resuelven iterativamente el problema maestro reducido y los subproblemas:

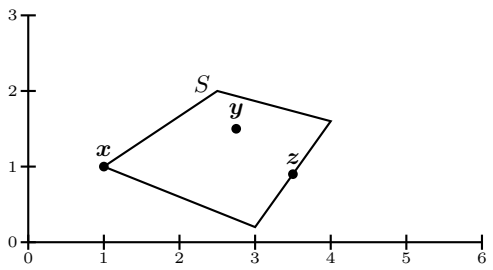
- El problema maestro reducido envía a los subproblemas información relativa a las variables/restricciones complicantes.
- Los subproblemas envían información relativa a variables o restricciones que deben ser añadidas al problema maestro reducido:
  - Variables en el caso de Dantzig-Wolfe, motivo por el que es una técnica de generación de columnas.
  - Restricciones en el caso de Benders, motivo por el que es una técnica de generación de filas.

6.2 Recordatorio de programación lineal

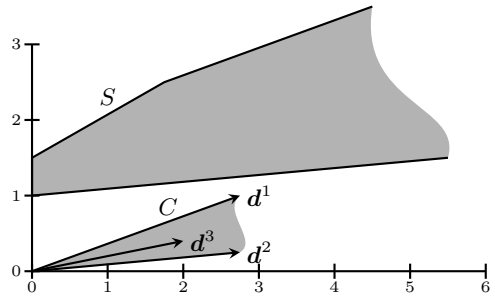
Antes de presentar formalmente los algoritmos de Dantzig-Wolfe y Benders es conveniente hacer un breve repaso de algunos conceptos y resultados de programación lineal en los que se apoyan dichos algoritmos.

6.2.1 Puntos extremos y direcciones extremas

Comenzamos recordando las definiciones de punto extremo y dirección extrema, ilustrados en la Figura 6.3 y que ya han sido usados previamente en estas notas.



(a) Puntos extremos de un conjunto convexo: $\boldsymbol{x}$, $\boldsymbol{y}$ y $\boldsymbol{z}$ son puntos de S, pero sólo $\boldsymbol{x}$ es un punto extremo.



(b) Direcciones extremas de un conjunto convexo: $\boldsymbol{d}^1$, $\boldsymbol{d}^2$ y $\boldsymbol{d}^3$ son direcciones del conjunto S, pero sólo $\boldsymbol{d}^1$ y $\boldsymbol{d}^2$ son direcciones extremas. C es el cono de direcciones de S.

Figura 6.3: Ilustración de los conceptos de punto extremo y dirección extrema.

Punto extremo. Dado un conjunto convexo S, un punto $\boldsymbol{z} \in S$ es un punto extremo de S si no puede ser representado como una combinación convexa estricta de dos puntos distintos de S. Formalmente, si $\boldsymbol{z} = \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}$ con $\lambda \in (0,1)$, entonces $\boldsymbol{z} = \boldsymbol{x} = \boldsymbol{y}$.

Dirección. Dado un conjunto convexo S, un vector $\boldsymbol{d} \neq \boldsymbol{0}$ es una *dirección* de S si para cualquier punto $\boldsymbol{x} \in S$, el rayo de vértice $\boldsymbol{x}$ y dirección $\boldsymbol{d}$ está contenido en S. Formalmente, para todo $\lambda \geq 0$, $\boldsymbol{x} + \lambda \boldsymbol{d} \in S$. Un conjunto acotado no tiene direcciones.

Dado un poliedro de la forma $S = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$ y una dirección $\boldsymbol{d} \neq \boldsymbol{0}$, entonces $\boldsymbol{d}$ es una dirección de S si y sólo si, para todo $\boldsymbol{x} \in S$ y todo $\lambda \geq 0$,

$$\boldsymbol{A}(\boldsymbol{x} + \lambda \boldsymbol{d}) \leq \boldsymbol{b} \quad \text{y} \quad \boldsymbol{x} + \lambda \boldsymbol{d} \geq \boldsymbol{0}.$$

Como $\boldsymbol{x} \in S$ tenemos que $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ y $\boldsymbol{x} \geq \boldsymbol{0}$ con lo que, usando que $\lambda \geq 0$, tendremos que $\boldsymbol{d} \neq \boldsymbol{0}$ es una dirección de S si y sólo si

$$\boldsymbol{A}\boldsymbol{d} \leq \boldsymbol{0} \quad \text{y} \quad \boldsymbol{d} \geq \boldsymbol{0}.$$

Decimos que dos direcciones $\boldsymbol{d}^1$ y $\boldsymbol{d}^2$ son distintas si no existe $\lambda > 0$ tal que $\boldsymbol{d}^1 = \lambda \boldsymbol{d}^2$. Entonces, es fácil ver que el conjunto de direcciones se puede representar como

$$D = \{\boldsymbol{d} : \boldsymbol{A}\boldsymbol{d} \leq \boldsymbol{0}, \boldsymbol{1}\boldsymbol{d} = 1, \boldsymbol{d} \geq \boldsymbol{0}\}. \tag{6.1}$$

Dirección extrema. Dado un conjunto convexo S, una dirección $\boldsymbol{d}$ de S es una dirección extrema si no puede ser representada como una combinación cónica estricta de dos direcciones distintas de S. Formalmente, $\boldsymbol{d}$ es una dirección extrema de S si, dadas dos direcciones $\boldsymbol{v}^1$ y $\boldsymbol{v}^2$ de S, la igualdad $\boldsymbol{d} = \lambda_1 \boldsymbol{v}^1 + \lambda_2 \boldsymbol{v}^2$ con $\lambda_1 > 0$ y $\lambda_2 > 0$ implica que $\boldsymbol{d}$, $\boldsymbol{v}^1$ y $\boldsymbol{v}^2$ representan la misma dirección.

6.2.2 Teorema de representación de Carathéodory

A continuación presentamos dos resultados clásicos relativos a la geometría de los poliedros, cuyas demostraciones pueden encontrarse en cualquier libro de programación lineal como Bazaraa y otros (2009).

Teorema 6.1 (Teorema de Carathéodory)**.** *Sea $S \subseteq \mathbb{R}^n$ el poliedro no vacío definido por $\{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$, entonces un punto $\bar{\boldsymbol{x}}$ pertenece a S si y sólo si puede ser representado como una combinación convexa de los puntos extremos de S más una combinación cónica de sus direcciones extremas.*

Equivalentemente, si denotamos por $\boldsymbol{x}^k$ con $k \in K$ a los puntos extremos de S y por $\boldsymbol{d}^l$ con $l \in L$ a sus direcciones extremas, entonces $\bar{\boldsymbol{x}} \in S$ si y sólo si

$$\begin{aligned} \bar{\boldsymbol{x}} &= \sum_{k\in K} \lambda_k \boldsymbol{x}^k + \sum_{l\in L} \mu_l \boldsymbol{d}^l, \quad \text{donde} \\ \sum_{k\in K} \lambda_k &= 1, \\ \lambda_k &\geq 0, \quad k \in K, \quad y \\ \mu_l &\geq 0, \quad l \in L. \end{aligned} \tag{6.2}$$

Proposición 6.2. *Dado un problema de programación lineal con región factible no vacía, las siguientes afirmaciones son ciertas:*

(i) *Si el problema tiene solución óptima finita, al menos uno de los puntos extremos de la región factible será óptimo.*

(ii) *Si el problema es de minimización, entonces tiene solución óptima finita si y sólo si no existe ninguna dirección extrema $\boldsymbol{d}$ de la región factible tal que $\boldsymbol{c}^\intercal \boldsymbol{d} < 0$.*[1]

6.2.3 Cálculo de una dirección extrema en un problema sin óptimo finito

La mayoría de los optimizadores lineales, en caso de encontrarnos con un problema sin óptimo finito, nos devolverán una dirección extrema del mismo en la cual podemos movernos haciendo que la función objetivo mejore indefinidamente. Esto es algo que el propio método símplex identifica en este tipo de problemas. A continuación presentamos una forma directa de calcular este tipo de direcciones extremas.

Supongamos que tenemos un problema de programación lineal de la forma

$$\begin{aligned} \text{minimizar} \quad & \boldsymbol{c}^\intercal \boldsymbol{x} \\ \text{sujeto a} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0}. \end{aligned}$$

Considérese ahora el siguiente problema

$$\begin{aligned} \text{minimizar} \quad & \boldsymbol{c}^\intercal \boldsymbol{d} \\ \text{sujeto a} \quad & \boldsymbol{A}\boldsymbol{d} \leq \boldsymbol{0} \\ & \boldsymbol{1}\boldsymbol{d} = 1 \\ & \boldsymbol{d} \geq \boldsymbol{0}. \end{aligned}$$

[1] En el caso de problemas de maximización, la condición cambia a $\boldsymbol{c}^\intercal \boldsymbol{d} > 0$.

El conjunto factible de este problema se corresponde con el conjunto de direcciones asociadas a la región factible del problema original, como vimos en la Ecuación (6.1). Además, como es habitual, las soluciones óptimas obtenidas al aplicar el método símplex serán direcciones extremas. Por último, si la función objetivo en el óptimo es menor que cero, entonces por el segundo apartado de la Proposición 6.2 tendremos que la dirección extrema óptima es una dirección en la cual la función objetivo del problema original mejora indefinidamente.

Para terminar, presentamos algunos comentarios relativos a este procedimiento de cálculo de direcciones extremas:

- Dado que, para todo $\lambda \geq 0$, $\boldsymbol{d}$ y $\lambda\boldsymbol{d}$ representan la misma dirección, la restricción $\mathbf{1}\boldsymbol{d} = 1$ es importante para asegurar que el problema tiene óptimo finito.
- En caso de que el problema original tenga una restricción de la forma $\boldsymbol{A}_i^{\mathrm{f}}\boldsymbol{x} \geq b_i$, en el problema auxiliar corresponderá también una restricción $\boldsymbol{A}_i^{\mathrm{f}}\boldsymbol{d} \geq 0$.[2]
- En caso de que el problema original tenga una restricción de la forma $\boldsymbol{A}_i^{\mathrm{f}}\boldsymbol{x} = b_i$, en el problema auxiliar corresponderá también una restricción de igualdad $\boldsymbol{A}_i^{\mathrm{f}}\boldsymbol{d} = 0$.[3]
- Si el problema original no tiene la restricción de que las variables sean no negativas, tampoco debe tener dicha restricción el problema auxiliar. Además, la condición $\mathbf{1}\boldsymbol{d} = 1$ ya no asegura que el problema auxiliar sea acotado, con lo que debe ser reemplazada, por ejemplo, por cotas inferiores y superiores para las variables. Para asegurar que no se elimina ninguna dirección del conjunto, las cotas inferiores deben ser negativas y las superiores positivas. Por ejemplo, exigiendo que, para todo j, $d_j \in [-1, 1]$.
- Si el problema original es de maximizar, entonces también el subproblema será de maximizar y las direcciones extremas en las que el objetivo mejora indefinidamente se corresponden son soluciones con función objetivo mayor que cero.

6.2.4 Dualidad y costes reducidos

Supongamos que tenemos un problema de programación lineal en forma estándar y consideremos también su problema dual:

Problema primal P		**Problema dual D**	
minimizar	$\boldsymbol{c}^\top\boldsymbol{x}$	maximizar	$\boldsymbol{b}^\top\boldsymbol{w}$
sujeto a	$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$	sujeto a	$\boldsymbol{A}^\top\boldsymbol{w} \leq \boldsymbol{c}$.
	$\boldsymbol{x} \geq \boldsymbol{0}$		

Método símplex y costes reducidos

Recordemos un poco la resolución del problema primal mediante el método símplex. En cada iteración del algoritmo tendremos una solución básica $\boldsymbol{x} = (\boldsymbol{x}_B, \boldsymbol{x}_N)$. Denotemos por $\boldsymbol{A}_B$ a

[2] Esta restricción para el problema auxiliar se obtiene sin más que ver $\boldsymbol{A}_i^{\mathrm{f}}\boldsymbol{x} \geq b_i$ como $-\boldsymbol{A}_i^{\mathrm{f}}\boldsymbol{x} \leq b_i$, obteniendo $-\boldsymbol{A}_i^{\mathrm{f}}\boldsymbol{d} \leq 0$ para el problema auxiliar, que es equivalente a $\boldsymbol{A}_i^{\mathrm{f}}\boldsymbol{d} \geq 0$.

[3] Esta restricción para el problema auxiliar se obtiene sin más que descomponer $\boldsymbol{A}_i^{\mathrm{f}}\boldsymbol{x} = b_i$ en $\boldsymbol{A}_i^{\mathrm{f}}\boldsymbol{x} \leq b_i$ y $\boldsymbol{A}_i^{\mathrm{f}}\boldsymbol{x} \geq b_i$ y combinar las restricciones asociadas para el problema auxiliar: $\boldsymbol{A}_i^{\mathrm{f}}\boldsymbol{d} \leq 0$ y $\boldsymbol{A}_i^{\mathrm{f}}\boldsymbol{d} \geq 0$.

la matriz básica, formada por las columnas de $\boldsymbol{A}$ asociadas a las variables básicas. Por tanto, $\boldsymbol{x}_B = \boldsymbol{A}_B^{-1} b$ y $\boldsymbol{x}_N = \boldsymbol{0}$.

El método símplex va saltando de solución básica factible en solución básica factible hasta que alcanza una solución óptima $\boldsymbol{x}$. Para comprobar la optimalidad de la solución $\boldsymbol{x}$ es necesario calcular, para cada variable no básica j, su coste reducido $z_j - c_j$, donde $z_j = \boldsymbol{c}_B^\intercal \boldsymbol{A}_B^{-1} \boldsymbol{A}_j^{\mathrm{c}}$. Una condición necesaria y suficiente de optimalidad, en la que se apoya el método símplex, es que todos los costes reducidos sean menores o iguales que cero.

Solución dual asociada a una base $\boldsymbol{A}_B$

Dada una base $\boldsymbol{A}_B$, además de tener asociada una solución básica del primal, tiene también asociada una solución del problema dual, dada por $\boldsymbol{w}^\intercal = \boldsymbol{c}_B^\intercal \boldsymbol{A}_B^{-1}$. Los costes reducidos del primal también pueden ser calculados con la ayuda de la solución asociada del dual, pues $z_j = \boldsymbol{c}_B^\intercal \boldsymbol{A}_B^{-1} \boldsymbol{A}_j^{\mathrm{c}} = \boldsymbol{w}^\intercal \boldsymbol{A}_j^{\mathrm{c}}$.

El siguiente resultado muestra que, en el caso de que $\boldsymbol{A}_B$ sea una base óptima para el primal, entonces también lo es para el dual.

Proposición 6.3. *Supongamos que el problema primal viene dado en forma estándar y que $\bar{\boldsymbol{x}}$ es una solución óptima con base asociada $\boldsymbol{A}_B$. Entonces,*

$$\bar{\boldsymbol{w}}^\intercal = \boldsymbol{c}_B^\intercal \boldsymbol{A}_B^{-1}$$

es una solución óptima del problema dual.

Demostración. El primal tiene por restricciones $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, $\boldsymbol{x} \geq \boldsymbol{0}$ y el dual $\boldsymbol{w}^\intercal \boldsymbol{A} \leq \boldsymbol{c}$. Tenemos que probar tanto factibilidad como optimalidad de $\bar{\boldsymbol{w}}$.

Factibilidad. Para cada variable no básica j tenemos $\bar{\boldsymbol{w}}^\intercal \boldsymbol{A}_j^{\mathrm{c}} = \boldsymbol{c}_B^\intercal \boldsymbol{A}_B^{-1} \boldsymbol{A}_j^{\mathrm{c}} = z_j$. Ahora, la condición de optimalidad del primal nos dice que $z_j - c_j \leq 0$ o, equivalentemente, que $\bar{\boldsymbol{w}}^\intercal \boldsymbol{A}_j^{\mathrm{c}} = z_j \leq c_j$, que es justamente la condición de factibilidad del dual.

Optimalidad. $\boldsymbol{c}_B^\intercal \bar{\boldsymbol{x}} = \boldsymbol{c}_B^\intercal \boldsymbol{A}_B^{-1} \boldsymbol{b} = \bar{\boldsymbol{w}}^\intercal \boldsymbol{b}$. Como la función objetivo de primal y dual coinciden, el teorema de dualidad débil asegura que $\bar{\boldsymbol{w}}$ es solución óptima del problema dual. □

Es importante notar que, en el resultado anterior se demuestra que es equivalente pedir que se cumpla una determinada restricción del dual, $\bar{\boldsymbol{w}}^\intercal \boldsymbol{A}_j^{\mathrm{c}} \leq c_j$ (factibilidad dual), a que el coste reducido de la variable asociada sea no negativo, $z_j - c_j \leq 0$ (optimalidad primal). Esta relación entre condiciones de factibilidad y optimalidad de los problemas primal y dual es de gran importancia en el diseño de algoritmos. Además, recordemos que no es una relación exclusiva de la programación lineal, pues ya se discutió en el Tema 5.

6.3 Métodos de generación de columnas. Restricciones complicantes y algoritmo de Dantzig-Wolfe

El algoritmo de Dantzig-Wolfe (Dantzig y Wolfe, 1960) fue diseñado para resolver problemas de programación lineal de gran tamaño pero con una estructura de bloques con restricciones

complicantes (Figura 6.1(a)). Consideremos el siguiente problema de programación lineal:

$$\begin{array}{ll} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimizar}} & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{sujeto a} & \boldsymbol{E}\boldsymbol{x} = \boldsymbol{e} \\ & \boldsymbol{B}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0}, \end{array} \tag{DW-Orig}$$

donde $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{c} \in \mathbb{R}^n$, $\boldsymbol{E} \in \mathbb{R}^{m_e \times n}$, $\boldsymbol{B} \in \mathbb{R}^{m_b \times n}$, $\boldsymbol{e} \in \mathbb{R}^{m_e}$ y $\boldsymbol{b} \in \mathbb{R}^{m_b}$. La matriz $\boldsymbol{E}$ se corresponde con las restricciones complicantes o de **E**nganche y la matriz $\boldsymbol{B}$ con las restricciones "fáciles", típicamente susceptibles de ser descompuestas por **B**loques. Dado que todo problema de programación lineal se puede transformar en uno equivalente en forma estándar, partir de un problema en forma estándar en la Ecuación (DW-Orig) no supone pérdida de generalidad.

6.3.1 Reformulación del problema: Problema maestro

La idea principal del algoritmo de Dantzig-Wolfe es apoyarse en el Teorema de Carathéodory (Teorema 6.1) para trabajar con la representación del conjunto $B = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{B}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$ mediante sus puntos y direcciones extremas.[4] Más concretamente, sean $\boldsymbol{x}^k$, con $k \in K$, los puntos extremos de B y $\boldsymbol{d}^l$, con $l \in L$, sus direcciones extremas, entonces todo $\boldsymbol{x} \in B$ se puede representar como

$$\begin{aligned} \boldsymbol{x} &= \sum_{k\in K} \lambda_k \boldsymbol{x}^k + \sum_{l\in L} \mu_l \boldsymbol{d}^l, \quad \text{donde} \\ \sum_{k\in K} \lambda_k &= 1, \\ \lambda_k &\geq 0, \quad k \in K, \quad \text{y} \\ \mu_l &\geq 0, \quad l \in L. \end{aligned}$$

Usamos ahora estas expresiones para sustituir las restricciones $\boldsymbol{B}\boldsymbol{x} = \boldsymbol{b}$, $\boldsymbol{x} \geq \boldsymbol{0}$ en el problema (DW-Orig), obteniendo

$$\begin{array}{ll} \underset{\boldsymbol{x}\in\mathbb{R}^n,\, \boldsymbol{\lambda}\in\mathbb{R}^{|K|},\, \boldsymbol{\mu}\in\mathbb{R}^{|L|}}{\text{minimizar}} & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{sujeto a} & \boldsymbol{E}\boldsymbol{x} = \boldsymbol{e} \\ & \boldsymbol{x} = \displaystyle\sum_{k\in K} \lambda_k \boldsymbol{x}^k + \sum_{l\in L} \mu_l \boldsymbol{d}^l \\ & \displaystyle\sum_{k\in K} \lambda_k = 1 \\ & \lambda_k \geq 0, \quad k \in K \\ & \mu_l \geq 0, \quad l \in L. \end{array} \tag{6.3}$$

[4] Aunque el Teorema de Carathéodory se enunció sobre conjuntos de la forma $\{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$, esto no supone un problema, pues toda restricción de igualdad $\boldsymbol{B}_i^{\mathrm{f}} = b_i$ se puede descomponer como $\boldsymbol{B}_i^{\mathrm{f}} \leq b_i$ y $-\boldsymbol{B}_i^{\mathrm{f}} \leq -b_i$. Por tanto, el Teorema de Carathéodory se puede aplicar sin problema sobre el conjunto B.

Podemos ir todavía un paso más allá, y sustituir $\boldsymbol{x}$ por $\sum_{k\in K}\lambda_k\boldsymbol{x}^k+\sum_{l\in L}\mu_l\boldsymbol{d}^l$ en la formulación anterior, obteniendo

$$\begin{aligned}
\underset{\boldsymbol{\lambda}\in\mathbb{R}^{|K|},\,\boldsymbol{\mu}\in\mathbb{R}^{|L|}}{\text{minimizar}} \quad & \sum_{k\in K}\lambda_k\boldsymbol{c}^\intercal\boldsymbol{x}^k+\sum_{l\in L}\mu_l\boldsymbol{c}^\intercal\boldsymbol{d}^l \\
\text{sujeto a} \quad & \sum_{k\in K}\lambda_k\boldsymbol{E}\boldsymbol{x}^k+\sum_{l\in L}\mu_l\boldsymbol{E}\boldsymbol{d}^l=\boldsymbol{e} \quad \text{(Restrs. complicantes)} \\
& \sum_{k\in K}\lambda_k=1 \quad \text{(Restr. de convexidad)} \\
& \lambda_k\geq 0, \quad k\in K \\
& \mu_l\geq 0, \quad l\in L.
\end{aligned} \tag{DW-Ma}$$

El problema (DW-Ma) es el problema maestro de Dantzig-Wolfe, que trabaja explícitamente con las restricciones complicantes, las asociadas a la matriz $\boldsymbol{E}$, e implícitamente con las restricciones asociadas a la matriz $\boldsymbol{B}$. El Teorema de Carathéodory nos asegura que los problemas (DW-Orig) y (DW-Ma) son equivalentes. Por el momento, esta reformulación no hace el problema más sencillo. El problema original tenía tamaño $(m_e+m_b)\cdot n$ y el problema maestro tiene tamaño $(m_e+1)\cdot(|K|+|L|)$. Aunque el número de restricciones se ha reducido, el número de variables es ahora muy grande, pues el número de puntos y direcciones extremas de un poliedro suele crecer exponencialmente como función del número de semiespacios que lo definen, m_b en este caso. En otras palabras, enumerar todas las variables del problema (DW-Ma) es en sí mismo un problema exponencial, con lo que simplemente escribir la formulación explícita del problema (DW-Ma) es algo intratable computacionalmente.

Es ahora donde entra la esencia del algoritmo de Dantzig-Wolfe: la generación de columnas. En vez de trabajar explícitamente con las $|K|+|L|$ variables del problema (DW-Ma), el algoritmo las irá añadiendo, a medida que vayan siendo necesarias, junto con las columnas correspondientes de la matriz de restricciones. En el óptimo del problema (DW-Ma) habrá m_e+1 variables básicas, tomando el resto de variables el valor 0: la mayoría de los puntos y direcciones extremas no entran en juego en el óptimo. Por tanto, una de las claves para un buen rendimiento del algoritmo de Dantzig-Wolfe es que únicamente sea necesario añadir una cantidad relativamente pequeña de puntos y direcciones extremas de B antes de encontrar el óptimo, de manera similar al funcionamiento del método símplex, que habitualmente sólo recorre una pequeña cantidad de los puntos extremos de la región factible.

6.3.2 Problema maestro reducido

El problema maestro reducido se define a partir del problema maestro (DW-Ma) sin más que partir de únicamente un subconjunto del conjunto total de variables. Hay distintas formas de identificar este subconjunto inicial y más adelante discutiremos explícitamente una de ellas. Por el momento, supongamos que ya disponemos de conjuntos reducidos de puntos y direcciones

extremas $K^{\mathrm{R}} \subset K$ y $L^{\mathrm{R}} \subset L$. Entonces, el problema maestro reducido se define como

$$\begin{aligned} \underset{\boldsymbol{\lambda}\in\mathbb{R}^{|K^{\mathrm{R}}|},\, \boldsymbol{\mu}\in\mathbb{R}^{|L^{\mathrm{R}}|}}{\text{minimizar}} \quad & \sum_{k\in K^{\mathrm{R}}} \lambda_k \boldsymbol{c}^\intercal \boldsymbol{x}^k + \sum_{l\in L^{\mathrm{R}}} \mu_l \boldsymbol{c}^\intercal \boldsymbol{d}^l \\ \text{sujeto a} \quad & \sum_{k\in K^{\mathrm{R}}} \lambda_k \boldsymbol{E}\boldsymbol{x}^k + \sum_{l\in L^{\mathrm{R}}} \mu_l \boldsymbol{E}\boldsymbol{d}^l = \boldsymbol{e} \\ & \sum_{k\in K^{\mathrm{R}}} \lambda_k = 1 \\ & \lambda_k \geq 0, \quad k \in K^{\mathrm{R}} \\ & \mu_l \geq 0, \quad l \in L^{\mathrm{R}}. \end{aligned} \tag{DW-Red}$$

Cualquier solución factible del problema (DW-Red) se corresponde con una solución factible del problema (DW-Ma), sin más que asignarle valor 0 a las variables faltantes. A su vez, la ecuación $\boldsymbol{x} = \sum_{k\in K^{\mathrm{R}}} \lambda_k \boldsymbol{x}^k + \sum_{l\in L^{\mathrm{R}}} \mu_l \boldsymbol{d}^l$ nos permite encontrar la solución correspondiente del problema original (DW-Orig).

La complejidad del algoritmo radica en determinar si una solución factible del problema reducido (DW-Red) es o no óptima en el problema completo (DW-Ma). Además, en caso de que no sea así, hay que identificar alguna variable faltante en el problema (DW-Red) (un punto extremo o una dirección extrema de B) que, una vez incluida, ayude a mejorar la función objetivo.

6.3.3 Subproblemas para la generación de columnas

Supongamos que tenemos una solución básica factible del problema (DW-Red). Esta solución se corresponde también con una solución básica factible del problema (DW-Ma) y también con una solución $(\boldsymbol{w}, \alpha)$ de su dual; donde $\boldsymbol{w}$ es el vector de variables duales asociadas a las m_e restricciones complicantes y α es la variable dual asociada a la restricción de convexidad.

Para determinar la optimalidad de la solución en el problema primal podemos calcular los costes reducidos asociados a las variables no básicas. Sin embargo, como ya hemos comentado, tener que manipular explícitamente todas las variables de este problema es algo computacionalmente prohibitivo. Por otro lado, en la Sección 6.2.4 vimos cómo calcular los costes reducidos usando la solución del dual. Aquí es clave el hecho de que, si bien los problemas (DW-Red) y (DW-Ma) tienen distinto número de variables, ambos tienen las mismas restricciones y, por tanto, sus duales tienen las mismas variables. Dado un punto extremo $\boldsymbol{x}^k$, su variable asociada en el problema (DW-Ma) es λ_k, con coste asociado $\boldsymbol{c}^\intercal \boldsymbol{x}^k$. La columna asociada de la matriz de restricciones es $\begin{pmatrix} \boldsymbol{E}\boldsymbol{x}^k \\ 1 \end{pmatrix}$. Por tanto, el coste reducido asociado se puede calcular como

$$(\boldsymbol{w}, \alpha)^\intercal \begin{pmatrix} \boldsymbol{E}\boldsymbol{x}^k \\ 1 \end{pmatrix} - \boldsymbol{c}^\intercal \boldsymbol{x}^k = \boldsymbol{w}^\intercal \boldsymbol{E}\boldsymbol{x}^k + \alpha - \boldsymbol{c}^\intercal \boldsymbol{x}^k.$$

Si este coste reducido es positivo, entonces λ_k es una candidata a entrar en la base para mejorar la solución actual. Habría que añadir λ_k al problema maestro reducido y resolverlo de nuevo. El proceso de añadir λ_k conlleva añadir la columna de la matriz de restricciones asociada al punto extremo $\boldsymbol{x}^k$, y de ahí el nombre de generación de columnas. Sin embargo,

esta formulación a través del dual no parece resolver el inconveniente de tener que verificar una cantidad prohibitiva de costes reducidos para poder establecer la optimalidad de una solución dada. La forma de sortear este inconveniente es calcular el punto extremo con un mayor coste reducido asociado mediante la resolución del siguiente problema de optimización:

$$\begin{array}{ll} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimizar}} & \boldsymbol{c}^\intercal\boldsymbol{x} - \boldsymbol{w}^\intercal\boldsymbol{E}\boldsymbol{x} - \alpha \\ \text{sujeto a} & \boldsymbol{B}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0}. \end{array}$$

Hemos cambiado el signo en la función objetivo para obtener un problema de minimización. Como α es una constante en este subproblema, podemos prescindir de ella, obteniendo

$$\begin{array}{ll} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimizar}} & \boldsymbol{c}^\intercal\boldsymbol{x} - \boldsymbol{w}^\intercal\boldsymbol{E}\boldsymbol{x} \\ \text{sujeto a} & \boldsymbol{B}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0}. \end{array} \tag{DW-SP}$$

La región factible de este problema es el conjunto B, con lo que si el subproblema (DW-SP) tiene un óptimo finito, dicha solución será un punto extremo $\bar{\boldsymbol{x}}$ de B, con función objetivo z_{SP}. Si $z_{\text{SP}} \geq \alpha$ sabremos que no hay ningún punto extremo de B con coste reducido positivo.

De modo similar, también podemos ver si alguna dirección extrema de B tiene coste reducido positivo. En este caso, el coste reducido asociado a una dirección extrema $\boldsymbol{d}$ viene dado directamente por $\boldsymbol{w}^\intercal\boldsymbol{E}\boldsymbol{d} - \boldsymbol{c}^\intercal\boldsymbol{d}$, pues la variable dual α se corresponde con la restricción de convexidad que no afecta a las direcciones extremas. Si hay una dirección $\boldsymbol{d}$ para la cual $\boldsymbol{w}^\intercal\boldsymbol{E}\boldsymbol{d} - \boldsymbol{c}^\intercal\boldsymbol{d} > 0$, entonces el subproblema (DW-SP) es no acotado, podemos movernos indefinidamente a lo largo de la dirección $\boldsymbol{d}$. De hecho, este razonamiento combinado con el apartado (II) de la Proposición 6.2 implica que, si $B \neq \emptyset$, entonces el subproblema (DW-SP) no tiene óptimo finito si y sólo si existe una dirección extrema $\boldsymbol{d}$ tal que $\boldsymbol{c}^\intercal\boldsymbol{d} - \boldsymbol{w}^\intercal\boldsymbol{E}\boldsymbol{d} < 0$.

Por tanto, podemos estudiar simultáneamente los costes reducidos de puntos y direcciones extremas de B a través del subproblema (DW-SP). Pueden darse los siguientes casos:

- El subproblema (DW-SP) no tiene soluciones factibles. Entonces $B = \emptyset$ y el problema original (DW-Orig) tampoco tiene soluciones factibles.
- El subproblema (DW-SP) tiene como solución óptima el punto extremo $\bar{\boldsymbol{x}}$, con variable asociada λ_k en el problema (DW-Ma):
  - Si $\boldsymbol{c}^\intercal\bar{\boldsymbol{x}} - \boldsymbol{w}^\intercal\boldsymbol{E}\bar{\boldsymbol{x}} \geq \alpha$, entonces el coste reducido de λ_k es menor o igual que cero y, por la optimalidad de $\bar{\boldsymbol{x}}$, también el asociado a cualquier otro punto extremo de B. Además, el hecho de tener un óptimo finito asegura que tampoco hay direcciones extremas con coste reducido positivo. Tendríamos asegurado que la solución actual del problema (DW-Red) es óptima en el problema (DW-Ma).
  - Si $\boldsymbol{c}^\intercal\bar{\boldsymbol{x}} - \boldsymbol{w}^\intercal\boldsymbol{E}\bar{\boldsymbol{x}} < \alpha$, entonces hay que añadir al problema (DW-Red) la variable λ_k, pues tiene coste reducido positivo.
- El subproblema (DW-SP) no tiene óptimo finito. En este caso tendremos una dirección extrema $\boldsymbol{d}$ de B asociada a una variable μ_l en el problema (DW-Red) con coste reducido positivo. Hay que añadir μ_l a dicho problema.

Desde el punto de vista computacional este procedimiento es mucho más eficiente que calcular explícitamente los costes reducidos de todos los puntos y direcciones extremas. La razón es que, en la resolución del problema de programación lineal (DW-SP) mediante el método símplex, este únicamente visitará una cantidad relativamente reducida de puntos extremos (salvo en problemas patológicos que rara vez aparecen en la práctica).

Ya estamos en condiciones de presentar el esquema general del algoritmo de Dantzig-Wolfe:

(i) Definir el problema (DW-Red) a partir de un conjunto inicial de puntos y direcciones extremas y resolverlo.

(ii) Estudiar los costes reducidos con respecto al problema (DW-Ma) mediante la resolución del subproblema (DW-SP):

- Si todos los costes son mayores o iguales a α, entonces la solución actual es óptima y el algoritmo termina.
- En caso contrario, habremos identificado un punto o dirección extrema que tendremos que añadir al problema (DW-Red) y resolverlo nuevamente.

La convergencia en una cantidad finita de iteraciones del algoritmo resultante está garantizada por el hecho de que el conjunto B tiene una cantidad finita de puntos y direcciones extremas. Antes de presentar formalmente el esquema completo del algoritmo de Dantzig-Wolfe, que veremos en la Figura 6.4 de la Sección 6.3.6, hay una serie de aspectos prácticos de gran relevancia que debemos discutir y a los que dedicamos las secciones 6.3.4 y 6.3.5.

6.3.4 Consideraciones adicionales relativas al algoritmo

Inicialización y factibilidad en el problema (DW-Red)

A la hora de definir el subconjunto de puntos y direcciones extremas que se usan para inicializar el problema (DW-Red), hay que tener cuidado con asegurar que dicho problema tenga soluciones factibles (siempre que las tenga el problema (DW-Ma)).

Una de las formas más habituales de inicializar el problema (DW-Red) es hacer algo similar a lo que hace el método de las dos fases en el algoritmo del símplex: introducir una primera fase centrada en conseguir una solución factible. La forma de hacer esto será trabajando inicialmente con la siguiente modificación del problema (DW-Red):

$$
\begin{aligned}
\underset{\substack{\boldsymbol{\lambda}\in\mathbb{R}^{|K^{\mathrm{R}}|},\,\boldsymbol{\mu}\in\mathbb{R}^{|L^{\mathrm{R}}|},\\ \boldsymbol{y}^a\in\mathbb{R}^{m_e},\,z^a\in\mathbb{R}}}{\text{minimizar}} \quad & \sum_{i=1}^{m_e} y_i^a + z^a \\
\text{sujeto a} \quad & \sum_{k\in K^{\mathrm{R}}} \lambda_k \boldsymbol{E}_i^{\mathrm{f}} \boldsymbol{x}^k + \sum_{l\in L^{\mathrm{R}}} \mu_l \boldsymbol{E}_i^{\mathrm{f}} \boldsymbol{d}^l + y_i^a - z^a = e_i,\ i\in\{1\dots,m_e\} \\
& \sum_{k\in K} \lambda_k = 1 \\
& \lambda_k \geq 0, \quad k\in K, \qquad \mu_l \geq 0, \quad l\in L \\
& y_i^a \geq 0, \quad i\in\{1\dots,m_e\}, \qquad z^a \geq 0.
\end{aligned}
\tag{DW-Red$^{\mathrm{fac}}$}
$$

El algoritmo comenzará con $K^{\mathrm{R}} = \emptyset$ y $L^{\mathrm{R}} = \emptyset$, y estos conjuntos se irán poblando durante la primera fase. Es fácil ver que, durante toda esta fase, las variables artificiales $\boldsymbol{y}^a$ y z^a aseguran la factibilidad del problema (DW-Red$^{\mathrm{fac}}$). A diferencia del problema (DW-Red), donde los puntos y direcciones extremas se van añadiendo para mejorar la función objetivo del problema original, en esta primera fase el objetivo es conseguir que todas las variables artificiales valgan cero. Si esto no se consigue, entonces el problema (DW-Orig) no tiene soluciones factibles. En otro caso, en cuanto todas las variables artificiales sean cero, podremos iniciar la segunda fase, pasando ya a resolver el problema (DW-Red) con los conjuntos K^{R} y L^{R} identificados en la primera fase.

Para tener completamente definida la primera fase tenemos que ver también cómo son los subproblemas a resolver. El cálculo de los costes reducidos asociados a los puntos y direcciones extremas todavía no añadidos es muy similar. Como el conjunto de restricciones es el mismo, las variables duales $\boldsymbol{w}$ y α no cambian y simplemente hay que tener en cuenta que el vector de costes $\boldsymbol{c}$ ahora no está en la función objetivo:

$$\begin{array}{rl} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimizar}} & -\boldsymbol{w}^\intercal \boldsymbol{E}\boldsymbol{x} \\ \text{sujeto a} & \boldsymbol{B}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0}. \end{array} \tag{DW-SP$^{\mathrm{fac}}$}$$

La regla para añadir puntos y direcciones extremas es análoga a la del problema (DW-SP):

- La infactibilidad del subproblema (DW-SP$^{\mathrm{fac}}$) implica la del problema (DW-Orig).
- El subproblema (DW-SP$^{\mathrm{fac}}$) tiene como solución óptima el punto extremo $\bar{\boldsymbol{x}}$:
  - Si $-\boldsymbol{w}^\intercal \boldsymbol{E}\bar{\boldsymbol{x}} \geq \alpha$ y alguna variable artificiales era distinta de cero en la solución del problema (DW-Red$^{\mathrm{fac}}$), entonces el problema (DW-Orig) es infactible.
  - Si $-\boldsymbol{w}^\intercal \boldsymbol{E}\bar{\boldsymbol{x}} < \alpha$, entonces añadimos un punto extremo a K^{R}.
- Si el subproblema (DW-SP$^{\mathrm{fac}}$) no tiene óptimo finito, entonces añadimos una dirección extrema a L^{R}.

Cálculo de direcciones extremas

Hemos visto que, cuando el subproblema (DW-SP) no tiene óptimo finito, entonces hay una dirección extrema con coste reducido positivo que debemos añadir al problema (DW-Red). Aunque en esta situación la mayoría de los optimizadores lineales actuales ya devolverían esa dirección extrema al intentar resolver el subproblema (DW-SP), siempre podríamos identificarla mediante el procedimiento descrito en la Sección 6.2.3.

También conviene destacar que, en la mayoría de los problemas que aparecen en la práctica, el conjunto B es acotado, típicamente porque se conocen cotas superiores para todas las variables del problema. En estos casos no habrá direcciones extremas en la formulación del problema (DW-Ma) y por tanto no tendremos que preocuparnos de su cálculo.

Cálculo de cotas

Un aspecto importante del algoritmo de Dantzig-Wolfe, que comparte con la mayoría de algoritmos que se apoyan en la dualidad, es que permite generar, iteración a iteración, cotas superiores e inferiores del valor óptimo de la función objetivo. Esto permite establecer criterios de parada basados en la distancia, *gap*, entre las dos cotas. A continuación mostramos cómo calcular estas cotas:

Cota superior. Cada resolución del problema (DW-Red) proporciona una solución factible del problema (DW-Ma). Su función objetivo, $\sum_{k\in K^{\mathrm{R}}} \lambda_k \boldsymbol{c}^\intercal \boldsymbol{x}^k + \sum_{l\in L^{\mathrm{R}}} \mu_l \boldsymbol{c}^\intercal \boldsymbol{d}^l$, será una cota superior del valor óptimo buscado.

Cota inferior. Supongamos que, en una cierta iteración del algoritmo de Dantzig-Wolfe, el óptimo del subproblema (DW-SP) tiene función objetivo finita z_{SP}. Dada una solución factible del problema original (DW-Orig), $\bar{\boldsymbol{x}}$, esta también es una solución factible del subproblema (DW-SP) y, además, $\boldsymbol{E}\bar{\boldsymbol{x}} = \boldsymbol{e}$. Entonces,

$$\boldsymbol{c}^\intercal \bar{\boldsymbol{x}} - \boldsymbol{w}^\intercal \boldsymbol{E}\bar{\boldsymbol{x}} \geq z_{\mathrm{SP}},$$

de donde

$$\boldsymbol{c}^\intercal \bar{\boldsymbol{x}} \geq z_{\mathrm{SP}} + \boldsymbol{w}^\intercal \boldsymbol{E}\bar{\boldsymbol{x}} \overset{\boldsymbol{E}\bar{\boldsymbol{x}}=\boldsymbol{e}}{=} z_{\mathrm{SP}} + \boldsymbol{w}^\intercal \boldsymbol{e}.$$

Por tanto, como la desigualdad anterior se cumple para cualquier solución factible del problema original, tenemos que $z_{\mathrm{SP}} + \boldsymbol{w}^\intercal \boldsymbol{e}$ es una cota inferior de la solución óptima buscada. Otra forma de identificar esta cota inferior es trabajando directamente con el dual del problema (DW-Ma):

$$\begin{array}{lll} \underset{\boldsymbol{v},\beta}{\text{minimizar}} & \boldsymbol{v}^\intercal \boldsymbol{e} + \beta & \\ \text{sujeto a} & \boldsymbol{v}^\intercal \boldsymbol{E}\boldsymbol{x}^k + \beta \leq \boldsymbol{c}^\intercal \boldsymbol{x}^k, & k \in K \\ & \boldsymbol{v}^\intercal \boldsymbol{E}\boldsymbol{d}^l \leq \boldsymbol{c}^\intercal \boldsymbol{d}^l, & l \in L. \end{array} \tag{DW-Ma^D}$$

El hecho de que z_{SP} sea el valor óptimo finito del subproblema (DW-SP) implica que, para todo $k \in K$, $z_{\mathrm{SP}} \leq \boldsymbol{c}^\intercal \boldsymbol{x}^k - \boldsymbol{w}^\intercal \boldsymbol{E}\boldsymbol{x}^k$ y, para todo $l \in L$, $\boldsymbol{c}^\intercal \boldsymbol{d}^l - \boldsymbol{w}^\intercal \boldsymbol{E}\boldsymbol{d}^l \geq 0$. Por tanto, $(\boldsymbol{w}, z_{\mathrm{SP}})$ es una solución factible del problema (DW-Ma^D) con función objetivo $\boldsymbol{w}^\intercal \boldsymbol{e} + z_{\mathrm{SP}}$. Por el Teorema de dualidad débil dicho valor es una cota inferior del valor óptimo del problema primal (DW-Ma).

A medida que el algoritmo progresa estas cotas se irán acercando, con la cota superior disminuyendo y la inferior aumentando. Es natural por tanto plantear criterios de parada basados en la diferencia, absoluta o relativa, entre estas dos cotas.

Generalizaciones a otras clases de problemas

Una de las razones que han hecho tan popular al algoritmo de Dantzig-Wolfe es que aprovecha al máximo toda la estructura del problema (DW-Orig). Esta misma característica hace que no sea inmediato extender el algoritmo a otras clases de problemas. Por ejemplo, su aplicación a

problemas donde algunas de las variables deben tomar valores enteros es especialmente problemática, pues la formulación del problema (DW-Ma) se basa en representar los puntos de $B = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{B}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$ mediante combinaciones convexas de sus puntos extremos. Cuando algunas o todas las variables han de tomar valores enteros se pierde la convexidad: ya no tenemos que las combinaciones convexas de puntos factibles sigan siendo puntos factibles.

Cabe mencionar que, desde este punto de vista, son mucho más flexibles el algoritmo de Benders, que veremos en la Sección 6.4, y los métodos de relajación lagrangiana, ya discutidos en la Sección 5.5.1, y que comentaremos de nuevo en la Sección 6.5.

6.3.5 Dantzig-Wolfe en problemas con estructura de bloques

Una de las claves para un buen rendimiento del algoritmo de Dantzig-Wolfe es que, en cada iteración, el subproblema (DW-SP) tenga una estructura que permita resolverlo eficientemente. La región factible de cada subproblema (DW-SP) es el conjunto B, que se obtiene al eliminar las restricciones complicantes del problema original (DW-Orig). Una de las situaciones más frecuentes en la práctica es aquella en la que la matriz $\boldsymbol{B}$ tiene una estructura de bloques como la de la Figura 6.1(a). En este apartado mostramos como esto permite descomponer el subproblema (DW-SP) en tantos subproblemas independientes como bloques tenga la matriz $\boldsymbol{B}$, lo que permitirá resolverlos más eficientemente. Además, el rendimiento computacional puede mejorar todavía más si disponemos de un entorno de computación que permita resolver los subproblemas en paralelo. En particular, nunca será necesario resolver un problema con todas las restricciones $\boldsymbol{B}\boldsymbol{x} = \boldsymbol{b}$. El problema maestro se encarga de las restricciones complicantes $\boldsymbol{E}\boldsymbol{x} = \boldsymbol{e}$, que típicamente son una pequeña proporción del total de restricciones, y cada subproblema se encarga de una de las componentes en las que se descompone $\boldsymbol{B}$.

Notación asociada a la estructura de bloques

Supongamos entonces que $\boldsymbol{B}$ puede descomponerse en n_{B} bloques, $\boldsymbol{B}_{[p]} \in \mathbb{R}^{m_p \times n_p}$, con $p \in \{1, \ldots, n_{\mathrm{B}}\}$. Tenemos las descomposiciones correspondientes para la matriz $\boldsymbol{E}$, $\boldsymbol{E}_{[p]} \in \mathbb{R}^{m_e \times n_p}$ y los vectores de variables, costes, y lados derechos: $\boldsymbol{x}_{[p]} \in \mathbb{R}^{n_p}$, $\boldsymbol{c}_{[p]} \in \mathbb{R}^{n_p}$, $\boldsymbol{b}_{[p]} \in \mathbb{R}^{m_p}$. Con esta estructura de bloques, el problema (DW-Orig) puede formularse como

$$
\begin{array}{rlcccccccl}
\underset{\boldsymbol{x}_{[1]},\ldots,\boldsymbol{x}_{[n_{\mathrm{B}}]}}{\text{minimizar}} & \boldsymbol{c}_{[1]}^{\mathsf{T}}\boldsymbol{x}_{[1]} & + & \boldsymbol{c}_{[1]}^{\mathsf{T}}\boldsymbol{x}_{[2]} & + & \ldots & + & \boldsymbol{c}_{[n_{\mathrm{B}}]}^{\mathsf{T}}\boldsymbol{x}_{[n_{\mathrm{B}}]} & & \\
\text{sujeto a} & \boldsymbol{E}_{[1]}\boldsymbol{x}_{[1]} & + & \boldsymbol{E}_{[2]}\boldsymbol{x}_{[2]} & + & \ldots & + & \boldsymbol{E}_{[n_{\mathrm{B}}]}\boldsymbol{x}_{[n_{\mathrm{B}}]} & = & \boldsymbol{e} \\
& \boldsymbol{B}_{[1]}\boldsymbol{x}_{[1]} & & & & & & & = & \boldsymbol{b}_{[1]} \\
& & & \boldsymbol{B}_{[2]}\boldsymbol{x}_{[2]} & & & & & = & \boldsymbol{b}_{[2]} \\
& & & & & \ddots & & & & \vdots \\
& & & & & & & \boldsymbol{B}_{[n_{\mathrm{B}}]}\boldsymbol{x}_{[n_{\mathrm{B}}]} & = & \boldsymbol{b}_{[n_{\mathrm{B}}]} \\
& & & & & & & \boldsymbol{x}_{[1]}, \boldsymbol{x}_{[2]}, \ldots, \boldsymbol{x}_{[n_{\mathrm{B}}]} & \geq & \boldsymbol{0},
\end{array}
$$

o, de manera más compacta, como

$$\begin{aligned} \underset{\boldsymbol{x}_{[1]},\ldots,\boldsymbol{x}_{[n_{\mathrm{B}}]}}{\text{minimizar}} \quad & \sum_{p=1}^{n_{\mathrm{B}}} \boldsymbol{c}_{[p]}^{\mathsf{T}} \boldsymbol{x}_{[p]} \\ \text{sujeto a} \quad & \sum_{p=1}^{n_{\mathrm{B}}} \boldsymbol{E}_{[p]} \boldsymbol{x}_{[p]} = \boldsymbol{e} \\ & \boldsymbol{B}_{[p]} \boldsymbol{x}_{[p]} = \boldsymbol{b}_{[p]}, \quad p \in \{1,\ldots,n_{\mathrm{B}}\} \\ & \boldsymbol{x}_{[p]} \geq \boldsymbol{0}, \quad p \in \{1,\ldots,n_{\mathrm{B}}\}. \end{aligned} \tag{DW-Orig$^{\mathrm{Desc}}$}$$

La estructura de bloques de la matriz $\boldsymbol{B}$ permite descomponer el conjunto $B = \{\boldsymbol{x} : \boldsymbol{B}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$ como el producto cartesiano $B = B_{[1]} \times \ldots \times B_{[n_{\mathrm{B}}]}$, donde $B_{[p]} = \{\boldsymbol{x}_{[p]} \in \mathbb{R}^{n_p} : \boldsymbol{B}_{[p]}\boldsymbol{x}_{[p]} = \boldsymbol{b}_{[p]}, \boldsymbol{x}_{[p]} \geq \boldsymbol{0}\}$. Una propiedad importante es que los puntos extremos de B son el producto cartesiano de puntos extremos de los conjuntos $B_{[p]}$. Además, es fácil ver que las direcciones extremas de B son el producto cartesiano de una dirección extrema de un conjunto $B_{[p]}$ y el vector cero para todas las demás componentes en las que se descompone el conjunto B. Para cada conjunto $B_{[p]}$, supondremos que sus puntos y direcciones extremas están indexados por $K_{[p]}$ y $L_{[p]}$, respectivamente. Además, cuando trabajemos con el problema maestro reducido, denotaremos los subconjuntos correspondientes como $K^{\mathrm{R}}_{[p]}$ y $L^{\mathrm{R}}_{[p]}$.

A continuación presentamos dos formulaciones del problema maestro que se aprovechan de manera ligeramente distinta de la estructura de bloques del problema original: la formulación agregada y la formulación desagregada.

Problema maestro: Formulación agregada

En su formulación agregada, el problema maestro es exactamente el mismo que el problema (DW-Ma). En este caso, la estructura de bloques no desempeña ningún papel en el problema maestro y se explotará en el subproblema, que podrá ser descompuesto en n_{B} subproblemas independientes. A continuación presentamos de nuevo la formulación del problema (DW-Ma), pero incluyendo de manera explícita la notación asociada a la estructura de bloques:

$$\begin{aligned} \underset{\boldsymbol{\lambda}\in\mathbb{R}^{|K|},\, \boldsymbol{\mu}\in\mathbb{R}^{|L|}}{\text{minimizar}} \quad & \sum_{k\in K} \lambda_k \sum_{p=1}^{n_{\mathrm{B}}} \boldsymbol{c}_{[p]}^{\mathsf{T}} \boldsymbol{x}_{[p]}^k + \sum_{l\in L} \mu_l \sum_{p=1}^{n_{\mathrm{B}}} \boldsymbol{c}_{[p]}^{\mathsf{T}} \boldsymbol{d}_{[p]}^l \\ \text{sujeto a} \quad & \sum_{k\in K} \lambda_k \sum_{p=1}^{n_{\mathrm{B}}} \boldsymbol{E}_{[p]} \boldsymbol{x}_{[p]}^k + \sum_{l\in L} \mu_l \sum_{p=1}^{n_{\mathrm{B}}} \boldsymbol{E}_{[p]} \boldsymbol{d}_{[p]}^l = \boldsymbol{e} \\ & \sum_{k\in K} \lambda_k = 1 \\ & \lambda_k \geq 0, \quad k \in K \\ & \mu_l \geq 0, \quad l \in L. \end{aligned} \tag{DW-Ma$^{\mathrm{Desc}_a}$}$$

Por otro lado, el subproblema (DW-SP) se puede reescribir como:

$$\begin{aligned} \underset{\boldsymbol{x}_{[1]},\ldots,\boldsymbol{x}_{[n_{\mathrm{B}}]}}{\text{minimizar}} \quad & \sum_{p=1}^{n_{\mathrm{B}}} \left(\boldsymbol{c}_{[p]}^{\mathsf{T}} \boldsymbol{x}_{[p]} - \boldsymbol{w}^{\mathsf{T}} \boldsymbol{E}_{[p]} \boldsymbol{x}_{[p]} \right) \\ \text{sujeto a} \quad & \boldsymbol{B}_{[p]} \boldsymbol{x}_{[p]} = \boldsymbol{b}_{[p]}, \quad p \in \{1,\ldots,n_{\mathrm{B}}\} \\ & \boldsymbol{x}_{[p]} \geq \boldsymbol{0}, \quad p \in \{1,\ldots,n_{\mathrm{B}}\}. \end{aligned} \tag{DW-SP$^{\mathrm{Desc}}$}$$

Claramente, este problema se puede resolver mediante la resolución de n_{B} subproblemas independientes. Para cada $p \in \{1,\ldots,n_{\mathrm{B}}\}$ tenemos el subproblema:

$$\begin{aligned} \underset{\boldsymbol{x}_{[p]}}{\text{minimizar}} \quad & \boldsymbol{c}_{[p]}^{\mathsf{T}} \boldsymbol{x}_{[p]} - \boldsymbol{w}^{\mathsf{T}} \boldsymbol{E}_{[p]} \boldsymbol{x}_{[p]} \\ \text{sujeto a} \quad & \boldsymbol{B}_{[p]} \boldsymbol{x}_{[p]} = \boldsymbol{b}_{[p]} \\ & \boldsymbol{x}_{[p]} \geq \boldsymbol{0}. \end{aligned} \tag{DW-SP$_p^{\mathrm{Desc}}$}$$

Denotemos por z_p a la función objetivo óptima de este subproblema. En este caso, la forma de proceder del algoritmo es la siguiente:

- Si cada subproblema (DW-SP$_p^{\mathrm{Desc}}$) tiene solución óptima, dada por un punto extremo $\bar{\boldsymbol{x}}_{[p]} \in B_{[p]}$, entonces podemos concatenarlos en un punto extremo $\bar{\boldsymbol{x}} = (\bar{\boldsymbol{x}}_{[1]},\ldots,\bar{\boldsymbol{x}}_{[n_{\mathrm{B}}]}) \in B$. Tenemos que $z_{\mathrm{SP}} = \sum_{p=1}^{n_{\mathrm{B}}} z_p$. Si $z_{\mathrm{SP}} \geq \alpha$ estamos ante una solución óptima y, en otro caso, el punto extremo $\bar{\boldsymbol{x}}$ debe ser añadido al problema maestro reducido.
- Si un subproblema (DW-SP$_p^{\mathrm{Desc}}$) no tiene óptimo finito, entonces tendremos una dirección extrema $\boldsymbol{d}_{[p]}$ de $B_{[p]}$ que podremos utilizar para obtener una dirección extrema de B, $\boldsymbol{d} = (\boldsymbol{0},\ldots,\boldsymbol{0},\boldsymbol{d}_{[p]},\boldsymbol{0},\ldots,\boldsymbol{0})$, que debe ser añadida al problema maestro reducido.

Problema maestro: Formulación desagregada

Presentamos ahora una formulación alternativa del problema maestro que explota más directamente la estructura de bloques y que suele producir mejores resultados en la práctica:

$$\begin{aligned} \underset{\substack{\boldsymbol{\lambda}_{[p]} \in \mathbb{R}^{|K_{[p]}|},\, \boldsymbol{\mu}_{[p]} \in \mathbb{R}^{|L_{[p]}|} \\ p \in \{1,\ldots,n_{\mathrm{B}}\}}}{\text{minimizar}} \quad & \sum_{p=1}^{n_{\mathrm{B}}} \left(\sum_{k \in K_{[p]}} \lambda_{[p]k} \boldsymbol{c}_{[p]}^{\mathsf{T}} \boldsymbol{x}_{[p]}^{k} + \sum_{l \in L_{[p]}} \mu_{[p]l} \boldsymbol{c}_{[p]}^{\mathsf{T}} \boldsymbol{d}_{[p]}^{l} \right) \\ \text{sujeto a} \quad & \sum_{p=1}^{n_{\mathrm{B}}} \left(\sum_{k \in K_{[p]}} \lambda_{[p]k} \boldsymbol{E}_{[p]} \boldsymbol{x}_{[p]}^{k} + \sum_{l \in L_{[p]}} \mu_{[p]l} \boldsymbol{E}_{[p]} \boldsymbol{d}_{[p]}^{l} \right) = \boldsymbol{e} \\ & \sum_{k \in K_{[p]}} \lambda_{[p]k} = 1, \quad p \in \{1,\ldots,n_{\mathrm{B}}\} \\ & \lambda_{[p]k} \geq 0, \quad k \in K_{[p]},\ p \in \{1,\ldots,n_{\mathrm{B}}\} \\ & \mu_{[p]l} \geq 0, \quad l \in L_{[p]},\ p \in \{1,\ldots,n_{\mathrm{B}}\}. \end{aligned} \tag{DW-Ma$^{\mathrm{Desc}_d}$}$$

Esta formulación presenta dos diferencias principales con respecto a la formulación agregada:

- Las variables $\lambda_{[p]k}$ y $\mu_{[p]l}$, en vez de ser una por cada punto y dirección extrema de B, son ahora una por cada punto y dirección extrema de cada componente $B_{[p]}$ de B.

- Ahora hay n_{B} restricciones de convexidad, una por cada componente $B_{[p]}$ de B. En particular esto implica que, además de las variables duales $\boldsymbol{w} \in \mathbb{R}^{m_e}$, tendremos una variable dual α_p para cada $p \in \{1, \ldots, n_{\mathrm{B}}\}$.

- El problema maestro reducido se definirá ahora a partir de subconjuntos de los conjuntos de variables $\lambda_{[p]k}$ y $\mu_{[p]l}$. Para cada $p \in \{1, \ldots, n_{\mathrm{B}}\}$ tendremos $K^{\mathrm{R}}_{[p]} \subset K_{[p]}$ y $L^{\mathrm{R}}_{[p]} \subset L_{[p]}$.

La clave para que esta formulación sea equivalente a la anterior es la ya mencionada de que $B = B_{[1]} \times \ldots \times B_{[n_{\mathrm{B}}]}$, lo que implica que los puntos extremos de B se pueden expresar como el producto cartesiano de puntos extremos de los conjuntos $\boldsymbol{B}_{[p]}$ y que las direcciones extremas de B se puedan expresar completando con ceros las direcciones extremas de cada conjunto $\boldsymbol{B}_{[p]}$.

Los subproblemas tienen exactamente la misma formulación que en el caso agregado, dada por el subproblema (DW-SP$^{\mathrm{Desc}}$), que se puede descomponer en los subproblemas (DW-SP$^{\mathrm{Desc}}_p$). En este caso, la forma de proceder del algoritmo es la siguiente:

- Si el subproblema (DW-SP$^{\mathrm{Desc}}_p$) tiene un óptimo finito y $z_p < \alpha_p$, la variable $\lambda_{[p]k}$ asociada al punto extremo correspondiente de $\boldsymbol{B}_{[p]}$ debe ser añadida al problema maestro reducido.

- Si todo subproblema (DW-SP$^{\mathrm{Desc}}_p$) tiene óptimo finito y en todos ellos $\boldsymbol{z}_p \geq \alpha_p$, entonces el algoritmo termina. Tenemos una solución óptima del problema (DW-Ma$^{\mathrm{Desc}_d}$).

- Si un subproblema (DW-SP$^{\mathrm{Desc}}_p$) es no acotado, la variable $\mu_{[p]l}$ asociada a la dirección extrema correspondiente de $\boldsymbol{B}_{[p]}$ debe ser añadida al problema maestro reducido.

Desde el punto de vista práctico, la principal diferencia entre ambas formulaciones es que en la agregada es necesario resolver todos los subproblemas para encontrar un punto extremo que añadir al problema maestro reducido. Por el contrario, en la formulación desagregada, como las variables del problema maestro están desagregadas por bloques, la resolución de cada subproblema (DW-SP$^{\mathrm{Desc}}_p$) permite ya decidir si debe o no añadirse algún punto o dirección extrema del conjunto $\boldsymbol{B}_{[p]}$ asociado. En la agregada las distintas componentes de un punto extremo $\boldsymbol{x}^k$ están atadas por la variable λ_k. En la desagregada puntos extremos añadidos a las distintas componentes en distintas iteraciones pueden ser "combinados" entre sí gracias a las variables $\lambda_{[p]k}$. Esta flexibilidad adicional es la clave del mejor rendimiento de la formulación desagregada en la mayoría de las aplicaciones, y compensa el incremento en el número de variables y restricciones del problema (DW-Ma$^{\mathrm{Desc}_d}$).

Para terminar, comentar que las consideraciones de la Sección 6.3.4 relativas a la factibilidad del problema maestro reducido y al cálculo de direcciones extremas y cotas siguen siendo totalmente válidas en cualquiera de las dos reformulaciones que acabamos de presentar.

6.3.6 Algoritmo de Dantzig-Wolfe

En la Figura 6.4 presentamos un esquema detallado del algoritmo de Dantzig-Wolfe, en el que incluimos tanto el cálculo de cotas, como el uso de variables artificiales para garantizar la factibilidad en el problema maestro reducido, como la descomposición en bloques en base a la formulación desagregada. A continuación presentamos una serie de comentarios relativos a este esquema:

ALGORITMO DE DANTZIG-WOLFE

INICIALIZACIÓN

Elegir un parámetro de terminación $\varepsilon > 0$. Inicializamos las cotas, $LB = -\infty$ y $UB = +\infty$ e inicializamos $K^{\mathrm{R}} = \emptyset$ y $L^{\mathrm{R}} = \emptyset$.

PRIMERA FASE. PASO 1.1. RESOLUCIÓN DEL PROBLEMA (DW-Red$^{\mathrm{fac}}$)

Obtenemos, para cada $p \in \{1, \ldots, n_{\mathrm{B}}\}$, $\boldsymbol{\lambda}_{[p]} \in \mathbb{R}^{|K^{\mathrm{R}}_{[p]}|}$, $\boldsymbol{\mu}_{[p]} \in \mathbb{R}^{|L^{\mathrm{R}}_{[p]}|}$. Además, obtenemos variables artificiales $\bar{\boldsymbol{y}}^a \in \mathbb{R}^{m_e}$ y $\bar{z}^a \in \mathbb{R}$ y duales $\boldsymbol{w} \in \mathbb{R}^{m_e}$ y $\alpha \in \mathbb{R}^{n_{\mathrm{B}}}$.

- Si $\bar{\boldsymbol{y}}^a = \boldsymbol{0}$ y $\bar{z}^a = 0$, pasar a la SEGUNDA FASE.
- En otro caso, ir al PASO 1.2.

PRIMERA FASE. PASO 1.2. RESOLUCIÓN DEL SUBPROBLEMA (DW-SP$^{\mathrm{fac}}$)

Para cada $p \in \{1, \ldots, n_{\mathrm{B}}\}$, resolver el subproblema de factibilidad asociado al bloque p:

(i) Si tiene óptimo finito y $z_p < \alpha_p$, añadir a $K^{\mathrm{R}}_{[p]}$ el punto extremo correspondiente de $\boldsymbol{B}_{[p]}$.

(ii) Si es no acotado, $z_p = -\infty$, añadir a $L^{\mathrm{R}}_{[p]}$ la dirección extrema correspondiente de $\boldsymbol{B}_{[p]}$.

Si algún problema p es infactible o todo subproblema p tiene óptimo finito con $\boldsymbol{z}_p \geq \alpha_p$, FIN, el problema (DW-Orig) no tiene soluciones factibles. En otro caso, ir al PASO 1.1.

SEGUNDA FASE. PASO 2.1. RESOLUCIÓN DEL PROBLEMA (DW-Red)

Obtenemos, para cada $p \in \{1, \ldots, n_{\mathrm{B}}\}$, $\boldsymbol{\lambda}_{[p]} \in \mathbb{R}^{|K^{\mathrm{R}}_{[p]}|}$, $\boldsymbol{\mu}_{[p]} \in \mathbb{R}^{|L^{\mathrm{R}}_{[p]}|}$. Además, obtenemos variables duales $\boldsymbol{w} \in \mathbb{R}^{m_e}$ y $\alpha \in \mathbb{R}^{n_{\mathrm{B}}}$.

Actualizamos UB al valor óptimo del problema (DW-Red).

SEGUNDA FASE. PASO 2. RESOLUCIÓN DEL SUBPROBLEMA (DW-SP$^{\mathrm{Desc}}$)

Para cada $p \in \{1, \ldots, n_{\mathrm{B}}\}$, resolver el subproblema (DW-SP$^{\mathrm{Desc}}_p$):

(i) Si tiene óptimo finito y $z_p < \alpha_p$, añadir a $K^{\mathrm{R}}_{[p]}$ el punto extremo correspondiente de $\boldsymbol{B}_{[p]}$.

(ii) Si es no acotado, $z_p = -\infty$, añadir a $L^{\mathrm{R}}_{[p]}$ la dirección extrema correspondiente de $\boldsymbol{B}_{[p]}$.

Actualizamos $LB = \max\{LB, \boldsymbol{w}^{\intercal}\boldsymbol{e} + \sum_{p=1}^{n_{\mathrm{B}}} z_p\}$.

- Si todo subproblema (DW-SP$^{\mathrm{Desc}}_p$) tiene óptimo finito con $\boldsymbol{z}_p \geq \alpha_p$, FIN, tenemos una solución óptima del problema maestro.
- Si $\frac{UB-LB}{|LB|+\varepsilon} < \varepsilon$, FIN, el *gap* relativo es suficientemente pequeño.
- En otro caso, ir al PASO 2.1.

Figura 6.4: Esquema del algoritmo de Dantzig-Wolfe.

- Los problemas en los que el conjunto B no admite descomposición en bloques se corresponden con $n_{\mathrm{B}} = 1$.
- Si el conjunto $B = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{B}\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0}\}$ es acotado, entonces no tiene direcciones extremas y no son necesarias las variables $\mu_{[p]l}$.
- En caso de querer garantizar que el algoritmo encuentra la solución óptima basta con eliminar el criterio de parada del *gap* relativo.
- La convergencia del algoritmo está garantizada por el hecho de que los conjuntos $B_{[p]}$ tienen una cantidad finita de puntos y direcciones extremas. Un punto o dirección extrema que ya ha sido añadido nunca será candidato a ser añadido de nuevo, pues el coste reducido de una variable presente en el problema (DW-Red) será menor o igual a cero.
- En la práctica podemos encontrarnos con variables que sólo afectan a las restricciones complicantes, con lo que habrá algún subproblema p que sólo tendrá restricciones de no negatividad, ya que la matriz $\boldsymbol{B}_{[p]}$ asociada no tiene ninguna fila. Un ejemplo lo tenemos en el siguiente apartado, donde para pasar el problema (PTM-RC) a forma estándar es necesario añadir variables de holgura a las restricciones complicantes. Variables que no afectarán a ninguna restricción de la matriz $\boldsymbol{B}$.
- En la práctica las implementaciones incorporan un número máximo de iteraciones.

6.3.7 Ejemplo ilustrativo: Problema de transporte (PTM-RC)

Veamos ahora cómo resolver el problema de transporte multimercancía con restricciones complicantes visto en la Sección 6.1.2 con el algoritmo de Dantzig-Wolfe. Comenzamos recordando la formulación matemática del problema (PTM-RC), pero añadiendo las variables de holgura necesarias para ponerlo en forma estándar, con todas las restricciones de igualdad:

PTM CON RESTRICCIONES COMPLICANTES

$$
\begin{aligned}
\text{minimizar} \quad & \sum_{m\in M}\sum_{i\in O}\sum_{j\in D} c_{ijm} f_{ijm} & & \\
\text{sujeto a} \quad & \sum_{j\in D} f_{ijm} = s_{im} & i \in O,\ m \in M & \\
& \sum_{i\in O} f_{ijm} = d_{jm} & j \in D,\ m \in M & \qquad (\text{PTM-RC}^{\text{est}}) \\
& \sum_{m\in M} p_m f_{ijm} + y^s_{ij} = p^{\max}_{ij} & i \in O,\ j \in D \quad \text{(Restrics. complicantes)} & \\
& f_{ijm} \geq 0 & i \in O,\ j \in D,\ m \in M & \\
& y^s_{ij} \geq 0 & i \in O,\ j \in D. &
\end{aligned}
$$

Con respecto a la formulación del problema (DW-Orig), tenemos que la matriz $\boldsymbol{E}$ está formada por $|O| \cdot |D|$ restricciones complicantes. El vector $\boldsymbol{e}$ viene definido por los pesos máximos $p^{\max}_{ij}$.

La matriz $\boldsymbol{B}$ está formada por $(|O| + |D|) \cdot |M|$ restricciones, y puede ser dividida en $|M|$ bloques, con cada bloque $\boldsymbol{B}_{[m]}$ formado por $|O| + |D|$ restricciones.

Es importante destacar que, como comentamos al final del apartado anterior, la presencia de las variables de holgura en las restricciones complicantes hace que tengamos un bloque $\boldsymbol{B}_{[m+1]}$ sin ninguna restricción asociada. Tendríamos por tanto que $B = \{\boldsymbol{f} \in \mathbb{R}^{O\times D\times M}, \boldsymbol{y}^s \in \mathbb{R}^M : \boldsymbol{B}\boldsymbol{f} = \boldsymbol{b},\ \boldsymbol{f} \geq \boldsymbol{0},\ \boldsymbol{y}^s \geq \boldsymbol{0}\}$ y la aplicación del algoritmo de Dantzig-Wolfe sería inmediata.

6.4 Métodos de Generación de filas. Variables complicantes y algoritmo de Benders

El algoritmo de Benders (Benders, 1962) fue diseñado para resolver problemas con variables complicantes, típicamente con una estructura de bloques como la de la Figura 6.1(b). A diferencia del algoritmo de Danztig-Wolfe, que no es fácil de extender más allá de los problemas de programación lineal, el algoritmo de Benders ya surgió para resolver problemas de programación lineal y entera, y sus ideas fueron rápidamente generalizadas para problemas no lineales Geoffrion (1972). Esta flexibilidad del algoritmo de Benders es la principal razón de que haya sido aplicado exitosamente en multitud de problemas reales en distintos campos, y también de que su popularidad siga creciendo año a año, como muestra la Figura 6.5. El trabajo Rahmaniani y otros (2017) contiene una detallada revisión de aplicaciones de esta técnica, así como de distintas variantes que han sido estudiadas para mejorar el rendimiento de la versión original, que es la que presentamos a continuación.

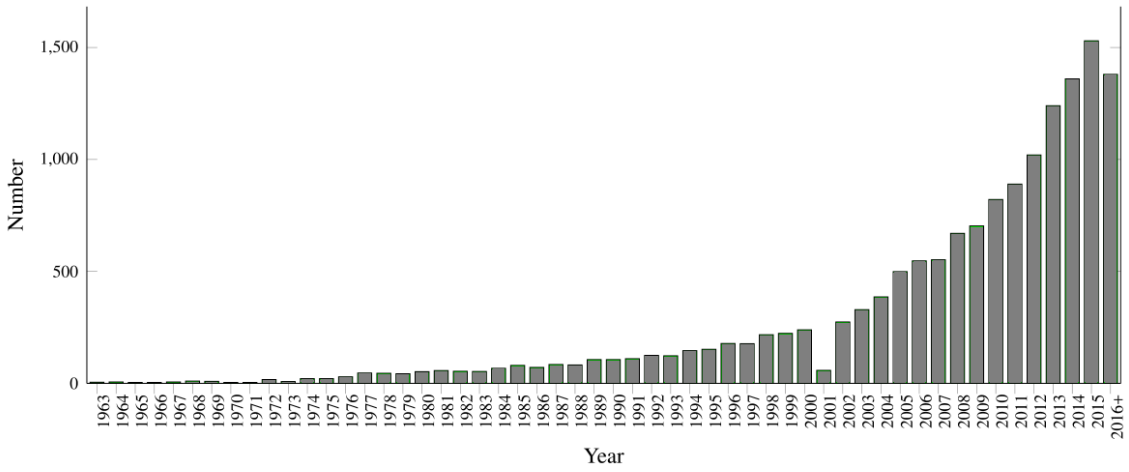


Figura 6.5: Menciones anuales en trabajos científicos a la descomposición de Benders según https://scholar.google.com/. Figura tomada del trabajo Rahmaniani y otros (2017).

Supongamos que tenemos el siguiente problema de programación lineal entera:

$$\begin{array}{rl} \underset{\boldsymbol{y}\in\mathbb{R}^{n_y},\,\boldsymbol{x}\in\mathbb{R}^{n_x}}{\text{minimizar}} & \boldsymbol{f}^\intercal\boldsymbol{y}+\boldsymbol{c}^\intercal\boldsymbol{x} \\ \text{sujeto a} & \boldsymbol{A}\boldsymbol{y}=\boldsymbol{b} \\ & \boldsymbol{E}\boldsymbol{y}+\boldsymbol{B}\boldsymbol{x}=\boldsymbol{e} \\ & \boldsymbol{x}\geq\boldsymbol{0}, \\ & \boldsymbol{y}\geq\boldsymbol{0},\ \boldsymbol{y}\in\mathbb{Z}^{n_y} \end{array} \tag{BD-Orig}$$

donde $\boldsymbol{y}\in\mathbb{Z}^{n_y}$ son las variables complicantes, que han de tomar valores enteros, $\boldsymbol{x}\in\mathbb{R}^{n_x}$, $\boldsymbol{f}\in\mathbb{R}^{n_y}$, $\boldsymbol{c}\in\mathbb{R}^{n_x}$, $\boldsymbol{A}\in\mathbb{R}^{m_a\times n_y}$, $\boldsymbol{b}\in\mathbb{R}^{m_a}$, $\boldsymbol{E}\in\mathbb{R}^{m_e\times n_y}$, $\boldsymbol{B}\in\mathbb{R}^{m_e\times n_x}$ y $\boldsymbol{e}\in\mathbb{R}^{m_e}$. Nuevamente no hay pérdida de generalidad asumir que el problema (BD-Orig) está en forma estándar. Las restricciones $\boldsymbol{A}\boldsymbol{y}=\boldsymbol{b}$ afectan únicamente a las variables complicantes y en muchas aplicaciones ni siquiera están presentes. Por otro lado, si fijamos las variables complicantes a un valor $\bar{\boldsymbol{y}}$ tal que $\boldsymbol{A}\bar{\boldsymbol{y}}=\boldsymbol{b}$, nos queda el siguiente problema:

$$\begin{array}{rl} \underset{\boldsymbol{x}\in\mathbb{R}^{n_x}}{\text{minimizar}} & \boldsymbol{c}^\intercal\boldsymbol{x} \\ \text{sujeto a} & \boldsymbol{B}\boldsymbol{x}=\boldsymbol{e}-\boldsymbol{E}\bar{\boldsymbol{y}} \\ & \boldsymbol{x}\geq\boldsymbol{0}, \end{array} \tag{BD-SP$_{\bar{\boldsymbol{y}}}$}$$

donde es habitual que la matriz $\boldsymbol{B}$ pueda ser descompuesta por **B**loques, lo que a su vez implica que el subproblema (BD-SP$_{\bar{\boldsymbol{y}}}$) podría resolverse mediante la resolución de tantos subproblemas independientes como bloques tenga $\boldsymbol{B}$. El subproblema (BD-SP$_{\bar{\boldsymbol{y}}}$) es un problema de optimización lineal continua y su dual viene dado por

$$\begin{array}{rl} \underset{\boldsymbol{v}\in\mathbb{R}^{m_e}}{\text{maximizar}} & (\boldsymbol{e}-\boldsymbol{E}\bar{\boldsymbol{y}})^\intercal\boldsymbol{v} \\ \text{sujeto a} & \boldsymbol{B}^\intercal\boldsymbol{v}\leq\boldsymbol{c}. \end{array} \tag{BD-SP$^D_{\bar{\boldsymbol{y}}}$}$$

6.4.1 Reformulación del problema: Problema maestro y subproblemas

Si definimos $Y=\{\boldsymbol{y}\in\mathbb{Z}^{n_y}:\boldsymbol{A}\boldsymbol{y}=\boldsymbol{b},\,\boldsymbol{y}\geq\boldsymbol{0}\}$ entonces, a la vista del subproblema (BD-SP$_{\bar{\boldsymbol{y}}}$), podemos escribir el problema (BD-Orig) como

$$\min_{\bar{\boldsymbol{y}}\in Y}\left\{\boldsymbol{f}^\intercal\bar{\boldsymbol{y}}+\min_{\boldsymbol{x}\in\mathbb{R}^{n_x}}\left\{\boldsymbol{c}^\intercal\boldsymbol{x}:\boldsymbol{B}\boldsymbol{x}=\boldsymbol{e}-\boldsymbol{E}\bar{\boldsymbol{y}},\,\boldsymbol{x}\geq\boldsymbol{0}\right\}\right\}.$$

EL Teorema de dualidad fuerte garantiza que la función objetivo en el óptimo coincide con la del dual, con lo que el problema anterior también es equivalente a

$$\min_{\bar{\boldsymbol{y}}\in Y}\left\{\boldsymbol{f}^\intercal\bar{\boldsymbol{y}}+\max_{\boldsymbol{v}\in\mathbb{R}^{m_e}}\left\{(\boldsymbol{e}-\boldsymbol{E}\bar{\boldsymbol{y}})^\intercal\boldsymbol{v}:\boldsymbol{B}^\intercal\boldsymbol{v}\leq\boldsymbol{c}\right\}\right\}.$$

Una de las claves del algoritmo de Benders es que la región factible del subproblema (BD-SP$^D_{\bar{\boldsymbol{y}}}$), $F=\{\boldsymbol{v}\in\mathbb{R}^{m_e}:\boldsymbol{B}^\intercal\boldsymbol{v}\leq\boldsymbol{c}\}$, es independiente de la elección de $\bar{\boldsymbol{y}}$. Sabemos por el Teorema de Carathéodory que el conjunto F se puede caracterizar mediante sus puntos y direcciones extremas, propiedad que también explota el algoritmo de Benders, aunque de manera distinta a la del algoritmo de Dantzig-Wolfe. Denotemos por $\boldsymbol{v}^k$, con $k\in K$, a los puntos extremos de F y por $\boldsymbol{d}^l$, con $l\in L$, a sus direcciones extremas.

Si el subproblema (BD-SP$^D_{\bar{\boldsymbol{y}}}$) es infactible, $F = \emptyset$, entonces hay dos posibilidades para el subproblema (BD-SP$_{\bar{\boldsymbol{y}}}$): i) no tiene óptimo finito para algún $\bar{\boldsymbol{y}} \in Y$, en cuyo caso el problema (BD-Orig) tampoco tiene óptimo finito (es no acotado) o ii) es infactible para todo $\bar{\boldsymbol{y}} \in Y$, en cuyo caso el problema (BD-Orig) no tiene soluciones factibles.

Supongamos ahora que $F \neq \emptyset$ y que para un cierto $\bar{\boldsymbol{y}} \in Y$ existe una dirección extrema de F, $\boldsymbol{d}^l$, tal que $(\boldsymbol{e} - \boldsymbol{E}\bar{\boldsymbol{y}})^\intercal \boldsymbol{d}^l > 0$. El apartado (II) de la Proposición 6.2 implica que el subproblema (BD-SP$^D_{\bar{\boldsymbol{y}}}$) es no acotado y, por tanto, el subproblema (BD-SP$_{\bar{\boldsymbol{y}}}$) no tiene soluciones factibles. Esto sugiere que podemos añadir $(\boldsymbol{e} - \boldsymbol{E}\boldsymbol{y})^\intercal \boldsymbol{d}^l > 0$ como restricción en el problema (BD-Orig), ya que si un vector $\bar{\boldsymbol{y}}$ incumple esta restricción, entonces el subproblema (BD-SP$_{\bar{\boldsymbol{y}}}$) es infactible. Si añadimos las restricciones asociadas a todas las direcciones extremas de F, entonces tenemos garantizado que el subproblema (BD-SP$^D_{\bar{\boldsymbol{y}}}$) nunca será un problema no acotado, con lo que tendrá un óptimo finito que se alcanzará en un punto extremo. Esto resulta en la siguiente reformulación del problema (BD-Orig):

$$\begin{array}{rl} \min\limits_{\bar{\boldsymbol{y}} \in Y} & \boldsymbol{f}^\intercal \bar{\boldsymbol{y}} + \max\limits_{k \in K} \{(\boldsymbol{e} - \boldsymbol{E}\bar{\boldsymbol{y}})^\intercal \boldsymbol{v}^k\} \\ \text{sujeto a} & (\boldsymbol{e} - \boldsymbol{E}\bar{\boldsymbol{y}})^\intercal \boldsymbol{d}^l \leq 0, \quad l \in L. \end{array} \tag{BD-Compacto}$$

Por último, una vez que el máximo se define sobre una cantidad finita de puntos, podemos recurrir a una transformación/"truco" habitual que consiste en pasar este tipo de máximos a las restricciones con la ayuda de una variable auxiliar, que llamaremos η, obteniendo la siguiente reformulación del problema (BD-Orig) que será equivalente siempre y cuando $F \neq \emptyset$:

$$\begin{array}{rl} \underset{\boldsymbol{y} \in \mathbb{R}^{n_y},\ \eta \in \mathbb{R}}{\text{minimizar}} & \boldsymbol{f}^\intercal \boldsymbol{y} + \eta \\ \text{sujeto a} & \boldsymbol{A}\boldsymbol{y} = \boldsymbol{b} \\ & (\boldsymbol{e} - \boldsymbol{E}\boldsymbol{y})^\intercal \boldsymbol{v}^k \leq \eta, \quad k \in K \\ & (\boldsymbol{e} - \boldsymbol{E}\boldsymbol{y})^\intercal \boldsymbol{d}^l \leq 0, \quad l \in L \\ & \boldsymbol{y} \geq \boldsymbol{0},\ \boldsymbol{y} \in \mathbb{Z}^{n_y}, \end{array} \tag{BD-Ma}$$

conocida como problema maestro de Benders. Al igual que pasaba con el problema maestro de Dantzig-Wolfe (DW-Ma), esta reformulación en sí misma no simplifica la resolución del problema original. En este caso, tenemos que el problema (BD-Ma) tiene $m_a + |K| + |L|$ restricciones y, como ya hemos comentado, tanto $|K|$ como $|L|$ suelen ser tan grandes que incluso enumerar explícitamente todas las restricciones puede ser algo computacionalmente prohibitivo. En su lugar, el algoritmo de Benders parte del problema maestro reducido, que definimos a continuación, y en el que las restricciones asociadas a puntos y direcciones extremas se van añadiendo a medida que van siendo identificadas mediante la resolución del subproblema (BD-SP$^D_{\bar{\boldsymbol{y}}}$).

$$\begin{array}{rl} \underset{\boldsymbol{y} \in \mathbb{R}^{n_y},\ \eta \in \mathbb{R}}{\text{minimizar}} & \boldsymbol{f}^\intercal \boldsymbol{y} + \eta \\ \text{sujeto a} & \boldsymbol{A}\boldsymbol{y} = \boldsymbol{b} \\ & \boldsymbol{y} \geq \boldsymbol{0},\ \boldsymbol{y} \in \mathbb{Z}^{n_y}. \end{array} \tag{BD-Red}$$

Ya estamos en condiciones de presentar el esquema general del algoritmo de Dantzig-Wolfe:

(i) Definir y resolver el problema (BD-Red).

(ii) Resolver el subproblema (BD-SP$^D_{\bar{\boldsymbol{y}}}$):

- Si el subproblema tiene un óptimo finito $\boldsymbol{v}^k$, con $k \in K$. Entonces, la restricción $(\boldsymbol{e} - \boldsymbol{E}\boldsymbol{y})^\intercal \boldsymbol{v}^k \leq \eta$, llamada *corte de optimalidad*, se añade al problema (BD-Red) y se resuelve nuevamente.
- Si el subproblema es no acotado tendremos una dirección extrema $\boldsymbol{d}^l$, con $l \in L$. Entonces, la restricción $(\boldsymbol{e} - \boldsymbol{E}\boldsymbol{y})^\intercal \boldsymbol{d}^l \leq 0$, llamada *corte de factibilidad*, se añade al problema (BD-Red) y se resuelve nuevamente.
- Si el subproblema es infactible el algoritmo termina. Hay que determinar si el problema (BD-Orig) es no acotado o infactible.

Nuevamente, la convergencia en una cantidad finita de iteraciones del algoritmo resultante se apoya en el hecho de que el conjunto F tiene una cantidad finita de puntos y direcciones extremas. Antes de presentar formalmente el esquema completo del algoritmo de Benders, que veremos en la Figura 6.6 de la Sección 6.4.4, hay una serie de aspectos prácticos de gran relevancia que debemos discutir y a los que dedicamos las secciones 6.4.2 y 6.4.3.

6.4.2 Consideraciones adicionales relativas al algoritmo

Problema (BD-Red) no acotado

Aunque el problema (BD-Orig) tenga óptimo finito, durante las primeras iteraciones del algoritmo de Benders el problema (BD-Red) puede ser no acotado. Por ejemplo, en su formulación inicial sin ningún corte de optimalidad no hay nada que impida hacer tender η a $-\infty$. Para evitar esto se puede conseguir un conjunto inicial de cortes resolviendo varias veces el subproblema (BD-SP$^D_{\bar{\boldsymbol{y}}}$) (variando $\bar{\boldsymbol{y}}$). Esto es similar a la primera fase de Dantzig-Wolfe para conseguir la factibilidad del problema (DW-Red). De hecho, una situación es dual a la otra.

La alternativa que aquí presentamos consiste en introducir en el problema (BD-Red) cotas auxiliares para aquellas variables que no las tengan y, a medida que progresa el algoritmo, detectar cuando estas cotas ya no son necesarias y eliminarlas de la formulación. Más concretamente, trabajaremos con el siguiente problema:

$$\begin{array}{ll} \underset{\boldsymbol{y}\in\mathbb{R}^{n_y},\ \eta\in\mathbb{R}}{\text{minimizar}} & \boldsymbol{f}^\intercal \boldsymbol{y} + \eta \\ \text{sujeto a} & \boldsymbol{A}\boldsymbol{y} = \boldsymbol{b} \\ & \boldsymbol{y} \geq \boldsymbol{0},\ \boldsymbol{y} \in \mathbb{Z}^{n_y} \\ & \boldsymbol{y} \leq M,\ \eta \geq -M, \qquad \text{(cotas auxiliares)} \end{array} \tag{BD-Red$^{\text{aux}}$}$$

donde, idealmente, $M > 0$ deberá tomar un valor suficientemente grande como para no dejar fuera ninguna solución factible del problema (BD-Orig). Si el algoritmo termina con alguna variable alcanzando el valor M, debería incrementarse dicho valor y continuar con el algoritmo. Eventualmente deberíamos poder eliminar estas cotas o concluir que el problema (BD-Orig) es no acotado.

Cálculo de direcciones extremas

El cálculo de una dirección extrema asociada a un problema (BD-SP$^D_{\bar{\boldsymbol{y}}}$) no acotado es similar al discutido en la Sección 6.3.4. La mayoría de los optimizadores lineales actuales ya devolverían

esa dirección extrema al intentar resolver el subproblema (BD-SP$^D_{\bar{\boldsymbol{y}}}$), pero siempre podríamos identificarla mediante el procedimiento descrito en la Sección 6.2.3.

Una diferencia con respecto al algoritmo de Dantzig-Wolfe es que, si bien para este decíamos que rara vez eran necesarias pues el conjunto B normalmente es acotado, en el algoritmo de Benders la situación es la contraria: casi siempre son necesarias. La razón es que siempre que elijamos una solución $\bar{\boldsymbol{y}}$ del problema (BD-Red) que dé lugar a un subproblema (BD-SP$_{\bar{\boldsymbol{y}}}$) infactible, entonces el problema (BD-SP$^D_{\bar{\boldsymbol{y}}}$) será no acotado (salvo que $F = \emptyset$).

Cálculo de cotas

El algoritmo de Benders también genera, iteración a iteración, cotas superiores e inferiores del valor óptimo de la función objetivo.

Cota inferior. El problema (BD-Red) es una relajación del problema (BD-Orig). Por tanto, dada una solución $\bar{\boldsymbol{y}}$, $\bar{\eta}$, $\boldsymbol{f}^\intercal\bar{\boldsymbol{y}}+\bar{\eta}$ nos da una cota inferior. En el caso de estar usando cotas auxiliares, lo será si ninguna variable alcanza el valor de su cota auxiliar.

Cota superior. Dado $\bar{\boldsymbol{y}}$ obtenido al resolver el problema (BD-Red) y un punto extremo $\boldsymbol{v}^k$ de F obtenido al resolver el problema (BD-SP$^D_{\bar{\boldsymbol{y}}}$), la formulación (BD-Compacto) nos asegura que $\boldsymbol{f}^\intercal\bar{\boldsymbol{y}} + (\boldsymbol{e} - \boldsymbol{E}\bar{\boldsymbol{y}})^\intercal\boldsymbol{v}^k$ es una cota superior. Otra forma de verlo es considerar la solución óptima $\bar{\boldsymbol{x}}$ del problema (BD-SP$_{\bar{\boldsymbol{y}}}$) asociada a $\boldsymbol{v}^k$. Entonces, $(\bar{\boldsymbol{y}}, \bar{\boldsymbol{x}})$ es una solución factible del problema (BD-Orig) y, por el Teorema de dualidad fuerte, el valor de la función objetivo es $\boldsymbol{f}^\intercal\bar{\boldsymbol{y}} + (\boldsymbol{e} - \boldsymbol{E}\bar{\boldsymbol{y}})^\intercal\boldsymbol{v}^k$, y por tanto tenemos una cota superior.

Repetición de puntos y direcciones extremas

Al igual que en el algoritmo de Dantzig-Wolfe, en el algoritmo de Benders no es necesario preocuparse por la posibilidad de que en un subproblema se obtenga un punto extremo o dirección factible cuyo corte asociado ya haya sido añadido al problema (BD-Red):

Repetición de puntos extremos. Denotemos por z_{BD} a la función objetivo óptima del problema (BD-Orig). Supongamos que en una cierto paso del algoritmo de Benders tenemos que, asociado a la solución $\bar{\boldsymbol{y}}$ del problema (BD-Red) obtenemos un punto extremo $\boldsymbol{v}^k$ cuyo corte de factibilidad asociado ya había sido añadido previamente. En este caso, por la resolución del subproblema la cota superior: $\boldsymbol{f}^\intercal\bar{\boldsymbol{y}} + (\boldsymbol{e} - \boldsymbol{E}\bar{\boldsymbol{y}})^\intercal\boldsymbol{v}^k \geq z_{\text{BD}}$. Además, la función objetivo asociada a $\bar{\boldsymbol{y}}$, $\boldsymbol{f}^\intercal\bar{\boldsymbol{y}} + \bar{\eta}$, es una cota inferior de z_{BD} lo cual, combinado con que el corte $(\boldsymbol{e} - \boldsymbol{E}\boldsymbol{y})^\intercal\boldsymbol{v}^k \leq \eta$ ya presente en el problema (BD-Red) resulta en

$$z_{\text{BD}} \overset{\text{Cota inferior}}{\geq} \boldsymbol{f}^\intercal\bar{\boldsymbol{y}} + \bar{\eta} \overset{\text{Corte}}{\geq} \boldsymbol{f}^\intercal\bar{\boldsymbol{y}} + (\boldsymbol{e} - \boldsymbol{E}\bar{\boldsymbol{y}})^\intercal\boldsymbol{v}^k \overset{\text{Cota superior}}{\geq} z_{\text{BD}}.$$

Por tanto, si en algún momento hubiese que repetir un punto extremo tendríamos que ya se habría cerrado el *gap* entre las dos cotas.

Repetición de direcciones extremas. Supongamos que el corte asociado a una dirección extrema $\boldsymbol{d}^l$, $(\boldsymbol{e} - \boldsymbol{E}\boldsymbol{y})^\intercal\boldsymbol{d}^l \leq 0$ está presente en el problema (BD-Red) y que la solución de este es $\bar{\boldsymbol{y}}$. Si el subproblema (BD-SP$^D_{\bar{\boldsymbol{y}}}$) es no acotado con $\boldsymbol{d}^l$ como dirección extrema asociada tendríamos, por el apartado (II) de la Proposición 6.2, que $(\boldsymbol{e} - \boldsymbol{E}\bar{\boldsymbol{y}})^\intercal\boldsymbol{d}^l > 0$, contradiciendo la factibilidad de $\bar{\boldsymbol{y}}$ en el problema (BD-Red).

Dificultad de resolución del problema (BD-Red)

Una diferencia importante entre los algoritmos de Dantzig-Wolfe y Benders radica en la dificultad de resolución del problema maestro reducido. En el algoritmo de Benders estamos ante un problema de programación lineal y entera, cuya resolución se complica a medida que vamos añadiendo restricciones. En el algoritmo de Benders, también se complica a medida que añadimos variables, pero nunca deja de ser un problema de programación lineal. Debido a esto, en la práctica es habitual encontrarse problemas donde la resolución del problema maestro de Benders es el gran cuello de botella del algoritmo.

Una línea de ataque bastante habitual para mitigar este problema es apoyarse en el hecho de que, una vez el algoritmo esté bastante avanzado, típicamente sucede que una gran parte de los cortes no están activos en las nuevas soluciones del problema (BD-Red). Por tanto, parece natural diseñar estrategias que hagan "limpieza" de los cortes y eliminen aquellos que se consideren irrelevantes. Por supuesto, no hay garantías de que no vuelvan a aparecer de nuevo más adelante, con lo que hay que tener mucho cuidado al diseñar estas estrategias por ejemplo, para evitar ciclados del algoritmo. El trabajo Rahmaniani y otros (2017) presenta de manera esquemática las referencias bibliográficas asociadas a esta y otras muchas estrategias para tratar de mejorar el rendimiento del algoritmo de Benders.

6.4.3 Benders en problemas con estructura de bloques

Nuevamente, una de las claves para un buen rendimiento del algoritmo de Benders es que el subproblema (BD-SP$^D_{\bar{\boldsymbol{y}}}$) se pueda resolver eficientemente. La situación más habitual es aquella en la que la matriz $\boldsymbol{B}$ tiene una estructura de bloques, propiedad que por tanto también tendrá la matriz $\boldsymbol{B}^\intercal$, y que permitirá la descomposición del problema (BD-SP$^D_{\bar{\boldsymbol{y}}}$) en subproblemas independientes. La notación que usaremos será la análoga a la introducida en la Sección 6.4.3 para el algoritmo de Dantzig-Wolfe, con cada bloque $\boldsymbol{B}_{[p]} \in \mathbb{R}^{m_p \times n_p}$.

En este caso, como la estructura de descomposición afectará a las variables $\boldsymbol{x}$ y estas no están presentes en el problema maestro, únicamente necesitamos discutir la reformulación de los subproblemas, que presentamos a continuación:

$$\begin{aligned} \underset{\boldsymbol{x}_{[1]},\ldots,\boldsymbol{x}_{[n_\mathrm{B}]}}{\text{minimizar}} \quad & \sum_{p=1}^{n_\mathrm{B}} \boldsymbol{c}_{[p]}^\intercal \boldsymbol{x}_{[p]} \\ \text{sujeto a} \quad & \boldsymbol{B}_{[p]}\boldsymbol{x}_{[p]} = (\boldsymbol{e} - \boldsymbol{E}\bar{\boldsymbol{y}})_{[p]}, \quad p \in \{1,\ldots,n_\mathrm{B}\} \\ & \boldsymbol{x}_{[p]} \geq \boldsymbol{0}, \quad p \in \{1,\ldots,n_\mathrm{B}\}, \end{aligned} \tag{BD-SP$^{\mathrm{Desc}}_{\bar{\boldsymbol{y}}}$}$$

Este problema se puede descomponer en n_B subproblemas independientes, obteniendo, para cada $p \in \{1,\ldots,n_\mathrm{B}\}$:

$$\begin{aligned} \underset{\boldsymbol{x}_{[p]} \in \mathbb{R}^{n_p}}{\text{minimizar}} \quad & \boldsymbol{c}_{[p]}^\intercal \boldsymbol{x}_{[p]} \\ \text{sujeto a} \quad & \boldsymbol{B}_{[p]}\boldsymbol{x}_{[p]} = (\boldsymbol{e} - \boldsymbol{E}\bar{\boldsymbol{y}})_{[p]} \\ & \boldsymbol{x}_{[p]} \geq \boldsymbol{0}. \end{aligned} \tag{BD-SP$^{\mathrm{Desc}}_{\bar{\boldsymbol{y}},p}$}$$

Para cada uno de estos problemas podemos escribir el dual correspondiente:

$$\begin{array}{rl} \underset{\boldsymbol{v}_{[p]}\in\mathbb{R}^{m_p}}{\text{maximizar}} & (\boldsymbol{e}-\boldsymbol{E}\bar{\boldsymbol{y}})_{[p]}^{\mathsf{T}}\boldsymbol{v}_{[p]} \\ \text{sujeto a} & \boldsymbol{B}_{[p]}^{\mathsf{T}}\boldsymbol{v}_{[p]}\leq \boldsymbol{c}_{[p]}, \end{array} \qquad (\text{BD-SP}_{\bar{\boldsymbol{y}},p}^{D,\text{Desc}})$$

Denotemos por z_p a la función objetivo óptima de este subproblema. En este caso, la forma de proceder del algoritmo es la siguiente:

- Si cada subproblema (BD-SP$_{\bar{\boldsymbol{y}},p}^{D,\text{Desc}}$) tiene solución óptima, dada por un punto extremo $\bar{\boldsymbol{v}}_{[p]}\in F_{[p]}$, entonces podemos concatenarlos en un punto extremo $\bar{\boldsymbol{v}}=(\bar{\boldsymbol{v}}_{[1]},\ldots,\bar{\boldsymbol{v}}_{[n_{\text{B}}]})\in F$ y definir el corte de optimalidad correspondiente.

- En otro caso, por cada subproblema (BD-SP$_{\bar{\boldsymbol{y}},p}^{D,\text{Desc}}$) sin óptimo finito, tendremos una dirección extrema $\boldsymbol{d}_{[p]}$ de $F_{[p]}$ que podremos utilizar para obtener una dirección extrema de F, $\boldsymbol{d}=(\boldsymbol{0},\ldots,\boldsymbol{0},\boldsymbol{d}_{[p]},\boldsymbol{0},\ldots,\boldsymbol{0})$, que podremos usar para añadir el corte de factibilidad correspondiente. Esto puede resultar en varios cortes de factibilidad en la misma iteración.

6.4.4 Algoritmo de Benders

En la Figura 6.6 presentamos un esquema detallado del algoritmo de Benders, en el que incluimos tanto el cálculo de cotas, como el uso de cotas auxiliares para garantizar que el problema maestro reducido tenga óptimo finito, como la descomposición en bloques. A continuación presentamos una serie de comentarios relativos a este esquema:

- Si el algoritmo termina con una solución óptima en la que alguna cota auxiliar está activa, será indicativo de que el problema (BD-Orig) no tiene óptimo finito. Para mayor seguridad, podría resolverse de nuevo el problema incrementando el valor de M.

- Si el conjunto F no admite descomposición en bloques basta tomar $n_{\text{B}}=1$.

- En caso de querer garantizar que el algoritmo encuentra la solución óptima basta con reemplazar el criterio de parada con el *gap* relativo por la condición $UB=LB$.

- Como ya argumentamos en la Sección 6.4.2, la convergencia del algoritmo está garantizada por la cantidad finita de puntos y direcciones extremas del conjunto F, y que ningún punto o dirección extrema cuyo corte ya haya sido añadido deberá añadirse de nuevo.

- En la práctica las implementaciones incorporan un número máximo de iteraciones.

6.4.5 Ejemplo ilustrativo: Problema de transporte (PTM-VC)

Si consideramos el problema de transporte multimercancía con variables complicantes visto en la Sección 6.1.2, la estructura asociada a la descomposición de Benders es muy fácil de

ALGORITMO DE BENDERS

INICIALIZACIÓN

Elegir un parámetro de terminación $\varepsilon > 0$, $\gamma > 1$, $M > 0$ y $M^{\text{máx}}$. Definir el problema maestro reducido con cotas auxiliares. Inicializamos las cotas, $LB = -\infty$ y $UB = +\infty$.

PASO 1. RESOLUCIÓN PROBLEMA MAESTRO REDUCIDO CON COTAS AUXILIARES

Obtenemos $\bar{\boldsymbol{y}} \in \mathbb{R}^{n_y}$ y $\bar{\eta} \in \mathbb{R}$.

- Si $\eta > -M$ e $y_i < M$ para todo $i \in \{1, \ldots, n_y\}$, eliminar las cotas auxiliares del problema maestro reducido.
- Si las cotas auxiliares ya han sido eliminadas, actualizamos $LB = \boldsymbol{f}^{\intercal}\bar{\boldsymbol{y}} + \bar{\eta}$.

PASO 2. RESOLUCIÓN DEL SUBPROBLEMA

Para cada $p \in \{1, \ldots, n_{\text{B}}\}$, resolver el subproblema (BD-SP$^{D,\text{Desc}}_{\bar{\boldsymbol{y}},p}$):

(i) Si un subproblema (BD-SP$^{D,\text{Desc}}_{\bar{\boldsymbol{y}},p}$) no tiene soluciones factibles, entonces $F = \emptyset$ y hay que determinar si el problema (BD-Orig) es no acotado o infactible.

(ii) Si cada subproblema (BD-SP$^{D,\text{Desc}}_{\bar{\boldsymbol{y}},p}$) tiene un punto extremo óptimo $\bar{\boldsymbol{v}}_{[p]}$, definir el punto extremo asociado de F, $\bar{\boldsymbol{v}}$, y añadir al problema maestro reducido el corte de optimalidad correspondiente.

(iii) En otro caso, por cada subproblema (BD-SP$^{D,\text{Desc}}_{\bar{\boldsymbol{y}},p}$) no acotado, identificar la dirección extrema de F asociada y añadir al problema maestro reducido el corte de factibilidad correspondiente.

Actualizamos $UB = \min\{UB, \boldsymbol{f}^{\intercal}\bar{\boldsymbol{y}} + (\boldsymbol{e} - \boldsymbol{E}\bar{\boldsymbol{y}})^{\intercal}\bar{\boldsymbol{v}}\}$. Si $\frac{UB-LB}{|LB|+\varepsilon} < \varepsilon$, entonces

- Si las cotas auxiliares han sido eliminadas, FIN, alcanzado *gap* relativo deseado.
- Si alguna cota auxiliar está activa, definir $M = \gamma M$ y $LB = -\infty$.

Y ahora,

- Si $M < M^{\text{máx}}$ volver al PASO 1.
- Si $M \geq M^{\text{máx}}$, FIN, puede ser que el problema sea no acotado.

Figura 6.6: Esquema del algoritmo de Benders.

identificar. Recordemos primero la formulación del problema:

PTM CON VARIABLES COMPLICANTES

$$
\begin{array}{lll}
\text{minimizar} & \displaystyle\sum_{i\in M} c^{\text{f}}_i y_i + \sum_{m\in M}\sum_{i\in O}\sum_{j\in D} c_{ijm} f_{ijm} & \\
\text{sujeto a} & -s_{im} y_i + \displaystyle\sum_{j\in D} f_{ijm} \leq 0 & i \in O,\ m \in M \\
& \displaystyle\sum_{i\in O} f_{ijm} = d_{jm} & j \in D,\ m \in M \\
& y_i \in \{0,1\} & i \in O \quad \text{(Variables complicantes)} \\
& f_{ijm} \geq 0 & i \in O,\ j \in D,\ m \in M.
\end{array}
\tag{PTM-VC}
$$

Con respecto a la formulación del problema (BD-Orig) no tenemos restricciones de la forma $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{b}$. Además, las matrices $\boldsymbol{E}$ y $\boldsymbol{B}$ tienen $(|O| + |D|) \cdot |M|$ filas, con n_y y n_x columnas, respectivamente. Además, las primeras $|O| \times |M|$ filas de $\boldsymbol{E}$ son nulas. El vector $\boldsymbol{f}$ viene dado

por los costes $c_i^{\mathrm{f}} y_i$ y el vector $\boldsymbol{c}$ por los costes c_{ijm}. Nuevamente, la matriz $\boldsymbol{B}$ puede ser dividida en $|M|$ bloques, con cada bloque $\boldsymbol{B}_{[m]}$ formado por $|O| + |D|$ restricciones. Una vez hecha esta identificación, la aplicación del algoritmo de Benders es inmediata.

Terminamos con un comentario adicional relativo a lo conveniente que resulta la descomposición en bloques en este problema. Fijado un valor $\bar{\boldsymbol{y}}$ para las variables complicantes, tenemos que el subproblema (BD-SP$_{\bar{\boldsymbol{y}}}$) se puede dividir en $|M|$ subproblemas independientes. Cada uno de estos se corresponde con un problema del transporte, en el cual la unimodularidad asegura que, aunque los flujos representen objetos indivisibles, todas las soluciones básicas factibles de cada subproblema tendrán todas las componentes enteras, y podemos tratar igualmente los subproblemas como problemas continuos. Sin embargo, en el caso de intentar resolver directamente el problema (PTM-VC), podríamos encontrarnos con ramificaciones innecesarias en las variables f_{ijm}. Esto es porque, en optimalidad, una vez que las variables $\bar{\boldsymbol{y}}$ tomen valores enteros, la unimodularidad asegurará que también lo hacen las variables f_{ijm}.

6.5 Generalizaciones a otras clases de problemas

Una de las técnicas más versátiles a la hora de intentar descomponer problemas es la relajación lagrangiana, siendo Everett (1963) uno de los trabajos pioneros en esta técnica, aunque debe su nombre actual al trabajo Geoffrion (1974). La idea de la relajación lagrangiana, que ya vimos en el problema de planificación energética de la Sección 5.5.1, es trabajar con un problema dual lagrangiano apropiadamente elegido, "dualizando" algunas de las restricciones de modo que el subproblema dual lagrangiano resultante se pueda descomponer en subproblemas independientes. Bajo convexidad del problema original esta técnica puede resultar muy efectiva, pues el Teorema de dualidad fuerte (Teorema 5.19) garantiza la equivalencia entre resolver el problema primal y el dual. En caso de que no haya convexidad, sigue siendo una técnica muy usada en la práctica, tanto como base de algoritmos heurísticos como para obtener cotas para el óptimo del problema original.

- En el caso de problemas no lineales con restricciones complicantes, la generalización más habitual del algoritmo de Dantzig-Wolfe es justamente la relajación lagrangiana. De hecho, en el caso lineal, Dantzig-Wolfe se puede reinterpretar como un algoritmo de relajación lagrangiana. Una versión más sofisticada de la técnica de relajación lagrangiana es el método de lagrangiano aumentado que veremos en la Sección 7.3 y cuyos primeros desarrollos fueron llevados a cabo en Hestenes (1969) y Powell (1969). Este método tiene propiedades de convergencia notablemente superiores, pero a cambio es más difícil usarlo como algoritmo de descomposición que se aproveche de estructuras de bloques en los subproblemas.

- En el caso de problemas con variables enteras y restricciones complicantes, la generalización de Dantzig-Wolfe se conoce como Branch and Price, nombre acuñado a partir de los trabajos unificadores de Vanderbeck y Wolsey (1996), Barnhart y otros (1998) y Vanderbeck (2000, 2011), pero con las ideas de generación de columnas subyacentes presentes ya desde los trabajos pioneros de Gilmore y Gomory (1961, 1963). Para este tipo de problemas la relajación lagrangiana también ha sido usada con éxito desde los trabajos

de Held y Karp (1970, 1971), en parte gracias al formalismo y resultados en Geoffrion (1974).

- En el caso de los problemas con variables complicantes, las generalizaciones del algoritmo de Benders se conocen precisamente como Benders generalizado, pues su estructura es muy similar a la del algoritmo original. Uno de los trabajos pioneros en esta dirección es Geoffrion (1972).

Tema 7

Optimización con restricciones. Algoritmos

Contenidos

7.1 Introducción

El objetivo principal de este tema es presentar algunos algoritmos para la resolución de problemas generales de programación matemática. Es decir, problemas de la forma

$$\begin{array}{rll} \text{minimizar} & f(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 & i = 1, \dots, m \\ & h_j(\boldsymbol{x}) = 0 & j = 1, \dots, l. \end{array}$$

Más concretamente, presentaremos algoritmos cuyo objetivo es identificar puntos que cumplan alguna condición de optimalidad, típicamente las condiciones necesarias de Karush-Kuhn-Tucker. Como ya sucedía en el caso de la optimización sin restricciones, también aquí nos encontramos con una condición únicamente local y en la práctica es habitual complementar este tipo de algoritmos con otras metodologías que intenten aumentar las posibilidades de identificar puntos con buenas probabilidades globales.

Es importante destacar que, a diferencia del tema en el que desarrollamos algoritmos de optimización sin restricciones, donde podríamos decir que hemos hecho una buena cobertura de los principales algoritmos de optimización en el caso diferenciable, aquí simplemente presentamos algunos ejemplos representativos. El abanico de algoritmos para problemas con restricciones es mucho más amplio, pudiendo aplicarse enfoques muy distintos entre sí, dependiendo de cómo gestionan las restricciones, del uso que se hace del dual, de si son algoritmos de punto interior o no,... Una cobertura exhaustiva de estos métodos excede los contenidos de este curso.

Al igual que hicimos en el Tema 4 al estudiar algoritmos de optimización sin restricciones, aquí nos centraremos especialmente en las ideas e intuiciones detrás de los métodos, omitiendo las demostraciones.

7.2 Métodos de penalización clásicos

Esta sección estará dedicada en gran medida a los métodos de *penalización exterior*, aunque en la parte final hablaremos brevemente de otra familia de algoritmos relacionada: los métodos de *penalización interior* o de *barrera*.

7.2.1 Definición y propiedades

Los métodos de penalización exterior tienen una idea muy natural: resolver los problemas de optimización con restricciones apoyándose en los algoritmos existentes para resolver problemas sin restricciones. Este objetivo se consigue "subiendo" todas las restricciones a la función objetivo, a través de una función que penalice las infactibilidades. Más concretamente, un método de penalización exterior se basa en elegir una sucesión de parámetros de penalización $\{\rho^t\}_{t\in\mathbb{N}} \subset \mathbb{R}^+$ tal que $\lim_{t\to\infty} \rho^t = \infty$ y resolver subproblemas de la forma:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) + \rho^t P(\boldsymbol{x}),$$

donde la función $P(\boldsymbol{x})$ se conoce como *función de penalización* y debe cumplir lo siguiente:

(i) Para todo $\boldsymbol{x} \in \mathbb{R}^n$, $P(\boldsymbol{x}) \geq 0$.

(ii) $P(\boldsymbol{x}) = 0$ si y sólo si $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$ y $\boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}$.

(iii) $P(\boldsymbol{x})$ es continua.

La idea del método es muy sencilla, cuanto más grande sea ρ^t más peso tendrá la función $P(\boldsymbol{x})$ en el subproblema de minimización y más difícil será que un punto no factible en el problema original pueda ser solución de este subproblema.[1] A continuación presentamos dos elecciones habituales para la función $P(\boldsymbol{x})$:

Penalización absoluta. $P^a(\boldsymbol{x}) = \sum_{i=1}^m \max\{0, g_i(\boldsymbol{x})\} + \sum_{j=1}^l |h_j(\boldsymbol{x})|$.

Penalización cuadrática. $P^c(\boldsymbol{x}) = \|\max\{0, \boldsymbol{g}(\boldsymbol{x})\}\|^2 + \|\boldsymbol{h}(\boldsymbol{x})\|^2 = \sum_{i=1}^m (\max\{0, g_i(\boldsymbol{x})\})^2 + \sum_{j=1}^l h_j(\boldsymbol{x})^2$.

En la práctica es habitual subir a la función objetivo únicamente las restricciones más difíciles, llegando a subproblemas penalizados que también son problemas de optimización con restricciones, pero mucho más sencillos de resolver que el problema original dada la naturaleza de las restricciones. En dicho caso podemos pensar que estamos resolviendo problemas de la forma:

$$\begin{array}{rll} \text{minimizar} & f(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 & i = 1, \ldots, m \\ & h_j(\boldsymbol{x}) = 0 & j = 1, \ldots, l \\ & \boldsymbol{x} \in S, & \end{array} \tag{7.1}$$

donde S es un conjunto definido a partir de las restricciones "fáciles" (por ejemplo restricciones lineales, que darían lugar a un conjunto S convexo). Entonces, tendríamos los subproblemas

$$\min_{\boldsymbol{x} \in S} f(\boldsymbol{x}) + \rho^t P(\boldsymbol{x}),$$

donde la función P penaliza únicamente las infactibilidades originadas en las funciones $\boldsymbol{h}$ y $\boldsymbol{g}$. También es importante destacar que, en general, la sucesión de soluciones de los subproblemas penalizados, $\{\boldsymbol{x}^t\}_{t \in \mathbb{N}}$, no tiene por qué estar bien definida ya que, por ejemplo, podemos encontrarnos que el subproblema de la iteración t es no acotado y no tiene óptimo finito. En la práctica este problema puede atajarse fácilmente cuando las funciones f, $\boldsymbol{h}$ y $\boldsymbol{g}$ son continuas sin más que asumir que el conjunto S es no vacío, cerrado y acotado (pues entonces el Teorema de Weierstrass nos asegura la existencia de un óptimo). Este supuesto no suele ser muy restrictivo, ya que en la práctica para la mayoría de los problemas se pueden definir fácilmente unas restricciones de cota para las variables $\boldsymbol{x} \in \mathbb{R}^n$ y restringir el problema de optimización al hipercubo resultante de la imposición de las mismas.

En la Figura 7.1 presentamos el esquema general de un algoritmo de penalización exterior. Una característica de este tipo de algoritmos, a diferencia de la mayoría de los que hemos visto hasta ahora, es que, desde el punto de vista teórico, los iterantes de la sucesión $\{\boldsymbol{x}^t\}_{t \in \mathbb{N}}$ son independientes entre sí, ya que los problemas a resolver dependen únicamente del parámetro de penalización. Sin embargo, en la práctica es habitual que el algoritmo usado para resolver

[1] En la práctica es habitual que la sucesión $\{\rho^t\}$ sea una sucesión de vectores, con una penalización específica para cada restricción o grupo de restricciones (para así poder tener en cuenta de forma explícita las escalas en las que se mueven las distintas restricciones).

MÉTODO DE PENALIZACIÓN EXTERIOR

INICIALIZACIÓN

Elegir la función de penalización $P(\boldsymbol{x})$ y un parámetro de terminación $\varepsilon > 0$. Elegir un parámetro inicial $\rho^1 > 0$ y un escalar $\gamma > 1$. Elegir un punto inicial $\boldsymbol{x}^1$. Definir $t = 1$.

PASO 1

Definir $\boldsymbol{x}^{t+1}$ como una solución óptima del problema $\min_{\boldsymbol{x}\in S} f(\boldsymbol{x}) + \rho^t P(\boldsymbol{x})$ (donde $\boldsymbol{x}^t$ se toma como solución inicial para resolver el problema).

- Si $\rho^t P(\boldsymbol{x}) < \varepsilon$, FIN. Devolvemos $\boldsymbol{x}^{t+1}$.
- En otro caso, definir $\rho^{t+1} = \gamma\rho^t$. Reemplazar t por $t+1$. Repetir el PASO 1.

Figura 7.1: Esquema del método de penalización exterior.

el problema $\min_{\boldsymbol{x}\in S} f(\boldsymbol{x}) + \rho^t P(\boldsymbol{x})$ y obtener el iterante $\boldsymbol{x}^{t+1}$ tome $\boldsymbol{x}^t$ como solución inicial. Como comentaremos más adelante, esto se debe a que una de las mayores limitaciones de los métodos de penalización es la dificultad de resolución de los subproblemas, especialmente a medida que se va incrementando el valor de ρ^t. Por tanto, es útil que el algoritmo que resuelve los subproblemas defina sus puntos iniciales a partir de las soluciones obtenidas en los pasos anteriores.

A continuación presentamos el principal resultado que sirve de justificación teórica de los métodos de penalización y que, tras la discusión previa, no debería resultar sorprendente.

Teorema 7.1. *Supongamos que las funciones del problema de optimización dado por la Ecuación (7.1) son continuas y que la sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$, construida al aplicarle el método de penalización exterior, está contenida en un subconjunto compacto de S. Supongamos además que, para todo $t > 1$, $\boldsymbol{x}^t$ es un óptimo global del subproblema de la iteración $t-1$. Entonces:*

(i) *Todo punto límite de la sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ es un óptimo global del problema $\min_{\boldsymbol{x}\in S} P(\boldsymbol{x})$.*

(ii) *Además, si el problema original tiene algún punto factible, entonces todo punto límite de $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ es un óptimo global de dicho problema.*

Nótese que I) nos dice que si nuestro problema es infactible los puntos límite serán aquellos que minimicen el nivel de infactibilidad, medido mediante la función $P(\boldsymbol{x})$. Por otro lado, II) nos dice que si el problema tiene una región factible no vacía, los puntos límite serán óptimos globales. La demostración de este resultado es justamente lo que se pide en el siguiente ejercicio.

•••**Ejercicio 7.1.** Demuestra el Teorema 7.1. ◁

A pesar de que el Teorema 7.1 puede verse como una justificación teórica muy fuerte para los métodos de penalización exterior, la implementación práctica de estos métodos tiene bastantes problemas que detallamos a continuación:

- En general es muy difícil resolver globalmente los subproblemas penalizados. Como vimos en los temas anteriores, salvo que estemos trabajando con funciones convexas, los algoritmos de optimización sin restricciones simplemente aseguran convergencia a puntos que cumplen alguna condición necesaria de optimalidad local (típicamente que se anule el gradiente).

- Incluso en el caso de que los puntos de la sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ cumplan alguna condición necesaria de optimalidad local para los subproblemas, el Teorema 7.1 no asegura sus puntos límite también también la cumplan para el problema original.

- Otro problema importante es que el óptimo del problema original puede no alcanzarse para ningún valor finito de ρ^t, lo que genera a su vez bastantes problemas:

  – No es fácil saber con antelación cuál es la sucesión $\{\rho^t\}_{t\in\mathbb{N}}$ más adecuada. En el caso del algoritmo tal y como lo presentamos en la Figura 7.1, esto se traduce en que no es fácil saber cuál es el valor adecuado para γ.

  – Cuanto mayor sea el parámetro ρ^t más difícil será resolver el subproblema penalizado correspondiente. Intuitivamente, lo que pasa cuando la penalización es muy grande es que la función objetivo de los subproblemas pone muy poco peso en la función objetivo del problema original, con lo que se vuelve más difícil la tarea de resolver dichos problemas globalmente (y puede haber cada vez más óptimos locales que únicamente "se preocupan" de la factibilidad). Siendo más precisos, valores altos de ρ^t dan lugar a problemas mal condicionados (mucha diferencia entre los autovalores de la matriz hessiana). En el Tema 4 (ver página 75), comentamos que que esto ocasiona varias dificultades prácticas: i) es fácil encontrarse con problemas de precisión numérica, lo que puede desestabilizar a los algoritmos y ii) aunque tuviéramos precisión infinita, los algoritmos de optimización sin restricciones pueden converger muy lentamente (vimos que el método de máximo descenso es especialmente sensible a esta problemática) y iii) dicha convergencia sólo suele estar garantizada a puntos donde se anula el gradiente, que pueden no ser buenos puntos si tenemos subproblemas con muchos óptimos locales. En el Ejemplo 7.1 que viene a continuación ilustramos este problema de condicionamiento.

Ejemplo 7.1. Consideremos el siguiente problema de optimización con restricciones:

$$\begin{array}{rl} \text{minimizar} & \frac{1}{2}x_1^2 + x_2^2 \\ \text{sujeto a} & x_1 - 1 = 0. \end{array}$$

Claramente, el único óptimo global de este problema $\bar{\boldsymbol{x}} = (1, 0)$. Veamos qué pasa si le aplicamos directamente el método de penalización exterior con penalización cuadrática, $P^c(\boldsymbol{x}) = (x_1-1)^2$. En cada iteración t tendremos que resolver el problema

$$\underset{\boldsymbol{x}\in\mathbb{R}^2}{\text{minimizar}} \;\; f_P^t(\boldsymbol{x}) = \frac{1}{2}x_1^2 + x_2^2 + \rho^t(x_1 - 1)^2.$$

Podemos comprobar fácilmente que estamos ante una función convexa (es suma de funciones convexas) y por tanto $\nabla f_P^t(\boldsymbol{x}) = \mathbf{0}$ es una condición necesaria y suficiente de optimalidad global (Corolario 2.5). El gradiente y la matriz hessiana de la función $f_P(\boldsymbol{x})$ vienen dados por

$$\nabla f_P^t(\boldsymbol{x}) = (x_1 + 2\rho^t(x_1 - 1), 2x_2) \quad \text{y} \quad \boldsymbol{H}_P^t(\boldsymbol{x}) = \begin{pmatrix} 1 + 2\rho^t & 0 \\ 0 & 2 \end{pmatrix}.$$

La ecuación $\nabla f_P^t(\boldsymbol{x}) = \mathbf{0}$ tiene una única solución, que es $x_1^{t+1} = \frac{2\rho^t}{1+2\rho^t}$ y $x_2^{t+1} = 0$. Vemos que, consistentemente con el Teorema 7.1, esta sucesión tiende al único óptimo global, $\bar{\boldsymbol{x}} = (1, 0)$. Sin embargo, este valor no se alcanza para ningún valor finito de la penalización ρ^t.

Además, si miramos la matriz hessiana, como es una matriz diagonal tenemos que sus autovalores son precisamente los elementos de la diagonal y, por tanto, el número de condicionamiento de la matriz $\boldsymbol{H}_P^t(\bar{\boldsymbol{x}})$ es

$$\kappa(\boldsymbol{H}_P^t(\bar{\boldsymbol{x}})) = \frac{1 + 2\rho^t}{2},$$

que tiende a infinito con ρ^t. Este mal condicionamiento es habitual en los métodos de penalización exterior y es el que ocasiona que, en la práctica, tengan muchas veces un mal comportamiento. ◇

Como hemos visto, el Ejemplo 7.1 también muestra que el óptimo $\bar{\boldsymbol{x}} = (1, 0)$ no se alcanza para ningún valor finito de ρ^t cuando se utiliza la penalización cuadrática. A continuación damos una intuición adicional para este resultado. Supongamos que estamos en el punto $\bar{\boldsymbol{x}} = (1, 0)$. En este punto, $\nabla f(\bar{\boldsymbol{x}}) = (1, 0)$. Entonces, para minimizar la función f nos gustaría movernos en la dirección $-\nabla f(\bar{\boldsymbol{x}}) = -(1, 0)$, lo que supone salir de la región factible. Supongamos ahora que estamos con el problema penalizado para un cierto valor ρ^t y veamos el valor de $f_P^t(\boldsymbol{x}_\varepsilon)$ con $\varepsilon > 0$ y $\boldsymbol{x}_\varepsilon = \bar{\boldsymbol{x}} - \varepsilon(1, 0) = (1 - \varepsilon, 0)$. Entonces tenemos que $f_P^t(\boldsymbol{x}_\varepsilon) = f(\boldsymbol{x}_\varepsilon) + \rho^t\varepsilon^2$. Ahora, utilizando la aproximación de Taylor de primer orden de f en $\bar{\boldsymbol{x}}$ tenemos que

$$f(\boldsymbol{x}_\varepsilon) = f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\intercal(\boldsymbol{x}_\varepsilon - \bar{\boldsymbol{x}}) + \varphi(\varepsilon) = f(\bar{\boldsymbol{x}}) + (1, 0)^\intercal(-\varepsilon, 0) + \varphi(\varepsilon) = f(\bar{\boldsymbol{x}}) - \varepsilon + \varphi(\varepsilon),$$

donde $\lim_{\varepsilon\to 0} \varphi(\varepsilon) = 0$. Por tanto, para $f_P^t(\boldsymbol{x}_\varepsilon)$ tenemos

$$f_P^t(\boldsymbol{x}_\varepsilon) = f(\boldsymbol{x}_\varepsilon) + \rho^t\varepsilon^2 = f(\bar{\boldsymbol{x}}) - \varepsilon + \rho^t\varepsilon^2 + \varphi(\varepsilon).$$

Esta expresión nos permite ver por qué $\bar{\boldsymbol{x}}$ no es un óptimo de f_P^t para ningún valor de t. No importa lo grande que sea ρ^t, una vez que está fijado tendremos que, para $\varepsilon > 0$ suficientemente pequeño, el término $-\varepsilon$ dominará al término $\rho^t\varepsilon^2$. Por tanto, para $\varepsilon > 0$ suficientemente pequeño, $f_P^t(\boldsymbol{x}_\varepsilon) < f_P^t(\bar{\boldsymbol{x}})$, con lo que $\bar{\boldsymbol{x}}$ no es un óptimo del problema penalizado. Este argumento que acabamos de presentar se puede formalizar de modo bastante general y, lo que es más importante, la situación que describe se suele presentar en la práctica.[2] Una excepción es el caso en el que $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$, en el que no hay ninguna dirección de descenso desde $\bar{\boldsymbol{x}}$.

Lo que acabamos de ver sugiere que la clave de la necesidad de hacer tender ρ^t a infinito para acercarnos al óptimo del problema original radica en tomar penalizaciones cuadráticas, que conlleva que $\rho^t\varepsilon^2$ acaba siendo dominado por $-\varepsilon$. Por tanto, podríamos preguntarnos si este efecto seguiría presente trabajando con penalizaciones absolutas, P^a, en vez de penalizaciones cuadráticas, P^c. En el ejemplo en cuestión, tomando penalizaciones absolutas tendríamos algo de la forma $\rho^t|\varepsilon|$ que dominará a $-\varepsilon$ siempre que $\rho^t > 1$. El siguiente resultado muestra que esto será cierto de forma general.

[2] Para formalizarlo de modo más general habría que trabajar también con los gradientes de las restricciones en el punto $\bar{\boldsymbol{x}}$.

Teorema 7.2. *Existe $\bar{\rho}$ tal que el óptimo global del problema*

$$\min_{\boldsymbol{x}\in S} f(\boldsymbol{x}) + \rho P^a(\boldsymbol{x}),$$

es el óptimo global del problema original para todo $\rho > \bar{\rho}$.

Además, se puede demostrar que es suficiente con que $\bar{\rho}$ sea mayor que el mayor, en valor absoluto, de los multiplicadores de Lagrange asociados al punto $\bar{\boldsymbol{x}}$. Desafortunadamente, en la práctica este resultado no es de tanta utilidad como podría parecer. Las razones por las cuales la penalización $P^a(\boldsymbol{x})$ rara vez se utiliza en este tipo de algoritmos son las siguientes:

- De antemano no podemos conocer los multiplicadores de Lagrange en el óptimo buscado $\bar{\boldsymbol{x}}$, con lo que igualmente habrá que trabajar con una sucesión creciente de valores ρ^t y hay cierto riesgo de caer en problemas de condicionamiento (aunque menor que con penalización cuadrática).

- El uso de las penalizaciones absolutas tiene el importante problema de que las funciones objetivo de los subproblemas no son diferenciables. Además, los puntos de no diferenciabilidad de dichas funciones son precisamente los puntos factibles en el caso de $|\boldsymbol{h}(\boldsymbol{x})|$ y puntos de la frontera en el caso de $\max\{\mathbf{0}, \boldsymbol{g}(\boldsymbol{x})\}$. Esto hace que la resolución de los subproblemas se complique considerablemente.

- En general no se pueden encontrar funciones que siendo diferenciables aseguren convergencia finita del método de penalización. Intuitivamente, la no diferenciabilidad se debe a que, dado que la función $P(\boldsymbol{x})$ es constante en la región factible, para asegurar la diferenciabilidad de $P(\boldsymbol{x})$ necesitamos que su gradiente sea cero al llegar a la región factible. En otras palabras, cuanto más cerca está un punto de ser factible más suave es la penalización $P(\boldsymbol{x})$, en el sentido de estar más cerca de cero y de tener crecimiento nulo, por lo que más grande tendrá que ser la penalización para compensar dicha suavidad en la penalización.

En la siguiente sección veremos el método de lagrangiano aumentado que, en cierto modo, es una versión sofisticada de los métodos de penalización clásicos y que consigue justamente lo que no se puede conseguir con ellos: subproblemas penalizados diferenciables cuyos óptimos convergen a puntos KKT del problema original sin necesidad de que la penalización tenga que tender a infinito. Antes de pasar a dicha sección presentamos un último apartado dedicado a otra familia relevante de métodos de penalización: los métodos de penalización interior o métodos de barrera.

7.2.2 Métodos de penalización interior o métodos de barrera

En este apartado presentamos brevemente otro tipo de algoritmos que guardan cierta similitud con los métodos de penalización exterior, pues también trabajan con una penalización en la función objetivo. Sin embargo, en vez de crear una sucesión de puntos que se aproximan al óptimo desde el exterior del conjunto factible, los métodos de penalización interior trabajan en

todo momento con puntos factibles. Más concretamente, estos métodos penalizan únicamente las restricciones de desigualdad en problemas de la forma

$$\begin{array}{rll} \text{minimizar} & f(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 & i = 1, \ldots, m \\ & \boldsymbol{x} \in S, & \end{array}$$

y la idea es definir una penalización que actúe como barrera para que los puntos definidos por el algoritmo no salgan de la región factible. En caso de que el problema original tenga restricciones de igualdad estas formarán parte de la definición del conjunto S, con lo que estos métodos no son muy efectivos con problemas que tienen restricciones de igualdad difíciles (ya que no permiten subirlas a la función objetivo).

Un método de penalización interior o de barrera se basa en elegir una sucesión de parámetros $\{\mu^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^+$ tal que $\lim_{k\to\infty} \mu^k = 0$ y resolver subproblemas de la forma:

$$\begin{array}{rll} \text{mín} & f(\boldsymbol{x}) + \mu^k B(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) < 0 & i = 1, \ldots, m \\ & \boldsymbol{x} \in S, & \end{array} \tag{7.2}$$

donde la función $B(\boldsymbol{x})$ se conoce como *función de barrera* y debe cumplir lo siguiente:

(i) Para todo $\boldsymbol{x} \in \mathbb{R}^n$, $B(\boldsymbol{x}) \geq 0$.

(ii) $B(\boldsymbol{x})$ tiende a $+\infty$ a medida que $\boldsymbol{x}$ se acerca a la frontera del conjunto $\{\boldsymbol{x} : \boldsymbol{g}(\boldsymbol{x}) \leq 0\}$ (desde dentro del mismo). Es decir, si para algún $i \in \{1, \ldots, m\}$ tenemos que $g_i(\boldsymbol{x})$ tiende a cero tomando valores negativos, entonces $B(\boldsymbol{x})$ tenderá a infinito.

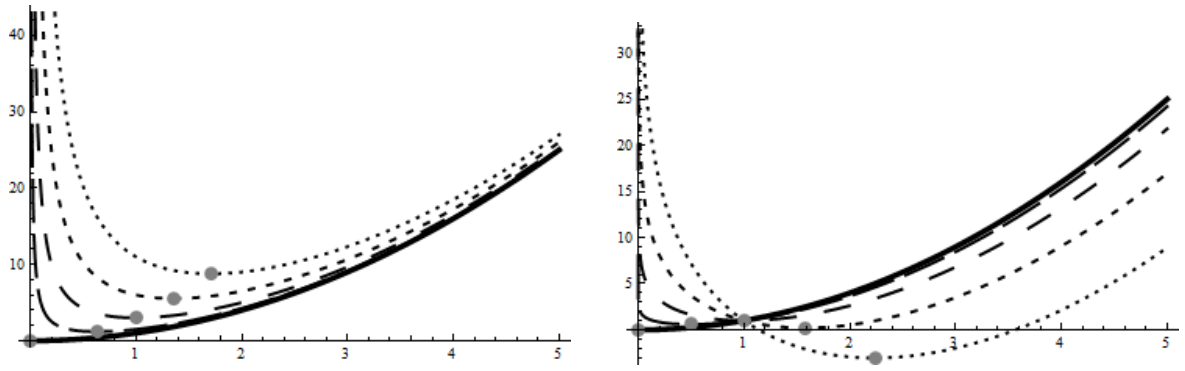
(iii) $B(\boldsymbol{x})$ es continua.

Idealmente nos gustaría trabajar con restriciones de la forma $g_i(\boldsymbol{x}) \leq 0$ en la Ecuación (7.2), pero esto no es posible pues la función $B(\boldsymbol{x})$ no está definida en los puntos con algún $g_i(\boldsymbol{x}) = 0$.

La idea del método es sencilla. Si partimos de un punto interior $\boldsymbol{x}^k$, es decir, un punto en el conjunto $G = \{\boldsymbol{x} : \boldsymbol{g}(\boldsymbol{x}) < 0\}$, entonces la función de barrera impedirá que en la iteración actual nos salgamos de dicho conjunto. Dado que los métodos de optimización no son capaces de trabajar con restricciones estrictas como las que definen el conjunto G, en la práctica es habitual no tener en cuenta explíticitamente las restricciones $g_i(\boldsymbol{x}) < 0$, pero a cambio hay que hacer alguna verificación para controlar que no nos salimos del conjunto G. En este sentido, métodos que se mueven localmente desde el punto actual deberían mantenerse siempre en el conjunto G, sin pasar nunca más allá de la "barrera". Una función de barrera natural es $B(\boldsymbol{x}) = \sum_{i=1}^m \frac{-1}{g_i(\boldsymbol{x})}$, aunque en la práctica es más utilizada $B(\boldsymbol{x}) = -\sum_{i=1}^m \log(-g_i(\boldsymbol{x}))$, pues da lugar a resultados de convergencia más fuertes. En la Figura 7.2 representamos los subproblemas que resultaría de aplicar estas dos funciones de barrera que acabamos de mencionar para el problema

$$\begin{array}{rl} \text{mín} & x^2 \\ \text{sujeto a} & -x \leq 0, \end{array}$$

cuyo óptimo es $\bar{x} = 0$ con $f(x) = 0$. En la figura ilustramos también los óptimos de los subproblemas y podemos ver como en ambos casos, a medida que el valor de μ se hace más



(a) Subproblemas con $B(x) = \frac{1}{x}$.

(b) Subproblemas con $B(x) = -\log(x)$.

Figura 7.2: Ilustrando los subproblemas resultantes de aplicar el método de barrera al problema de minimizar $f(x) = x^2$ sujeto a $-x \leq 0$. La línea sólida representa la función original, $f(x) = x^2$, y las líneas discontinuas los subproblemas con μ tomando los valores 10, 5, 2 y 0.5.

pequeño dichos óptimos se acercan a $\bar{x} = 0$ y lo hacen desde dentro del conjunto factible ($x \geq 0$).

En la Figura 7.3 presentamos el esquema general de un algoritmo de penalización interior o de barrera. Puede observarse que el esquema del método es muy similar al método de penalización exterior. Sin embargo, el funcionamiento del método es muy distinto debido a que uno se mueve por fuera de la región factible y otro por dentro de la misma. Se puede probar que, bajos los supuestos adecuados, un método de barrera converge a una solución del problema original, pero no vamos a entrar en los detalles.

Método de penalización interior o de barrera

Inicialización

Elegir la función de barrera $B(\boldsymbol{x})$ y un parámetro de terminación $\varepsilon > 0$. Elegir un parámetro inicial $\mu^1 > 0$ y un escalar $\gamma \in (0, 1)$. Elegir un punto inicial $\boldsymbol{x}^1 \in S$ tal que $\boldsymbol{g}(\boldsymbol{x}^1) < \boldsymbol{0}$. Definir $k = 1$.

Paso 1

Partiendo de $\boldsymbol{x}^k$, resolver el problema

$$\begin{array}{rl} \text{mín} & f(\boldsymbol{x}) + \mu^k B(\boldsymbol{x}) \\ \text{sujeto a} & \boldsymbol{g}(\boldsymbol{x}) < 0 \\ & \boldsymbol{x} \in S, \end{array}$$

y denotemos por $\boldsymbol{x}^{k+1}$ a la solución óptima encontrada.

- Si $\mu^k B(\boldsymbol{x}) < \varepsilon$, Fin. Devolvemos $\boldsymbol{x}^{k+1}$.
- En otro caso, definir $\mu^{k+1} = \gamma \mu^k$. Reemplazar k por $k + 1$. Repetir el Paso 1.

Figura 7.3: Esquema del método de barrera.

Los métodos de barrera también tienen importantes limitaciones, que comentamos a continuación:

- Necesitamos problemas en los que el conjunto G no sea vacío y ser capaces de identificar un punto inicial dentro del mismo. En general, esta tarea no es sencilla.
- Estos métodos no sirven para resolver problemas con restricciones de igualdad. Teóricamente podemos reemplazar una restricción de la forma $h_j(\boldsymbol{x}) = 0$ por las restricciones $h_j(\boldsymbol{x}) \geq 0$ y $h_j(\boldsymbol{x}) \leq 0$ y transformar cualquier problema en un problema únicamente con restricciones de igualdad. Sin embargo, aunque algunos algoritmos permiten hacer esto, este no es el caso de los métodos de barrera, pues en los subproblemas tendríamos que verificar simultáneamente $h_j(\boldsymbol{x}) > 0$ y $h_j(\boldsymbol{x}) < 0$, lo que no es posible. Dicho de otra forma, el conjunto G sería vacío.
- Al igual que los métodos de penalización exterior, los métodos de barrera también tienen problemas de condicionamiento. En este caso, los problemas aparecen cuando μ^k se hace muy pequeño.

7.3 Método del lagrangiano aumentado

7.3.1 Motivación

Como comentamos en la sección anterior, una de las mayores limitaciones de los métodos de penalización exterior radica en la necesidad de trabajar con penalizaciones muy grandes, lo que resulta en subproblemas mal condicionados y, por tanto, de difícil solución. Una forma de evitar este problema es a través de la función de penalización absoluta, $P^a(\boldsymbol{x})$, pero que resulta en subproblemas no diferenciables, lo que tampoco es deseable.

Una alternativa bastante usada en la práctica es recurrir a técnicas de relajación lagrangiana, ya discutidas brevemente en las secciones 5.5.1 y 6.5, donde se trabaja con el dual lagrangiano y el papel del parámetro de penalización lo asumen los multiplicadores de Lagrange. Si bien estas técnicas tienen muchas virtudes, por ejemplo a la hora de descomponer el problema original en subproblemas, los resultados de convergencia asociados suelen requerir supuestos bastante fuertes. La resolución del dual lagrangiano normalmente pasa por el uso de algoritmos iterativos basados en direcciones de ascenso (en la línea del método de máximo descenso para problemas de optimización sin restricciones). Como vimos en la Sección 5.4.6, la principal dificultad radica en que el problema dual suele ser no diferenciable, lo que obliga a usar subgradientes y en la práctica no siempre es fácil encontrar subgradientes que sean direcciones de ascenso.

En esta sección presentamos un nuevo "método de penalización", que es un refinamiento de las técnicas de relajación lagrangiana. Permite atajar el problema de la actualización de las penalizaciones sin tener que pasar por la resolución de subproblemas no diferenciables. Se trata del método de lagrangiano aumentado, cuya idea pasamos a desarrollar a continuación. Para facilitar la exposición, comenzamos trabajando con problemas que tienen únicamente restricciones de igualdad:

Problema P

$$\begin{aligned} \text{minimizar} \quad & f(\boldsymbol{x}) \\ \text{sujeto a} \quad & h_j(\boldsymbol{x}) = 0 \qquad j = 1, \ldots, l. \end{aligned}$$

Supongamos ahora que, en vez de resolver este problema, estudiamos el problema

$$\begin{array}{c}\text{Problema PP}\\ \begin{array}{rl}\text{minimizar} & f(\boldsymbol{x})+\frac{\rho}{2}\sum_{j=1}^{l}h_j(\boldsymbol{x})^2=f(\boldsymbol{x})+\frac{\rho}{2}P^c(\boldsymbol{x})\\ \text{sujeto a} & h_j(\boldsymbol{x})=0 \qquad j=1,\dots,l.\end{array}\end{array}$$

El problema PP es una versión penalizada del problema P; realmente se podría penalizar con ρ en vez de $\frac{\rho}{2}$, pero para los desarrollos resulta más cómoda esta formulación. Tenemos la penalización cuadrática en la función objetivo por no cumplir las restricciones, pero mantenemos las restricciones explícitas en el problema. Nótese que los problemas P y PP son equivalentes. Ambos tienen la misma región factible y la función objetivo toma el mismo valor en cualquier punto factible.

El carácter estrictamente convexo de la penalización cuadrática le da una mayor regularidad al problema. En particular, es importante para poder apoyarse en el apartado (II) del Teorema 2.2 y que, con bastante generalidad, la función dual, dados los valores óptimos del dual, tenga un óptimo único. En este caso, la Proposición 5.28 garantiza la diferenciabilidad de la función dual en el óptimo.

Por otro lado, tenemos que la función lagrangiana del problema P viene dada por

$$L(\boldsymbol{x},\boldsymbol{v})=f(\boldsymbol{x})+\sum_{j=1}^{l}v_jh_j(\boldsymbol{x}),$$

mientras que la función lagrangiana del problema PP, conocida como *lagrangiano aumentado*, viene dada por

$$L_\rho^A(\boldsymbol{x},\boldsymbol{v})=f(\boldsymbol{x})+\sum_{j=1}^{l}v_jh_j(\boldsymbol{x})+\frac{\rho}{2}\sum_{j=1}^{l}h_j(\boldsymbol{x})^2.$$

Además de ver esta función como la función lagrangiana del problema PP, también se puede ver como la función objetivo que obtendríamos si aplicamos el método de penalización exterior al problema de minimizar la función lagrangiana $f(\boldsymbol{x})+\sum_{j=1}^{l}v_jh_j(\boldsymbol{x})$, sujeto a $\boldsymbol{h}(\boldsymbol{x})=\boldsymbol{0}$, que también es un problema equivalente a P.

Supongamos ahora que tenemos un punto KKT del problema P, $\bar{\boldsymbol{x}}$, con multiplicadores de Lagrange asociados dados por $\bar{\boldsymbol{v}}$. Esto quiere decir que, fijado $\bar{\boldsymbol{v}}$, $\bar{\boldsymbol{x}}$ minimiza la función lagrangiana, lo que implica que

$$\nabla_{\boldsymbol{x}}L(\bar{\boldsymbol{x}},\bar{\boldsymbol{v}})=\nabla f(\bar{\boldsymbol{x}})+\sum_{j=1}^{l}\bar{v}_j\nabla h_j(\bar{\boldsymbol{x}})=\boldsymbol{0}.$$

Por otro lado, por ser $\bar{\boldsymbol{x}}$ factible también se cumple que

$$\nabla_{\boldsymbol{x}}L_\rho^A(\bar{\boldsymbol{x}},\bar{\boldsymbol{v}})=\nabla f(\bar{\boldsymbol{x}})+\sum_{j=1}^{l}\bar{v}_j\nabla h_j(\bar{\boldsymbol{x}})+\rho\sum_{j=1}^{l}h_j(\boldsymbol{x})\nabla h_j(\bar{\boldsymbol{x}})=\boldsymbol{0}.$$

Esta última igualdad es cierta independientemente del valor del parámetro de penalización ρ. Por otro lado, trabajando con el método de penalización exterior (con penalización cuadrática) tendríamos que el gradiente de la función penalizada en $\bar{\boldsymbol{x}}$ vendría dado por $\nabla f(\bar{\boldsymbol{x}})+$

$\rho \sum_{j=1}^{l} h_j(\boldsymbol{x}) \nabla h_j(\bar{\boldsymbol{x}}) = \nabla f(\bar{\boldsymbol{x}})$, que sólo se anulará si $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$, lo que contribuía a que, en general, $\bar{\boldsymbol{x}}$ no fuese un óptimo del problema penalizado para ningún valor finito de ρ. Nótese que la función L_ρ^A mantiene cualquier propiedad de diferenciabilidad que tengan las funciones que definen el problema P.

El razonamiento que acabamos de presentar sugiere que uno puede minimizar el lagrangiano aumentado para resolver el problema original, sin la limitación de necesitar que ρ tienda a infinito. La siguiente sección está dedicada a dar soporte teórico a esta intuición.

En cierta manera, podemos ver el algoritmo de lagrangiano aumentado como una mejora de la relajación lagrangiana en la que se mitigan los problemas de diferenciabilidad en la función dual y, al mismo tiempo, como una mejora de los métodos penalizados en la que, manteniendo la diferenciabilidad, no necesitamos incrementar indefinidamente las penalizaciones.

7.3.2 Soporte teórico del método de lagrangiano aumentado

Comenzamos el desarrollo de los fundamentos teóricos del método del lagrangiano aumentado con un lema auxiliar.

Lema 7.3. *Sean $\boldsymbol{A}$ y $\boldsymbol{B}$ matrices simétricas con $\boldsymbol{B}$ semidefinida positiva y $\boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x} > 0$ para todo $\boldsymbol{x} \neq \mathbf{0}$ con $\boldsymbol{x}^\intercal \boldsymbol{B}\boldsymbol{x} = 0$. Entonces existe $\bar{\rho} > 0$ tal que $\boldsymbol{A} + \rho \boldsymbol{B}$ es definida positiva para todo $\rho > \bar{\rho}$.*

Demostración. Supongamos que el resultado no es cierto. Consideremos entonces una sucesión no negativa $\{\rho^t\}_{t\in\mathbb{N}}$ tal que $\lim_{t\to\infty} \rho^t = \infty$. Entonces, para cada $t \in \mathbb{N}$ tenemos $\boldsymbol{x}^t$ tal que $(\boldsymbol{x}^t)^\intercal(\boldsymbol{A} + \rho \boldsymbol{B})\boldsymbol{x}^t \leq 0$. Sin pérdida de generalidad podemos tomar la sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ tal que $\|\boldsymbol{x}^t\| = 1$. Por tanto, esta sucesión está contenida en un compacto y tendrá una subsucesión convergente. Supongamos, sin pérdida de generalidad, que la propia sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ es convergente y sea $\bar{\boldsymbol{x}}$ su límite.

Entonces tenemos que, para todo $t \in \mathbb{N}$, $(\boldsymbol{x}^t)^\intercal \boldsymbol{A}\boldsymbol{x}^t + \rho^t (\boldsymbol{x}^t)^\intercal \boldsymbol{B}\boldsymbol{x}^t \leq 0$ y $(\boldsymbol{x}^t)^\intercal \boldsymbol{B}\boldsymbol{x}^t \geq 0$. En particular, para todo $t \in \mathbb{N}$, $(\boldsymbol{x}^t)^\intercal \boldsymbol{A}\boldsymbol{x}^t \leq 0$. Tomando límites obtenemos que $(\bar{\boldsymbol{x}})^\intercal \boldsymbol{B}\bar{\boldsymbol{x}} = 0$ pues, de otro modo, $\rho^t (\boldsymbol{x}^t)^\intercal \boldsymbol{B}\boldsymbol{x}^t$ tendería a $+\infty$, mientras que $(\boldsymbol{x}^t)^\intercal \boldsymbol{A}\boldsymbol{x}^t$ tiende a $(\bar{\boldsymbol{x}})^\intercal \boldsymbol{A}\bar{\boldsymbol{x}}$, contradiciendo que $(\boldsymbol{x}^t)^\intercal \boldsymbol{A}\boldsymbol{x}^t + \rho^t (\boldsymbol{x}^t)^\intercal \boldsymbol{B}\boldsymbol{x}^t \leq 0$ para todo $t \in \mathbb{N}$.

Ahora bien, si $(\bar{\boldsymbol{x}})^\intercal \boldsymbol{B}\bar{\boldsymbol{x}} = 0$, entonces los supuestos sobre $\boldsymbol{A}$ nos aseguran que $(\bar{\boldsymbol{x}})^\intercal \boldsymbol{A}\bar{\boldsymbol{x}} > 0$ lo que contradice que para todo $t \in \mathbb{N}$, $(\boldsymbol{x}^t)^\intercal \boldsymbol{A}\boldsymbol{x}^t \leq 0$. □

Antes de presentar el resultado matemático en el que se sustenta el método del lagrangiano aumentado, veamos cómo queda la condición suficiente de KKT de segundo orden del Teorema 5.8 para el problema P. Supongamos que $\bar{\boldsymbol{x}}$ es un punto KKT de P con multiplicadores de Lagrange dados por $\bar{\boldsymbol{v}}$. Si la hessiana de la función lagrangiana restringida al primal en $\bar{\boldsymbol{v}}$,

$$\nabla^2 L^P(\bar{\boldsymbol{x}}) = \nabla^2 f(\bar{\boldsymbol{x}}) + \sum_{j=1}^{l} \bar{v}_i \nabla^2 h_j(\bar{\boldsymbol{x}}),$$

cumple que $\boldsymbol{d}^\intercal \nabla^2 L^P(\bar{\boldsymbol{x}})\boldsymbol{d} > 0$ para todo $\boldsymbol{d} \in C(\bar{\boldsymbol{x}})$, el punto $\bar{\boldsymbol{x}}$ es un mínimo local estricto. Recordemos que $C(\bar{\boldsymbol{x}}) = \{\boldsymbol{d} \in \mathbb{R}^n : \boldsymbol{d} \neq \mathbf{0} \text{ y } \nabla h_j(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} = 0, \forall j \in \{1, \ldots, l\}\}$. La demostración del siguiente resultado se apoyará en aplicar el Lema 7.3 con $\nabla^2 L^P(\bar{\boldsymbol{x}})$ desempeñando el rol de $\boldsymbol{A}$ y $\sum_{j=1}^{l} \nabla h_j(\bar{\boldsymbol{x}}) \nabla h_j(\bar{\boldsymbol{x}})^\intercal$ el de $\boldsymbol{B}$.

Teorema 7.4. *Supongamos que $\bar{\boldsymbol{x}}$ es un punto de KKT con multiplicadores $\bar{\boldsymbol{v}}$ que cumple la condición suficiente de KKT de segundo orden. Entonces existe $\bar{\rho} > 0$ tal que, para todo $\rho \geq \bar{\rho}$, la función $L_\rho^A(\cdot, \bar{\boldsymbol{v}})$ tiene un mínimo local estricto en $\bar{\boldsymbol{x}}$.*

Demostración. El resultado nos dice que la función

$$L_\rho^A(\boldsymbol{x}, \boldsymbol{v}) = f(\boldsymbol{x}) + \sum_{j=1}^{l} v_j h_j(\boldsymbol{x}) + \frac{\rho}{2} \sum_{j=1}^{l} h_j(\boldsymbol{x})^2,$$

una vez fijado $\boldsymbol{v} = \bar{\boldsymbol{v}}$, como función de $\boldsymbol{x}$ tiene un mínimo local estricto en $\bar{\boldsymbol{x}}$. Vamos a evaluar el gradiente y la hessiana de la función $L_\rho^A(\cdot, \bar{\boldsymbol{v}})$ en el punto $\bar{\boldsymbol{x}}$.

$$\begin{aligned}\nabla_{\boldsymbol{x}} L_\rho^A(\bar{\boldsymbol{x}}, \bar{\boldsymbol{v}}) &= \nabla f(\bar{\boldsymbol{x}}) + \sum_{j=1}^{l} \bar{v}_j \nabla h_i(\bar{\boldsymbol{x}}) + \rho \sum_{j=1}^{l} h_j(\bar{\boldsymbol{x}}) \nabla h_j(\bar{\boldsymbol{x}}) \\ &\overset{h_j(\bar{\boldsymbol{x}})=0}{=} \nabla f(\bar{\boldsymbol{x}}) + \sum_{j=1}^{l} \bar{v}_j \nabla h_j(\bar{\boldsymbol{x}}) = \nabla L^P(\bar{\boldsymbol{x}}) \overset{\bar{\boldsymbol{x}} \text{ KKT}}{=} 0.\end{aligned}$$

Ya tenemos la condición de primer orden del problema de optimización sin restricciones asociado a minimizar $L_\rho^A(\cdot, \bar{\boldsymbol{v}})$. Veamos ahora la condición de segundo orden.

$$\begin{aligned}\nabla^2_{\boldsymbol{xx}} L_\rho^A(\bar{\boldsymbol{x}}, \bar{\boldsymbol{v}}) &= \nabla^2 f(\bar{\boldsymbol{x}}) + \sum_{j=1}^{l} \left(\bar{v}_j + \rho h_j(\bar{\boldsymbol{x}})\right) \nabla^2 h_j(\bar{\boldsymbol{x}}) + \rho \sum_{j=1}^{l} \nabla h_j(\bar{\boldsymbol{x}}) \nabla h_j(\bar{\boldsymbol{x}})^\intercal \\ &\overset{h_j(\bar{\boldsymbol{x}})=0}{=} \nabla^2 L^P(\bar{\boldsymbol{x}}) + \rho \sum_{j=1}^{l} \nabla h_j(\bar{\boldsymbol{x}}) \nabla h_j(\bar{\boldsymbol{x}})^\intercal.\end{aligned}$$

La condición suficiente de KKT nos asegura que $\nabla^2 L^P(\bar{\boldsymbol{x}})$ cumple que $\boldsymbol{d}^\intercal \nabla^2 L^P(\bar{\boldsymbol{x}}) \boldsymbol{d} > 0$ para todo $\boldsymbol{d} \in C(\bar{\boldsymbol{x}}) = \{\boldsymbol{d} \in \mathbb{R}^n : \boldsymbol{d} \neq \boldsymbol{0} \text{ y } \nabla h_j(\bar{\boldsymbol{x}})^\intercal \boldsymbol{d} = 0, \forall j \in \{1, \ldots, l\}\}$. Como la matriz $\sum_{j=1}^{l} \nabla h_j(\bar{\boldsymbol{x}}) \nabla h_j(\bar{\boldsymbol{x}})^\intercal$ es semidefinida positiva por ser suma de matrices semidefinidas positivas, podemos aplicar el Lema 7.3 con $\boldsymbol{A} = \nabla^2 L^P(\bar{\boldsymbol{x}})$ y $\boldsymbol{B} = \sum_{j=1}^{l} \nabla h_j(\bar{\boldsymbol{x}}) \nabla h_j(\bar{\boldsymbol{x}})^\intercal$.

Por tanto, $\nabla^2_{\boldsymbol{xx}} L_\rho^A(\bar{\boldsymbol{x}}, \bar{\boldsymbol{v}})$ es definida positiva y tenemos que $\bar{\boldsymbol{x}}$ es un mínimo local estricto de $L_\rho^A(\cdot, \bar{\boldsymbol{v}})$. □

A la vista de este resultado, parece natural minimizar la función $L_\rho^A(\cdot, \bar{\boldsymbol{v}})$ en vez de minimizar las funciones penalizadas $f_P(\boldsymbol{x})$ como hacía el método de penalización exterior. De cara al diseño de un algoritmo para encontrar un óptimo de P, el Teorema 7.4 nos asegura que no será necesario hacer tender ρ a infinito y que mantendremos las condiciones de diferenciabilidad de P en los subproblemas. Aunque todo esto es esencialmente cierto, hay un problema que el algoritmo debe solventar, y es que no conocemos $\bar{\boldsymbol{v}}$. Es por este motivo que el método de lagrangiano aumentado irá actualizando iteración a iteración no sólo el valor de $\bar{\boldsymbol{x}}$, sino también el valor de $\bar{\boldsymbol{v}}$ (y $\bar{\boldsymbol{u}}$ cuando haya restricciones de desigualdad).

Antes de definir formalmente el algoritmo presentamos una interpretación adicional de la función L_ρ^A y también algún ejemplo para ilustrar los conceptos desarrollados hasta el momento.

7.3.3 Discusión y ejemplos

Consideremos el lagrangiano aumentado:

$$L_\rho^A(\boldsymbol{x}, \boldsymbol{v}) = f(\boldsymbol{x}) + \sum_{j=1}^{l} v_j h_j(\boldsymbol{x}) + \frac{\rho}{2} \sum_{j=1}^{l} h_j(\boldsymbol{x})^2 = f(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})^\intercal \boldsymbol{v} + \frac{\rho}{2} \|\boldsymbol{h}(\boldsymbol{x})\|^2.$$

No es difícil ver que esta función se puede escribir, equivalentemente, de la siguiente forma:

$$L_\rho^A(\boldsymbol{x}, \boldsymbol{v}) = f(\boldsymbol{x}) + \frac{\rho}{2} \sum_{j=1}^{l} (h_j(\boldsymbol{x}) + \frac{v_j}{\rho})^2 + c = f(\boldsymbol{x}) + \frac{\rho}{2} \|\boldsymbol{h}(\boldsymbol{x}) + \frac{\boldsymbol{v}}{\rho}\|^2 + c.$$

donde c no depende de $\boldsymbol{x}$, con lo que no tendrá ningún impacto a la hora de minimizar la función lagrangiana con respecto de $\boldsymbol{x}$. Con esta representación, la función $L_\rho^A(\boldsymbol{x}, \boldsymbol{v})$ se puede pensar como un método de penalización en el que se está penalizando la desviación no con respecto a $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$, sino con respecto a $\boldsymbol{h}(\boldsymbol{x}) + \frac{\boldsymbol{v}}{\rho} = \boldsymbol{0}$. Es decir, se está desplazando la restricción original. Sin embargo, a medida que ρ tiende a infinito, este desplazamiento se converge a cero, con lo que en el límite estaríamos resolviendo el problema P. El Teorema 7.4 asegura que, si elegimos de forma adecuada el vector $\boldsymbol{v}$, no es necesario hacer tender ρ a infinito.

Con respecto a las técnicas habituales de relajación lagrangiana, la principal aportación del método de lagrangiano aumentado es la inclusión en el lagrangiano del término de regularización $\|\boldsymbol{h}(\boldsymbol{x})\|^2$. Esto tiene dos efectos importantes:

- En el lado positivo, el término de regularización asegura que, bajo supuestos bastante generales, el problema dual será diferenciable, lo que facilita su resolución mediante métodos de ascenso y mejora notablemente la convergencia del algoritmo.

- En el lado negativo, aunque las restricciones $\boldsymbol{h}(\boldsymbol{x})$ tengan una estructura subyacente que permitiría descomponer la función lagrangiana como funciones independientes (como en el problema de planificación energética visto en la Sección 5.5.1), esta estructura de descomposición se pierde al incluir el término $\|\boldsymbol{h}(\boldsymbol{x})\|^2$ que "estropea" cualquier estructura de separabilidad de las funciones $\boldsymbol{h}(\boldsymbol{x})$. Esta es una de las razones por las cuales la relajación lagrangiana sigue siendo todavía una técnica muy utilizada en la práctica. Asimismo, cabe destacar que hay variantes del algoritmo de lagrangiano aumentado que se han usado exitosamente para descomponer clases especiales de problemas. Un ejemplo destacado es el algoritmo *progressive hedging* introducido en Rockafellar y Wets (1991) para resolver problemas de optimización estocástica.

Ejemplo 7.2. Volvamos ahora el problema del Ejemplo 7.1:

$$\begin{array}{rl} \text{minimizar} & \frac{1}{2}x_1^2 + x_2^2 \\ \text{sujeto a} & x_1 - 1 = 0, \end{array}$$

para el que habíamos visto que el método de penalización exterior con penalización cuadrática no alcanzaba el óptimo $\bar{\boldsymbol{x}} = (1, 0)$ para ningún valor finito de ρ, lo cual además ponía de manifiesto los problemas de condicionamiento del método.

La condición de KKT aplicada al punto $\bar{\boldsymbol{x}} = (1, 0)$ nos dice que ha de existir un multiplicador de Lagrange $\bar{v}$ tal que

$$\mathbf{0} = \nabla f(\bar{\boldsymbol{x}}) + \bar{v}\,\nabla h(\bar{\boldsymbol{x}}) = (1, 0) + \bar{v}(1, 0).$$

De donde obtenemos $\bar{v} = -1$. Por otro lado, el lagrangiano aumentado es de la forma

$$L_\rho^A(\boldsymbol{x}, v) = \frac{1}{2}x_1^2 + x_2^2 + v(x_1 - 1) + \frac{\rho}{2}(x_1 - 1)^2.$$

Si intentamos minimizar $L_\rho^A(\boldsymbol{x}, v)$ como función de $\boldsymbol{x}$ obtenemos la condición

$$\nabla_{\boldsymbol{x}}\, L_\rho^A(\boldsymbol{x}, v) = (x_1 + v + \rho(x_1 - 1), 2x_2) = (0, 0),$$

de donde sacamos

$$x_1 = \frac{\rho - v}{\rho + 1} \qquad x_2 = 0.$$

En particular, tenemos que si hacemos tender v a $\bar{v}$, x_1 tiende a 1. Además, esto es cierto para cualquier $\rho > 0$. A la vista de esto, está claro que el lagrangiano aumentado resulta prometedor como mejora de los métodos de penalización exterior. Por supuesto, todavía está abierta la pregunta de cómo definir el método de lagrangiano aumentado cuando, como será el caso en la práctica, no conocemos el multiplicador asociado al óptimo que estamos buscando. La respuesta la veremos en la siguiente sección. ◇

A continuación presentamos otro ejemplo que muestra que si el punto $\bar{\boldsymbol{x}}$ del enunciado del Teorema 7.4 no cumple la condición de suficiente de KKT de segundo orden, entonces puede ser que también para la función L_ρ^A necesitemos hacer tender ρ a infinito.

Ejemplo 7.3. Consideremos el problema de optimización

$$\begin{array}{rl} \text{minimizar} & x_1^4 + x_1 x_2 \\ \text{sujeto a} & x_2 = 0. \end{array}$$

Claramente, $\bar{\boldsymbol{x}} = (0, 0)$ es la solución óptima. Como $\nabla f(\boldsymbol{x}) = (4x_1^3 + x_2, x_1)$ tenemos que $\nabla f(\bar{\boldsymbol{x}}) = (0, 0)$, obtenemos que el multiplicador asociado es $\bar{v} = 0$. Para la matriz de lagrangiano L^P, como $\nabla^2 h(\boldsymbol{x}) = \mathbf{0}$ para todo $\boldsymbol{x}$, tenemos

$$\nabla^2 L^P(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) = \begin{pmatrix} 12x_1^2 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{y} \quad \nabla^2 L^P(0, 0) = \nabla^2 f(0, 0) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Además, como $\nabla h(0, 0) = (0, 1)$, tenemos $C(0, 0) = \{\mathbf{0} \neq \boldsymbol{d} \in \mathbb{R}^2 : (0, 1) \cdot \boldsymbol{d} = 0\}$. En particular, $(1, 0) \in C(0, 0)$ y

$$(1, 0)\,\nabla^2 L^P(\boldsymbol{x})(1, 0)^\intercal = (1, 0) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} (1, 0)^\intercal = 0.$$

Por tanto, no se cumple la condición suficiente de KKT de segundo orden y no estamos en los supuestos del Teorema 7.4. Veamos que pasa si intentamos minimizar el lagrangiano aumentado $L_\rho^A(\boldsymbol{x}, v)$ tomando $v = \bar{v} = 0$.

$$L_\rho^A(\boldsymbol{x}, 0) = x_1^4 + x_1 x_2 + \frac{\rho}{2}x_2^2.$$

Ahora tenemos

$$\nabla L_\rho^A(\boldsymbol{x},0) = (4x_1^3 + x_2, x_1 + \rho x_2) \quad \text{y} \quad \nabla^2 L_\rho^A(\boldsymbol{x},0) = \begin{pmatrix} 12x_1^2 & 1 \\ 1 & \rho x_2 \end{pmatrix}.$$

El gradiente se anula en $\bar{\boldsymbol{x}} = (0,0)$ y también en $\hat{\boldsymbol{x}} = (\frac{1}{2\sqrt{\rho}}, -\frac{1}{2\rho\sqrt{\rho}})$ y en $-\hat{\boldsymbol{x}}$. Además, se puede comprobar fácilmente que $\nabla^2 L_\rho^A$ no es semidefinida positiva en $\bar{\boldsymbol{x}} = (0,0)$, con lo que $\bar{\boldsymbol{x}}$ no es un mínimo local para ningún $\rho > 0$. Sin embargo $\nabla^2 L_\rho^A$ sí que es definida positiva en $\hat{\boldsymbol{x}}$ que de hecho es el mínimo de $L_\rho^A(\boldsymbol{x},0)$ para todo $\rho > 0$. Además, $\hat{\boldsymbol{x}}$ tiende a $\boldsymbol{x}$ cuando ρ tiende a infinito.

Lo que hemos visto en este ejemplo es que si no se cumple la condición suficiente de KKT de segundo orden en el punto $\bar{\boldsymbol{x}}$, entonces el lagrangiano aumentado podría llegar a tener un comportamiento parecido al del método de penalización exterior (aunque también podría comportarse bien, dependiendo del problema). ◇

7.3.4 Incorporando restricciones de desigualdad

Supongamos ahora que tenemos un problema con restricciones de desigualdad de la forma

$$\begin{aligned} \text{minimizar} \quad & f(\boldsymbol{x}) \\ \text{sujeto a} \quad & g_i(\boldsymbol{x}) \leq 0 \qquad i = 1, \ldots, m. \end{aligned}$$

Estas desigualdades se pueden transformar en igualdades sin más que añadir variables auxiliares z_i y poner las restricciones

$$h_i(x) = g_i(x) + z_i^2 = 0.$$

Como regla general estas restricciones no suelen ser nada buenas. La razón es que la derivada con respecto a z_i se anula justamente cuando $z_i = 0$ lo que dificultará que el problema resultante tenga buenas propiedades de regularidad. En la práctica esta transformación se suele hacer añadiendo directamente las variables z_i a las restricciones y exigiendo que $z_i \geq 0$.

Sin embargo, en nuestro caso esta transformación es un artificio teórico cuyo comportamiento será bueno porque el término z_i^2 lo vamos a "calcular" explícitamente. Con la formulación que vimos al principio de la Sección 7.3.3, para estas nuevas restricciones de desigualdad h_i el lagragiano aumentado queda de la forma

$$L_\rho^A(\boldsymbol{x},\boldsymbol{u}) = f(\boldsymbol{x}) + \frac{\rho}{2}\sum_{i=1}^m (h_i(\boldsymbol{x}) + \frac{u_i}{\rho})^2 + c = f(\boldsymbol{x}) + \frac{\rho}{2}\sum_{i=1}^m (g_i(\boldsymbol{x}) + z_i^2 + \frac{u_i}{\rho})^2 + c.$$

El método de lagrangiano aumentado tendrá como paso principal la minimización en $\boldsymbol{x}$ de $L_\rho^A(\boldsymbol{x},\boldsymbol{u})$ para distintos valores de $\boldsymbol{u}$. En este caso, dado un vector $\boldsymbol{x}$ tendremos que

$$z_i = \begin{cases} 0 & \text{si } g_i(\boldsymbol{x}) + \frac{u_i}{\rho} \geq 0 \\ \sqrt{-g_i(\boldsymbol{x}) - \frac{u_i}{\rho}} & \text{en otro caso.} \end{cases}$$

Por tanto, cada término $(g_i(\boldsymbol{x}) + z_i^2 + \frac{u_i}{\rho})^2$ tomará el valor $\max\{0, g_i(\boldsymbol{x}) + \frac{u_i}{\rho}\}^2$. Entonces, para trabajar con restricciones de desigualdad bastará tomar

$$L_\rho^A(\boldsymbol{x},\boldsymbol{u}) = f(\boldsymbol{x}) + \frac{\rho}{2}\sum_{i=1}^m \max\left\{0, g_i(\boldsymbol{x}) + \frac{u_i}{\rho}\right\}^2 + c.$$

En el caso de problemas con restricciones de igualdad y desigualdad tendremos

$$L_\rho^A(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) = f(\boldsymbol{x}) + \frac{\rho}{2}\Big(\sum_{i=1}^{m} \text{máx}\left\{0, g_i(\boldsymbol{x}) + \frac{u_i}{\rho}\right\}^2 + \sum_{j=1}^{l}\left(h_j(\boldsymbol{x}) + \frac{v_j}{\rho}\right)^2\Big) + c.$$

7.3.5 Descripción del método de lagrangiano aumentado

Tras lo discutido en los apartados anteriores, ya estamos en condiciones de presentar el método de lagrangiano aumentado. Supongamos que queremos resolver un problema de programación no lineal

Problema P

$$\begin{array}{rll} \text{minimizar} & f(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 & i = 1,\dots,m \\ & h_j(\boldsymbol{x}) = 0 & j = 1,\dots,l. \end{array}$$

El funcionamiento del método será el siguiente. En cada iteración t tendremos un parámetro de penalización ρ^t y aproximaciones $\boldsymbol{u}^t$ y $\boldsymbol{v}^t$ a los multiplicadores de Lagrange que estamos buscando. En ese momento buscaremos el mínimo en $\boldsymbol{x}$ del lagrangiano aumentado $L_{\rho^t}^A(\boldsymbol{x},\boldsymbol{u}^t,\boldsymbol{v}^t)$ que, olvidándonos del término constante c, es equivalente a minimizar

$$f(\boldsymbol{x}) + \frac{\rho^t}{2}\Big(\sum_{i=1}^{m} \text{máx}\left\{0, g_i(\boldsymbol{x}) + \frac{u_i^t}{\rho^t}\right\}^2 + \sum_{j=1}^{l}\left(h_j(\boldsymbol{x}) + \frac{v_j^t}{\rho^t}\right)^2\Big).$$

Supongamos que $\boldsymbol{x}^{t+1}$ es el mínimo de este problema de minimización. Es importante destacar que, en la práctica, el problema de minimización de la iteración t se resolverá utilizando, por ejemplo, un método de Newton, cuasi-Newton o gradiente conjugado (dependiendo de las características del problema). Además, este método de optimización sin restricciones se suele resolver tomando como solución inicial $\boldsymbol{x}^t$, la solución de la iteración anterior, lo que puede mejorar sustancialmente la velocidad del algoritmo completo. Para el mínimo $\boldsymbol{x}^{t+1}$ tendremos que $\nabla_{\boldsymbol{x}} L_{\rho^t}^A(\boldsymbol{x}^{t+1},\boldsymbol{u}^t,\boldsymbol{v}^t) = 0$ con lo que

$$\nabla f(\boldsymbol{x}^{t+1}) + \sum_{i=1}^{m} \text{máx}\{0, u_i^t + \rho^t g_i(\boldsymbol{x}^{t+1})\}\,\nabla g_i(\boldsymbol{x}^{t+1}) + \sum_{j=1}^{l}(v_j^t + \rho^t h_j(\boldsymbol{x}^{t+1}))\,\nabla h_j(\boldsymbol{x}^{t+1}) = 0.$$

Esta condición es equivalente a la condición de KKT del problema P de partida. Sin embargo, se cumple con respecto a unos "multiplicadores" distintos de $\boldsymbol{u}^t$ y $\boldsymbol{v}^t$, pero que yo no conocía de antemano pues dependen de $\boldsymbol{x}^{t+1}$. De hecho, nada nos garantiza que el $\boldsymbol{x}^{t+1}$ sea factible en el problema P.

Por tanto, el método de lagrangiano aumentado actualizará los valores de cada u_i^t a $u_i^{t+1} = \text{máx}\{0, u_i^t + \rho^t g_i(\boldsymbol{x}^{t+1})\} \geq 0$ y los de cada v_j^t a $v_j^{t+1} = v_j^t + \rho^t h_j(\boldsymbol{x}^{t+1})$. Esta actualización, además de ser natural, también puede ser justificada partir de las propiedades que conocemos del dual. En la Sección 5.4.6 (Proposición 5.29) vimos que, dado un óptimo $\bar{\boldsymbol{x}}$ de la función

$$\mathcal{L}^D(\boldsymbol{u},\boldsymbol{v}) = \inf_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) + \sum_{i=1}^{m} u_i g_i(\boldsymbol{x}) + \sum_{j=1}^{l} v_j h_j(\boldsymbol{x}),$$

el vector $s = (\boldsymbol{g}(\bar{\boldsymbol{x}}), \boldsymbol{h}(\bar{\boldsymbol{x}}))$ es un subgradiente de $\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v})$. Por tanto, la actualización que hemos comentado de los u_i^t y v_j^t puede verse como un paso en la dirección del subgradiente de la función dual que, recordemos, queremos maximizar. De hecho, en el caso de que el óptimo $\bar{\boldsymbol{x}}$ de un problema $\mathcal{L}^D(\boldsymbol{u}, \boldsymbol{v})$ sea único, $s = (\boldsymbol{g}(\bar{\boldsymbol{x}}), \boldsymbol{h}(\bar{\boldsymbol{x}}))$, es el único subgradiente y la actualización propuesta es un paso en la dirección de máximo ascenso (y la regularización mediante la penalización cuadrática en el lagrangiano aumentado propicia esta unicidad).

Además de actualizar los multiplicadores para intentar que se acerquen a su valor en el óptimo buscado (solución del dual), tenemos que actualizar el parámetro ρ^t para que los iterantes $\boldsymbol{x}^t$ se vayan acercando a la región factible.

Como ya comentamos, en caso de que las sucesiones $\{\boldsymbol{u}^t\}_{t\in\mathbb{N}}$ y $\{\boldsymbol{v}^t\}_{t\in\mathbb{N}}$ converjan, la propiedad $\nabla_{\boldsymbol{x}} L^A_{\rho^t}(\boldsymbol{x}^{t+1}, \boldsymbol{u}^t, \boldsymbol{v}^t) = 0$ asegurará que la condición de KKT se cumplirá en el límite, por tanto la condición de factibilidad del dual de KKT estará asegurada en este caso.

Además, para la convergencia a un punto KKT del problema original, queda por controlar la condición de holguras complementarias, pues el sumatorio en $\nabla_{\boldsymbol{x}} L^A_{\rho^t}(\boldsymbol{x}^{t+1}, \boldsymbol{u}^t, \boldsymbol{v}^t) = 0$ para las funciones g_i recorre todas las restricciones y no sólo las activas. Sin embargo, nótese que si la sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ converge a un punto $\bar{\boldsymbol{x}}$ con $g_i(\bar{\boldsymbol{x}}) < 0$ y la sucesión $\{u_i^t\}_{t\in\mathbb{N}}$ está acotada, entonces, para ρ^t suficientemente grande u_i^t será cero y la condición de complementariedad se cumplirá.

Una de las claves del método de lagrangiano aumentado es que no es necesario actualizar la penalización ρ en todas las iteraciones. Lo que haremos será mirar si en la última iteración hemos mejorado en el cumplimiento de las condiciones de factibilidad y complementariedad y aumentar ρ únicamente cuando no haya sido el caso (sólo actualizamos ρ después de "malas" iteraciones). Para hacer esto nos hace falta una función que mida de alguna manera el grado de violación de estas condiciones. Para las restricciones de igualdad se tomará simplemente $\|\boldsymbol{h}(\boldsymbol{x})\|$. Para las restricciones de desigualdad se tomará $\|V(\boldsymbol{x}^{t+1}, \boldsymbol{u}^t, \rho^t)\|$ donde, para cada restricción de desigualdad, la medida de violación $V_i(\boldsymbol{x}^{t+1}, \boldsymbol{u}^t, \rho^t)$ viene dada por

$$V_i(\boldsymbol{x}^{t+1}, \boldsymbol{u}^t, \rho^t) = \text{máx}\{g_i(\boldsymbol{x}^{t+1}), -\frac{u_i^t}{\rho^t}\}.$$

No es difícil ver que $V_i(\boldsymbol{x}^{t+1}, \boldsymbol{u}^t, \rho^t) = 0$ si y sólo si i) $g_i(\boldsymbol{x}^{t+1}) \leq 0$ y ii) $u_i^t = 0$ cuando $g_i(\boldsymbol{x}^{t+1}) < 0$.

En la Figura 7.4 presentamos el método de lagrangiano aumentado, que integra todas las consideraciones hechas hasta el momento. Hacemos ahora un par de comentarios relativos al método:

- **Resolución de los subproblemas.** Esta resolución pasará por aplicar un método de optimización sin restricciones que se inicializará en el iterante $\boldsymbol{x}^t$, solución del subproblema de la iteración anterior.

- **Criterio de parada.** Se pueden poner parámetros distintos ε_1, ε_2 y ε_3 para las condiciones sobre C_1, C_2 y C_3.

- **Necesidad de $\boldsymbol{u}^{\text{máx}}$, $\boldsymbol{v}^{\text{mín}}$ y $\boldsymbol{v}^{\text{máx}}$.** Estos parámetros del algoritmo son necesarios para asegurar que los multiplicadores permanecen acotados y que la penalización ρ puede

MÉTODO DE LAGRANGIANO AUMENTADO

INICIALIZACIÓN

Elegir un parámetro de terminación $\varepsilon > 0$. Elegir $\rho^0 > 0$ y escalares $\delta \in (0,1)$ y $\gamma > 1$. Elegir $\boldsymbol{u}^0 \geq 0$ y $\boldsymbol{v}^0$, $\boldsymbol{u}^{\text{máx}}$, $\boldsymbol{v}^{\text{mín}}$ y $\boldsymbol{v}^{\text{máx}}$ para los multiplicadores. Elegir un punto inicial $\boldsymbol{x}^1$.

Calcular, para cada $i \in \{1,\dots,m\}$, $V_i(\boldsymbol{x}^1, \boldsymbol{u}^0, \rho^0) = \text{máx}\{g_i(\boldsymbol{x}^1), -\frac{u_i^0}{\rho^0}\}$.

Definir $t = 1$, $\rho^1 = \rho^0$, $\boldsymbol{u}^1 = \boldsymbol{u}^0$ y $\boldsymbol{v}^1 = \boldsymbol{v}^0$.

PASO 1. RESOLUCIÓN SUBPROBLEMA

Definir $\boldsymbol{x}^{t+1}$ como una solución óptima del problema $\text{mín}_{\boldsymbol{x}}\, L^A_{\rho^t}(\boldsymbol{x}, \boldsymbol{u}^t, \boldsymbol{v}^t)$; donde $\boldsymbol{u}^t$ y $\boldsymbol{v}^t$ están fijos y $\boldsymbol{x}^t$ se toma como iterante inicial para resolver el problema. Ir al PASO 2.

PASO 2. ACTUALIZACIÓN DE PARÁMETROS

Calcular, para cada $i \in \{1,\dots,m\}$, $V_i(\boldsymbol{x}^{t+1}, \boldsymbol{u}^t, \rho^t) = \text{máx}\{g_i(\boldsymbol{x}^{t+1}), -\frac{u_i^t}{\rho^t}\}$.

- Si $\text{máx}\{\|V(\boldsymbol{x}^{t+1}, \boldsymbol{u}^t, \rho^t)\|, \|\boldsymbol{h}(\boldsymbol{x}^{t+1})\|\} > \delta\, \text{máx}\{\|V(\boldsymbol{x}^t, \boldsymbol{u}^{t-1}, \rho^{t-1})\|, \|\boldsymbol{h}(\boldsymbol{x}^t)\|\}$, entonces definir $\rho^{t+1} = \gamma\rho^t$.
- En otro caso, $\rho^{t+1} = \rho^t$.

Para cada $i \in \{1,\dots,m\}$, definir $u_i^{t+1} = \text{mín}\left\{u_i^{\text{máx}}, \text{máx}\{0, u_i^t + \rho^t g_i(\boldsymbol{x}^{t+1})\}\right\}$.

Para cada $j \in \{1,\dots,l\}$:

- Si $v_j^t + \rho^t h_j(\boldsymbol{x}^{t+1}) \geq 0$, definir $v_j^{t+1} = \text{mín}\{v_j^{\text{máx}}, v_j^t + \rho^t h_j(\boldsymbol{x}^{t+1})\}$.
- En otro caso, definir $v_j^{t+1} = \text{máx}\{v_j^{\text{mín}}, v_j^t + \rho^t h_j(\boldsymbol{x}^{t+1})\}$.

Ir al PASO 3.

PASO 3. CRITERIO DE PARADA

Calculamos los siguientes parámetros:

Factibilidad dual KKT: $C_1 = \left\| \nabla f(\boldsymbol{x}^{t+1}) + \sum_{i=1}^{m} u_i^{t+1} \nabla g_i(\boldsymbol{x}^{t+1}) + \sum_{j=1}^{l} v_j^{t+1} \nabla h_j(\boldsymbol{x}^{t+1}) \right\|$.

Holguras complementarias: $C_2 = \max_{i\in\{1,\dots,m\}} \left\{ \left| \text{mín}\{-g_i(\boldsymbol{x}^{t+1}), u_i^{t+1}\} \right| \right\}$.

Factibilidad primal: $C_3 = \text{máx}\left\{ \max_{i\in\{1,\dots,m\}} g_i(\boldsymbol{x}^{t+1}), \max_{j\in\{1,\dots,l\}} |h_j(\boldsymbol{x}^{t+1})| \right\}$.

- Si $\text{máx}\{C_1, C_2, C_3\} < \varepsilon$, FIN. Devolvemos $\boldsymbol{x}^{t+1}$ y los multiplicadores $\boldsymbol{u}^{t+1}$ y $\boldsymbol{v}^{t+1}$.
- En otro caso, reemplazar t por $t+1$. Volver al PASO 1.

Figura 7.4: Esquema del método de lagrangiano aumentado.

acabar dominándolos en caso de ser necesario. Los resultados teóricos asociados a este algoritmo aseguran que, si elegimos estos parámetros suficientemente grandes, el algoritmo funcionará, aunque no es fácil saber a priori cómo de grandes deben ser. Sería suficiente tomarlos mayores que el valor de los multiplicadores en el óptimo buscado, pero no son conocidos a priori.

- **Definición de L^A.** Durante la exposición en esta sección hemos supuesto que todas las restricciones se penalizan en función objetivo. Sin embargo, el método de lagrangiano aumentado, al igual que el de penalización exterior, también se puede implementar subiendo sólo algunas de las restricciones a la función objetivo, típicamente las más difíciles. De este modo los subproblemas a resolver en cada iteración serán problemas con restricciones, pero posiblemente no muy difíciles. En particular, existen buenos algoritmos cuando todas las restricciones son de cota e incluso cuando son lineales.

Terminamos presentando el resultado relativo a las propiedades de la sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ construida por el método.

Teorema 7.5. *Si aplicamos el método de lagrangiano aumentado al problema P, entonces las sucesiones generadas tienen las siguientes propiedades:*

- *Los puntos de acumulación de la sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ son puntos KKT del siguiente problema:* minimizar $\sum_{i=1}^{m} \max\{0, g_i(\boldsymbol{x})\}^2 + \sum_{j=1}^{l} h_j(\boldsymbol{x})^2$.

- *Si un punto de acumulación $\bar{\boldsymbol{x}}$ es factible y los vectores $\nabla g_i(\bar{\boldsymbol{x}})$, con $i \in I(\bar{\boldsymbol{x}})$, y $\nabla h_j(\bar{\boldsymbol{x}})$ con $j \in \{1,\ldots,l\}$ son linealmente independientes (condición de regularidad), entonces $\bar{\boldsymbol{x}}$ es un punto KKT y las sucesiones $\{\boldsymbol{u}^t\}_{t\in\mathbb{N}}$ y $\{\boldsymbol{v}^t\}_{t\in\mathbb{N}}$ convergen a los multiplicadores de Lagrange asociados a $\bar{\boldsymbol{x}}$.*

- *Si además en $\bar{\boldsymbol{x}}$ se cumple la condición suficiente de KKT de segundo orden, entonces la sucesión $\{\rho^t\}_{t\in\mathbb{N}}$ está acotada.*

7.4 Programación lineal sucesiva

Se podría decir que los métodos de penalización vistos en la sección anterior, incluido el método de lagrangiano aumentado, intentan resolver los problemas de optimización no lineal con restricciones pasándolos a problemas sin restricciones (o con restricciones "fáciles"). En este sentido, las técnicas de programación lineal sucesiva que veremos en esta sección buscan algo parecido. Si los métodos de penalización buscan apoyarse en subproblemas no lineales sin restricciones, en esta sección los subproblemas seguirán teniendo restricciones, pero tanto éstas como la función objetivo serán lineales. En ambos casos pasamos a subproblemas más sencillos de resolver que el problema original.

Para empezar, recordemos algún concepto ya desarrollado en la Sección 5.3.8. A continuación presentamos la formulación general de un problema de programación no lineal y su versión

linealizada.

$$\begin{array}{rll} & \text{Problema P} & \\ \text{minimizar} & f(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 & i = 1, \ldots, m \\ & h_j(\boldsymbol{x}) = 0 & j = 1, \ldots, l \end{array} \qquad \begin{array}{rll} & \text{Problema LP}(\bar{\boldsymbol{x}}) & \\ \text{minimizar} & f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}) & \\ \text{sujeto a} & g_i(\bar{\boldsymbol{x}}) + \nabla g_i(\bar{\boldsymbol{x}})^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq 0 & i = 1, \ldots, m \\ & h_j(\bar{\boldsymbol{x}}) + \nabla h_j(\bar{\boldsymbol{x}})^\intercal (\boldsymbol{x} - \bar{\boldsymbol{x}}) = 0 & j = 1, \ldots, l. \end{array}$$

El Teorema 5.12 nos asegura que, en el caso de que todas las restricciones sean de desigualdad, es equivalente que un punto $\bar{\boldsymbol{x}}$ sea un KKT de P a que sea un óptimo de LP($\bar{\boldsymbol{x}}$). Además, por la forma en la que trabajaremos con las restricciones de igualdad, este resultado será suficiente para nosotros.

La idea del método de programación lineal sucesiva es muy natural. En cada iteración t tendremos un cierto iterante $\boldsymbol{x}^t$, resolveremos LP($\boldsymbol{x}^t$), obteniendo $\boldsymbol{x}^{t+1}$, y así sucesivamente. Si en algún momento tenemos que $\boldsymbol{x}^t$ es un óptimo de LP($\boldsymbol{x}^t$), entonces el algoritmo termina y el Teorema 5.12 nos garantiza que tenemos un punto KKT.

Aunque así expresada la idea es muy sencilla, hay varios posibles problemas que requieren modificar un poco la idea original. Uno de ellos, ilustrado en el siguiente ejemplo, es que LP($\boldsymbol{x}^t$) pueda no tener puntos factibles (aunque P sí los tenga).

Ejemplo 7.4. Consideremos el siguiente problema de optimización con restricciones:

$$\begin{array}{rl} \text{minimizar} & (x_1 - 5)^2 + (x_2 - 3)^2 \\ \text{sujeto a} & x_1^2 + x_2^2 - 2 = 0. \end{array}$$

Supongamos que aplicamos nuestra idea de programación lineal sucesiva y en una cierta iteración llegamos a $\boldsymbol{x}^t = (0, 0)$. En este caso, tendremos que la linealización de la restricción $x_1^2 + x_2^2 - 2 = 0$ se traduce en $-2 = 0$, que no se puede cumplir independientemente de los valores de x_1 y x_2. Es decir, tenemos que LP($\boldsymbol{x}^t$) tiene región factible vacía. ◇

No es difícil modificar la programación lineal sucesiva para asegurar que todos los iterantes de la sucesión $\{\boldsymbol{x}^t\}_{t \in \mathbb{N}}$ están bien definidos. La forma más habitual consiste en pasar a una versión *penalizada* del método, en el que las restricciones linealizadas se pasan penalizadas a la función objetivo, usando la penalización absoluta P^a. Para fijar notaciones, supondremos que el problema P es de la forma

$$\begin{array}{rll} & \text{Problema P} & \\ \text{minimizar} & f(\boldsymbol{x}) & \\ \text{sujeto a} & g_i(\boldsymbol{x}) \leq 0 & i = 1, \ldots, m \\ & h_j(\boldsymbol{x}) = 0 & j = 1, \ldots, l \\ & \boldsymbol{x} \in S = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}, & \end{array}$$

donde todas las restricciones lineales han sido incluidas en el conjunto S, pues con ellas no será necesario hacer modificación alguna. Consideremos ahora el problema P$^\rho$, la versión penalizada de P:

$$\text{Problema P}^\rho$$

$$\underset{\boldsymbol{x} \in S}{\text{minimizar}}\, f^\rho(\boldsymbol{x}) = f(\boldsymbol{x}) + \rho P(\boldsymbol{x}) = f(\boldsymbol{x}) + \rho\Big(\sum_{i=1}^{m} \max\{0, g_i(\boldsymbol{x})\} + \sum_{j=1}^{l} |h_j(\boldsymbol{x})| \Big).$$

El problema P^ρ es un problema penalizado con restricciones lineales. Además, el Teorema 7.2 nos asegura que si ρ es suficientemente grande el óptimo global de P^ρ es un óptimo global de P. En su momento vimos que el mayor problema de la penalización P^a es su carácter no diferenciable, y de ahí su poca aplicación práctica como parte del método de penalización exterior. A continuación vamos a ver que el problema P^ρ puede transformarse de una forma conveniente para el método de programación lineal sucesiva y que permite salvar la no diferenciabilidad de P^ρ.

La idea es aprovecharse del "truco" habitual en programación lineal para trabajar con máximos y valores absolutos en la función objetivo. Esto consiste en pasar estos máximos y valores absolutos a restricciones con la ayuda de variables artificiales. Para cada $i \in \{1, \dots, m\}$ definimos $y_i = \max\{0, g_i(\boldsymbol{x})\}$ y, para cada $j \in \{1, \dots, l\}$, tomamos variables $z_j^+ \geq 0$ y $z_j^- \geq 0$ tales que $|h_j(\boldsymbol{x})| = z_j^+ + z_j^-$. Ahora el problema P^ρ puede escribirse equivalentemente como

$$\begin{array}{rll} & \text{Problema } \mathrm{P}^\rho & \\ \text{minimizar} & f(\boldsymbol{x}) + \rho\Big(\displaystyle\sum_{i=1}^m y_i + \sum_{j=1}^l (z_j^+ + z_j^-)\Big) & \\ \text{sujeto a} & y_i \geq g_i(\boldsymbol{x}) & i = 1, \dots, m \\ & z_j^+ - z_j^- = h_j(\boldsymbol{x}) & j = 1, \dots, l \\ & y_i \geq 0 & i = 1, \dots, m \\ & z_j^+ \geq 0, z_j^- \geq 0 & j = 1, \dots, l \\ & \boldsymbol{x} \in S = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}. & \end{array}$$

Claramente, fijado $\rho > 0$, para cada $\boldsymbol{x} \in S$, la forma óptima de elegir $\boldsymbol{y}$, $\boldsymbol{z}^+$ y $\boldsymbol{z}^-$ será tomando $y_i = \max\{0, g_i(\boldsymbol{x})\}$ para cada $i \in \{1, \dots, m\}$ y, para cada $j \in \{1, \dots, l\}$, tendremos: $z_j^+ = h_j(\boldsymbol{x})$ y $z_j^- = 0$ si $h_j(\boldsymbol{x}) \geq 0$ y $z_j^+ = 0$ y $z_j^- = -h_j(\boldsymbol{x})$ si $h_j(\boldsymbol{x}) \leq 0$.

Nótese que este truco no sería de ninguna utilidad para el método de penalización exterior, pues volvemos a tener todas las funciones g_i y h_j como restricciones. La gran ventaja es que, una vez que las linealicemos, las variables y_i, z_j^+ y z_j^- nos asegurarán que la región factible de los subproblemas es no vacía.

Si ahora resolvemos el problema P^ρ y obtenemos una solución infactible de P, entonces no tendremos más que incrementar el valor de las penalizaciones y el Teorema 7.2 nos asegura que en algún momento obtendremos soluciones factibles de P (siempre que exista alguna).

Ya estamos en condiciones de presentar una primera versión del algoritmo de programación lineal sucesiva que se basa en resolver, sucesivamente, la versión linealizada de P^ρ, $\mathrm{LP}^\rho(\boldsymbol{x}^t)$ que viene dada por

$$\begin{array}{rcl} & & \text{Problema } \mathrm{LP}^\rho(\boldsymbol{x}^t) \\ \underset{\boldsymbol{x}\in S}{\text{minimizar}}\, f_{L,t}^\rho(\boldsymbol{x}) & = & f(\boldsymbol{x}^t) + \nabla f(\boldsymbol{x}^t)^\intercal(\boldsymbol{x} - \boldsymbol{x}^t) \\ & + & \rho \sum_{i=1}^m \max\{0, g_i(\boldsymbol{x}^t) + \nabla g_i(\boldsymbol{x}^t)^\intercal(\boldsymbol{x} - \boldsymbol{x}^t)\} \\ & + & \rho \sum_{j=1}^l \left|h_j(\boldsymbol{x}^t) + \nabla h_j(\boldsymbol{x}^t)^\intercal(\boldsymbol{x} - \boldsymbol{x}^t)\right| \end{array}$$

y que, por la discusión anterior, se puede expresar equivalentemente como

$$\begin{array}{rll} & \text{Problema } \mathrm{LP}^\rho(\boldsymbol{x}^t) & \\ \text{minimizar} & \nabla f(\boldsymbol{x}^t)^\intercal(\boldsymbol{x}-\boldsymbol{x}^t)+\rho\Big(\sum_{i=1}^m y_i+\sum_{j=1}^l (z_j^+ + z_j^-)\Big) & \\ \text{sujeto a} & y_i \geq g_i(\boldsymbol{x}^t)+\nabla g_i(\boldsymbol{x}^t)^\intercal(\boldsymbol{x}-\boldsymbol{x}^t) & i=1,\dots,m \\ & z_j^+ + z_j^- = h_j(\boldsymbol{x}^t)+\nabla h_j(\boldsymbol{x}^t)^\intercal(\boldsymbol{x}-\boldsymbol{x}^t) & j=1,\dots,l \\ & y_i \geq 0 & i=1,\dots,m \\ & z_j^+ \geq 0, z_j^- \geq 0 & j=1,\dots,l \\ & \boldsymbol{x}\in S=\{\boldsymbol{x}:\boldsymbol{A}\boldsymbol{x}\leq \boldsymbol{b}\}. & \end{array}$$

Nótese que hemos obviado el término constante $f(\boldsymbol{x}^t)$. El problema $\mathrm{LP}^\rho(\boldsymbol{x}^t)$ siempre tiene soluciones factibles. Por tanto, podemos apoyarnos en él para definir una primera versión del método de programación lineal sucesiva, que describimos en la Figura 7.5.

MÉTODO DE PROGRAMACIÓN LINEAL SUCESIVA (SIN REGIÓN DE CONFIANZA)

INICIALIZACIÓN

Elegir la penalización $\rho > 0$ y un parámetro de terminación $\varepsilon > 0$. Elegir un punto inicial $\boldsymbol{x}^1 \in S$. Definir $t = 1$.

PASO 1

Definir $\boldsymbol{x}^{t+1}$ como una solución óptima del problema $\mathrm{LP}^\rho(\boldsymbol{x}^t)$.

- Si $\frac{\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|}{\|\boldsymbol{x}^t\|+\varepsilon} \leq \varepsilon$, FIN. Devolvemos $\boldsymbol{x}^{t+1}$.
- En otro caso, reemplazar t por $t+1$. Repetir el PASO 1.

Figura 7.5: Método de programación lineal sucesiva sin región de confianza para el problema P^ρ.

A continuación presentamos dos resultados clave para justificar matemáticamente el método de programación lineal sucesiva:

Teorema 7.6. *Si un punto $\bar{\boldsymbol{x}}$ es un punto KKT del problema P^ρ y además es factible para el problema P, entonces $\bar{\boldsymbol{x}}$ es un punto KKT del problema P.*

Demostración. Ejercicio. □

••Ejercicio 7.2. Demuestra el Teorema 7.6. ◁

Teorema 7.7. *Si P tiene soluciones factibles, ρ es suficientemente grande y la sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ definida por el método de programación lineal sucesiva sin región de confianza converge a un punto $\bar{\boldsymbol{x}}$, entonces $\bar{\boldsymbol{x}}$ es un punto KKT del problema P.*

Demostración. El Teorema 5.12 nos asegura que, si la sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ converge a $\bar{\boldsymbol{x}}$, entonces $\bar{\boldsymbol{x}}$ es un KKT de P^ρ. Además, si ρ es suficientemente grande y P tiene soluciones factibles, $\bar{\boldsymbol{x}}$ será factible en P. Por último, el Teorema 7.6 asegura ahora que $\bar{\boldsymbol{x}}$ es un KKT de P. □

Antes de continuar presentamos un par de observaciones relativas a la implementación práctica de este algoritmo.

- En caso de que el algoritmo termine en un punto no factible, tendríamos que incrementar el valor de ρ y volver a aplicar el algoritmo (a partir de la última solución obtenida si se desea).

- Para que aplique el Teorema 7.2, la penalización ρ tiene que ser al menos tan grande como el mayor de los valores absolutos de cualquier multiplicador de Lagrange asociado con las restricciones $\boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}$ y $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$. Debido a esto, tiene sentido trabajar con $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{m+l})$, de tal manera que cada restricción tiene un parámetro de penalización específico.

La versión del método de programación lineal sucesiva presentada en la Figura 7.5 tiene dos importantes limitaciones, relacionadas entre sí, que hacen que en la práctica no se suela implementar de esa manera:

- En primer lugar, nada asegura que la función objetivo mejore iteración a iteración. Esto es porque $\boldsymbol{x}^{t+1}$ podría estar muy lejos de $\boldsymbol{x}^t$, con lo que la linealización alrededor de $\boldsymbol{x}^t$ podría ser muy mala aproximación del problema en $\boldsymbol{x}^{t+1}$.

- Así definido, el algoritmo puede no tener buenas propiedades de convergencia y, en particular, podría ciclarse, oscilando entre dos o más soluciones muy distantes entre sí. En este caso, todas estas soluciones son puntos de acumulación de $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$, pero nada garantiza que sean KKT del problema P^ρ. Esto es porque cada una de estas soluciones resuelve una linealización de P^ρ alrededor de un punto muy distante de dicha solución, con lo que, nuevamente, la aproximación dada por la linealización podría ser muy mala en la solución en cuestión.

El enfoque clásico para atajar los problemas que acabamos de describir consiste incluir una región de confianza en el método, de tal manera que aseguremos que $\boldsymbol{x}^{t+1}$ mejore $\boldsymbol{x}^t$ con respecto al problema P^ρ. Dicho de otra forma, buscaremos asegurar que $f^\rho(\boldsymbol{x}^{t+1}) \leq f^\rho(\boldsymbol{x}^t)$. Dado que en cada iteración del algoritmo resolvemos una aproximación lineal del problema P^ρ, es fácil ver que si $\hat{\boldsymbol{x}}$ representa el óptimo de la linealización alrededor de $\boldsymbol{x}^t$, entonces la dirección $\boldsymbol{d}^t = \hat{\boldsymbol{x}} - \boldsymbol{x}^t$ es una dirección de descenso de f^ρ. Por tanto, si obtenemos $\boldsymbol{x}^{t+1}$ moviéndonos suficientemente poco en esa dirección, tendremos que, efectivamente, $f^\rho(\boldsymbol{x}^{t+1}) \leq f^\rho(\boldsymbol{x}^t)$. Esto es justamente lo que garantiza el método de programación lineal sucesiva con región de confianza,

que se basa en trabajar con el problema

$$\begin{array}{rll} & \text{Problema LP}^\rho(\boldsymbol{x}^t, \boldsymbol{r}^t) & \\ \text{minimizar} & \nabla f(\boldsymbol{x}^t)^\intercal \boldsymbol{d} + \rho\Big(\sum_{i=1}^m y_i + \sum_{j=1}^l (z_j^+ + z_j^-)\Big) & \\ \text{sujeto a} & y_i \geq g_i(\boldsymbol{x}^t) + \nabla g_i(\boldsymbol{x}^t)^\intercal \boldsymbol{d} & i = 1, \ldots, m \\ & z_j^+ + z_j^- = h_j(\boldsymbol{x}^t) + \nabla h_j(\boldsymbol{x}^t)^\intercal \boldsymbol{d} & j = 1, \ldots, l \\ & y_i \geq 0 & i = 1, \ldots, m \\ & z_j^+ \geq 0, z_j^- \geq 0 & j = 1, \ldots, l \\ & -r_k^t \leq d_k \leq r_k^t & k = 1, \ldots, m \\ & \boldsymbol{x} \in S = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}. & \end{array}$$

Con respecto al problema $\text{LP}^\rho(\boldsymbol{x}^t)$ hemos añadido una región de confianza delimitada por el vector $\boldsymbol{r}^t$ y reemplazado los términos $(\boldsymbol{x} - \boldsymbol{x}^t)$ con $\boldsymbol{d}$. Por tanto, la minimización se hace con respecto a este vector $\boldsymbol{d}$ y los vectores $\boldsymbol{y}$, $\boldsymbol{z}^+$ y $\boldsymbol{z}^-$.

El método de programación lineal sucesiva con región de confianza, representado en la Figura 7.6, decide en cada iteración si aceptar o no $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \boldsymbol{d}^t$ dependiendo del decrecimiento en f^ρ en comparación con el decrecimiento estimado $f_{L,t}^\rho$. Además, también usa esta misma información para ajustar dinámicamente $\boldsymbol{r}^t$, aumentando y reduciendo el tamaño de la región de confianza. Más concretamente, trabaja con las diferencias

$$\Delta f_t^\rho = f^\rho(\boldsymbol{x}^t) - f^\rho(\boldsymbol{x}^t + \boldsymbol{d}^t) \quad y \quad \Delta f_{L,t}^\rho = f_{L,t}^\rho(\boldsymbol{x}^t) - f_{L,t}^\rho(\boldsymbol{x}^t + \boldsymbol{d}^t)$$

y con el ratio asociado

$$\gamma^t = \frac{\Delta f_t^\rho}{\Delta f_{L,t}^\rho}.$$

Claramente, como $\boldsymbol{d}^t = \boldsymbol{0}$ es una solución factible de $\text{LP}^\rho(\boldsymbol{x}^t, \boldsymbol{r}^t)$, $\Delta f_{L,t}^\rho \geq 0$. Por tanto, un valor negativo de γ^t nos indica un que la función f^ρ aumenta al pasar de $\boldsymbol{x}^t$ a $\boldsymbol{x}^t + \boldsymbol{d}^t$, con lo que no aceptamos $\boldsymbol{x}^t + \boldsymbol{d}^t$ como nuevo iterante. Asimismo, si γ^t es muy pequeño, quiere decir que la mejora en f^ρ es mucho más pequeña de lo estimado por $f_{L,t}^\rho$, con lo que tampoco se considera bueno el paso a $\boldsymbol{x}^t + \boldsymbol{d}^t$. En estas situaciones, se reducirá el tamaño de la región de confianza, limitando la distancia a la que nos podemos mover desde $\boldsymbol{x}^t$ para así mejorar el comportamiento de $\boldsymbol{x}^t + \boldsymbol{d}^t$ con respecto a la linealización $f_{L,t}^\rho$.

Para el método de programación lineal sucesiva con región de confianza de puede obtener el siguiente resultado:

Teorema 7.8. *Si ρ es suficientemente grande y $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ es la sucesión definida por el método de programación lineal sucesiva con región de confianza, entonces todo punto de acumulación de $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ es un punto KKT del problema P.*

Para terminar, presentamos varias observaciones relativas al comportamiento de este método.

- El Teorema 7.8 es sensiblemente más fuerte que el Teorema 7.7 pues, en este caso, aunque la sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ se ciclase entre dos o más puntos, todos ellos serían puntos de acumulación y, por tanto, puntos KKT.

MÉTODO DE PROGRAMACIÓN LINEAL SUCESIVA (CON REGIÓN DE CONFIANZA)

INICIALIZACIÓN

Elegir $\rho > 0$ y $\varepsilon > 0$. Elegir α_1, α_2 y α_3 con $0 < \alpha_1 < \alpha_2 < \alpha_3 < 1$ y $\beta \in (0, 1)$. Elegir $\boldsymbol{r}^{\text{mín}} > \mathbf{0}$. Elegir un punto inicial $\boldsymbol{x}^1 \in S$ y un vector inicial $\boldsymbol{r}^1 > \mathbf{0}$. Definir $t = 1$.

PASO 1

Definir $\boldsymbol{d}^t$ como una solución óptima del problema $\text{LP}^\rho(\boldsymbol{x}^t, \boldsymbol{r}^t)$.

- Si $\Delta f^\rho_{L,t} \leq \varepsilon$ y $\|\boldsymbol{d}^t\| < \varepsilon$, FIN. Devolvemos $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \boldsymbol{d}^t$.
- En otro caso se evalúa γ^t:
  - Si $\gamma^t < \alpha_1$, definir $\boldsymbol{r}^t = \beta \boldsymbol{r}^t$. Repetimos la iteración actual, ir al PASO 1.
  - Si $\gamma^t \geq \alpha_1$, ir al PASO 2.

PASO 2

Definir $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \boldsymbol{d}^t$. Reajustamos la región de confianza:

- Si $\alpha_1 \leq \gamma^t < \alpha_2$, definir $\boldsymbol{r}^{t+1} = \beta \boldsymbol{r}^t$.
- Si $\alpha_2 \leq \gamma^t < \alpha_3$, definir $\boldsymbol{r}^{t+1} = \boldsymbol{r}^t$.
- Si $\gamma^t > \alpha_3$, definir $\boldsymbol{r}^{t+1} = \frac{\boldsymbol{r}^t}{\beta}$.

En cualquier caso definir, para cada $k \in \{1, \ldots, n\}$, $r_k^{t+1} = \text{máx}\{r_k^{\text{mín}}, r_k^{t+1}\}$. Reemplazar t por $t + 1$. Repetir el PASO 1.

Figura 7.6: Método de programación lineal sucesiva con región de confianza.

- Además, en caso de que el algoritmo genere una sucesión contenida en un compacto, cosa que en la práctica se puede asegurar definiendo unas restricciones de caja adecuadas, entonces la sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ tendrá al menos un punto de acumulación, que será un punto KKT. Aunque a nivel teórico este resultado es bastante positivo, en la práctica, en el caso de que la sucesión $\{\boldsymbol{x}^t\}_{t\in\mathbb{N}}$ no converja, no es una tarea sencilla el identificar un punto de acumulación en la misma.

- En lo que respecta al criterio de parada, $\Delta f^\rho_{L,t} = 0$ y $\boldsymbol{d}^t = \mathbf{0}$ son condiciones suficientes para que $\boldsymbol{x}^t$ sea un KKT de P, con lo que teóricamente bastaría con expresar el criterio de parada con respecto a uno de los dos criterios usados.

- Una desventaja de este método con respecto a la versión sin región de confianza es que depende de la especificación de más parámetros: ρ, ε, α_1, α_2, α_3, $\boldsymbol{r}^{\text{mín}}$, β.[3]

- Como ya hemos comentado, el método de programación lineal sucesiva se puede interpretar como un método que busca direcciones de descenso basándose en aproximaciones lineales del problema original. Es por esto que, al igual que los métodos de gradiente para problemas sin restricciones, puede no tener buenas propiedades de convergencia cerca del óptimo, y que aparezca el típico fenómeno de zigzag. Es por esto que en la práctica es habitual usar técnicas que se basen en aproximaciones cuadráticas en vez de lineales, como discutimos brevemente en la siguiente sección.

[3] En particular, en el libro "Nonlinear Programming: Theory and Algorithms", escrito por Mokhtar S. Bazaraa, Hanif D. Sherali y C. M. Shetty, sugieren $\alpha_1 = 10^{-6}$, $\alpha_2 = 0.25$, $\alpha_3 = 0.75$ y $\beta = 0.5$. Además, también comentan otras formas de jugar con estos parámetros y también con el criterio de parada.

7.4.1 Programación cuadrática sucesiva

Para terminar esta sección presentamos las ideas detrás de una mejora de la programación lineal sucesiva, basada en trabajar con aproximaciones cuadráticas en vez de aproximaciones lineales, y que se conoce con el nombre de programación cuadrática sucesiva. Sin embargo, uno no puede simplemente reemplazar todas las aproximaciones de grado uno por aproximaciones de grado dos, pues es necesario que los subproblemas sigan siendo fáciles de resolver. El objetivo del método entonces será mantener la linealidad de las restricciones y que los términos cuadráticos de la función objetivo tengan asociada una matriz hessiana definida positiva, para así tener subproblemas cuadráticos de programación convexa con restricciones lineales para los que existen algoritmos bastante eficientes.[4]

Además, para que la aproximación cuadrática sea no sólo de la función f sino también de las restricciones $\boldsymbol{g}$ y $\boldsymbol{h}$, el método trabajará con la aproximación de grado dos de la función lagrangiana:

$$\nabla^2_{\boldsymbol{xx}} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = \nabla^2 f(\boldsymbol{x}) + \sum_{i=1}^{m} u_i \nabla^2 g_i(\boldsymbol{x}) + \sum_{j=1}^{l} \boldsymbol{v}_j \nabla^2 h_j(\boldsymbol{x}).$$

Por comodidad, cuando $\boldsymbol{u}$ y $\boldsymbol{v}$ estén fijos, denotaremos $\nabla^2_{\boldsymbol{xx}} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$ por $\nabla^2 L^P(\boldsymbol{x})$. Entonces, el método de programación cuadrática sucesiva se basará en resolver, en cada iteración t, un problema de la forma:

$$\begin{array}{rll} & \text{Problema QP}(\boldsymbol{x}^t, \boldsymbol{u}^t, \boldsymbol{v}^t) & \\ \text{minimizar} & f(\boldsymbol{x}^t) + \nabla f(\boldsymbol{x}^t)^\intercal \boldsymbol{d} + \frac{1}{2} \boldsymbol{d}^\intercal \nabla^2 L^P(\boldsymbol{x}^t) \boldsymbol{d} & \\ \text{sujeto a} & g_i(\boldsymbol{x}^t) + \nabla g_i(\boldsymbol{x}^t)^\intercal \boldsymbol{d} \leq 0 & i = 1, \ldots, m \\ & h_j(\boldsymbol{x}^t) + \nabla h_j(\boldsymbol{x}^t)^\intercal \boldsymbol{d} = 0 & j = 1, \ldots, l. \end{array}$$

Nótese que esta aproximación cuadrática además de incorporar información de primer orden de función objetivo y restricciones, también incorpora información de la curvatura de las mismas a través del término $\nabla^2 L^P(\boldsymbol{x}^t)$.

Debido al uso conjunto de la función lagrangiana y de aproximaciones cuadráticas en el espíritu del método de Newton, los métodos de programación cuadrática sucesiva también se conocen como métodos de Lagrange-Newton o de lagrangiano proyectado (para reflejar que la solución tiene que tomarse dentro de la región factible delimitada por la linealización de las restricciones).

Aunque la intuición detrás del método de programación lineal sucesiva es bastante natural, a la hora de implementarlo hay una serie de detalles que hay que tener en consideración:

- En cada iteración, al resolver el problema QP$(\boldsymbol{x}^t, \boldsymbol{u}^t, \boldsymbol{v}^t)$ minimizando con respecto de $\boldsymbol{x}$, se obtendrá un nuevo iterante $\boldsymbol{x}^{t+1}$. Pero también habrá que evaluar las condiciones de KKT del problema QP para obtener nuevos valores de los multiplicadores $\boldsymbol{u}^{t+1}$ y $\boldsymbol{v}^{t+1}$. Por tanto, este algoritmo genera una sucesión de soluciones del primal y también del dual.

[4] De hecho, los optimizadores punteros para problemas de programación lineal como Gurobi y CPLEX pueden resolver este tipo de problemas.

- Al igual que pasaba con el método de Newton, es preciso poder trabajar con matrices definidas positivas. Es por esto que la matriz $\nabla^2 L^P(\boldsymbol{x}^t)$ se reemplaza por aproximaciones cuasi-Newton, V^t, como las descritas en la Sección 4.8.3.

- Al igual que pasaba con el método de programación lineal sucesiva, $\mathrm{QP}(\boldsymbol{x}^t, \boldsymbol{u}^t, \boldsymbol{v}^t)$ puede tener región factible vacía incluso cuando P tiene soluciones factibles. Para evitar esto, el método de programación cuadrática sucesiva suele implementarse también en versión penalizada, apoyándose en la penalización absoluta P^a. Al igual que entonces, estas penalizaciones deberán ser "retransferidas" a las restricciones mediante las mismas transformaciones para tener subproblemas diferenciables.

- El método de programación cuadrática sucesiva también suele ir acompañado de una región de confianza.

- En la práctica es habitual que este método se complemente con una búsqueda de línea (al igual que se hacía con el método de Newton en problemas sin restricciones). Una vez que encuentra una dirección candidata $\boldsymbol{d}^t$, en vez de definir $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \boldsymbol{d}^t$, se suele hacer una búsqueda de línea para minimizar la función f^ρ en la dirección $\boldsymbol{d}^t$ (aunque hay que elegir cuidadosamente el método de búsqueda de línea, dado que f^ρ no es diferenciable).

Como puede verse, hay bastantes consideraciones a tener en cuenta para llevar a cabo una implementación solvente de un método de programación cuadrática sucesiva.

Tema 8
Optimización Global y heurísticas

Contenidos

PENDIENTE

Bibliografía

Armijo, L. (1966). Minimization of functions having lipschitz continuous first-partial derivatives. *Pacific Journal of Mathematics*, 16(1-3).

Barnhart, C., Johnson, E. L., Nemhauser, G. L., Savelsbergh, M. W. P., y Vance, P. H. (1998). Branch-and-price: Column generation for solving huge integer programs. *Operations Research*, 46, pp. 316–329.

Bazaraa, M. S., Jarvis, J. J., y Sherali, H. D. (2009). *Linear Programming and Network Flows*. John Wiley & Sons. 4th Edition.

Bazaraa, M. S., Sherali, H. D., y Shetty, C. M. (2006). *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons. 3rd Edition.

Ben-Tal, A. (1980). Second-order and related extremality conditions in nonlinear programming. *Journal of Optimization Theory and Applications*, 31, pp. 143–165.

Benders, J. F. (1962). Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4, pp. 238–252.

Boisvert, R. F., Howe, S. E., y Kahaner, D. K. (1985). GAMS: A framework for the management of scientific software. *ACM Trans. Math. Softw.*, 11, pp. 313–355.

Boser, B. E., Guyon, I. M., y Vapnik, V. (1992). A training algorithm for optimal margin classifiers. En Haussler, D., editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152. Pittsburgh, PA. ACM Press.

Broyden, C. G. (1967). Quasi-Newton methods and their application to function minimization. *Mathematics of Computation*, 21, pp. 368–381.

Broyden, C. G. (1970). The convergence of a class of double rank minimization algorithms: The new algorithm. *Journal of the Institute of Mathematics and Its Applications*, 6, pp. 222–231.

Conejo, A. J., Castillo, E., Minguez, R., y Garcia-Bertrand, R. (2006). *Decomposition Techniques in Mathematical Programming*. Springer.

Conn, A. R., Gould, N. I. M., y Toint, P. L. (2000). *Trust-Region Methods*. SIAM, Philadelphia, PA.

Cortes, C. y Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20, pp. 273–297.

Cournot, A. A. (1838). *Recherches sur les Principes Mathematiques de la Théorie des Richesses.* L. Hachette. Paris, France. (English translation by N. T. Bacon in Economic Classics, New York, NY: Macmillan, 1897; reprinted by Augustus M. Kelly, 1960).

Cristianini, N. y Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press.

Dantzig, G. B. y Wolfe, P. (1960). Decomposition principle for linear programs. *Operations Research*, 8, pp. 101–111.

Davidon, W. C. (1959). Variable metric method for minimization. Technical report, AEC Research Development Report ANL-5990.

Dono-Caramés, A. (2019). Problemas de regresión e clasificación usando SVM. Trabajo fin de grado, Facultad de Matemáticas. Universidad de Santiago de Compostela.

Dunning, I., Huchette, J., y Lubin, M. (2017). JuMP: A modeling language for mathematical optimization. *SIAM Review*, 59, pp. 295–320.

Everett, H. (1963). Generalized Lagrange multiplier method for solving problems of optimum allocation of resources. *Operations Research*, 11, pp. 399–417.

Fletcher, R. (1970). A new approach to variable metric algorithms. *Computer Journal*, 13, pp. 317–322.

Fletcher, R. (1987). *Practical Methods of Optimization.* Wiley.

Fletcher, R. y Powell, M. J. D. (1963). A rapidly convergent descent method for minimization. *Computer Journal*, 6, pp. 163–168.

Fletcher, R. y Reeves, C. (1964). Function minimization by conjugate gradients. *Computer Journal*, 7, pp. 149–154.

Fourer, R., Gay, D. M., y Kernighan, B. W. (1990). A modeling language for mathematical programming. *Management Science*, 36, pp. 519–554.

Geoffrion, A. M. (1972). Generalized Benders decomposition. *Journal of Optimization Theory and Application*, 10, pp. 237–260.

Geoffrion, A. M. (1974). *Approaches to Integer Programming*, volumen 2 de *Mathematical Programming Studies*, capítulo Lagrangean relaxation for integer programming. Springer, Berlin, Heidelberg.

Gilmore, P. y Gomory, R. (1961). A linear programming approach to the cutting stock problem. *Operations Research*, 9, pp. 849–859.

Gilmore, P. y Gomory, R. (1963). A linear programming approach to the cutting stock problem: Part II. *Operations Research*, 11, pp. 863–888.

Goldfarb, D. (1970). A family of variable metric methods derived by variational means. *Mathematics of Computation*, 24, pp. 23–26.

Gonzaga, C. C. (1990). Polynomial affine algorithms for linear programming. *Mathematical Programming*, 49, pp. 7–21.

González-Díaz, J., García-Jurado, I., y Fiestras-Janeiro, G. (2010). *An Introductory Course on Mathematical Game Theory*, volumen 115 de *Graduate Studies in Mathematics*. American Mathematical Society.

Hart, W. E., Laird, C. D., Watson, J.-P., Woodruff, D. L., Hackebeil, G. A., Nicholson, B. L., y Siirola, J. D. (2012). *Pyomo-Optimization Modeling in Python*. Springer Science & Business Media.

Held, M. y Karp, R. M. (1970). The traveling salesman problem and minimum spanning trees. *Operations Research*, 18, pp. 1138–1162.

Held, M. y Karp, R. M. (1971). The traveling salesman problem and minimum spanning trees: Part ii. *Mathematical Programming*, 1, pp. 6–25.

Held, M., Wolfe, P., y Crowder, H. (1974). Validation of subgradient optimization. *Mathematical Programming*, 6, pp. 62–88.

Hestenes, M. R. (1969). Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4, pp. 303–320.

Hestenes, M. R. y Stiefel, E. (1952). Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49, pp. 409–436.

Hooke, R. y Jeeves, T. A. (1961). Direct search solution of numerical and statistical problems. *Journal of the Association for Computing Machinery*, 8, pp. 212–229.

John, F. (1948). *Studies and Essays, Courant Anniversary Volume*, capítulo Extremum problems with inequalities as side conditions, pp. 187–204. Wiley (Interscience). K. O. Friedrichs, O. E. Neugebauer and J. J. Stoker (eds.).

Karush, W. (1939). Minima of functions of several variables with inequalities as side conditions. Tesis de máster, Department of Mathematics, University of Chicago.

Kiefer, J. (1953). Sequential minimax search for a maximum. En *Proceedings of the American Mathematical Society*, volumen 4, pp. 502–506.

Kuhn, H. W. y Tucker, A. W. (1951). Nonlinear programming. En Neyman, J., editor, *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, CA.

Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, 2, pp. 164–168.

Mangasarian, O. L. y Fromovitz, S. (1967). The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. *Journal of Mathematical Analysis and Applications*, 17, pp. 37–47.

Marquardt, D. W. (1963). An algorithm for least squares estimation of nonlinear parameters. *SIAM Journal of Industrial and Applied Mathematics*, 11, pp. 431–441.

McCormick, G. P. (1967). Second order conditions for constrained minima. *SIAM Journal on Applied Mathematics*, 15, pp. 641–652.

Nash, J. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America*, 36, pp. 48–49.

Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, 54, pp. 286–295.

Polak, E. y Ribière, G. (1969). Note sur la convergence de méthodes de directions conjuguées. *Revue Francaise Information Recherche Operationelle*, 16, pp. 35–43.

Polyak, B. T. (1967). A general method for solving extremum problems. *Soviet Mathematics*, 8, pp. 593–597.

Polyak, B. T. (1969). Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics (English translation)*, 9, pp. 14–29.

Powell, M. J. D. (1969). *Optimization*, capítulo A method for nonlinear constraints in minimization problems, pp. 283–298. Academic Press, New York.

Powell, M. J. D. (2003). On trust region methods for unconstrained minimization without derivatives. *Mathematical Programming B*, 97, pp. 605–623.

Rahmaniani, R., Crainic, T. G., Gendreau, M., y Rei, W. (2017). The Benders decomposition algorithm: A literature review. *European Journal of Operational Research*, 259, pp. 801–817.

Rockafellar, R. T. y Wets, R. J.-B. (1991). Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, 16, pp. 119–147.

Rosenbrock, H. H. (1960). An automatic method for finding the greatest or least value of a function. *Computer Journal*, 3, pp. 175–184.

Ruszczyński, A. (2006). *Nonlinear Optimization*. Princeton University Press.

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4, pp. 25–55.

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimizations. *Mathematics of Computation*, 24, pp. 641–656.

Vanderbeck, F. (2000). On dantzig-wolfe decomposition in integer programming and ways to perform branching in a branch-and-price algorithm. *Operations Research*, 48, pp. 111–128.

Vanderbeck, F. (2011). Branching in branch-and-price: a generic scheme. *Mathematical Programming*, 130(2), pp. 249–294.

Vanderbeck, F. y Wolsey, L. A. (1996). An exact algorithm for ip column generation. *Operations Research Letters*, 19, pp. 151–159.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, NY.

Vapnik, V. y Chervonenkis, A. (1964). A note on one class of perceptrons. *Automation and Remote Control*, 25.

Vapnik, V. y Chervonenkis, A. (1974). *Theory of Pattern Recognition [in Russian]*. Nauka. Moscow.

Vapnik, V. y Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24, pp. 774–780.

von Stackelberg, H. (1934). *Marktform und Gleichgewicht (Market Structure and Equilibrium)*. Springer-Verlag, Berlin-Vienna.

Wilde, D. J. (1964). *Optimum Seeking Methods*. Prentice-Hall.