Optimización semidefinida

Santiago Laplagne

Índice general

Prefacio

Estas notas fueron elaboradas para el curso de Optimización Semidefinida dictado en el Segundo Cuatrimestre de 2021 en la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires. La mayor parte del material en estas notas es traducción de otros textos, que se referencian al principio de la sección correspondiente. No hay ninguna intención de originalidad en estas notas sino que se presentan como una recopilación y traducción de material para facilitar la cursada a los alumnos y alumnas del curso.

Capítulo 1

Programación lineal

1.1. Introducción

Programación lineal es el problema de minimizar o maximizar una función lineal sujeta a restricciones lineales.

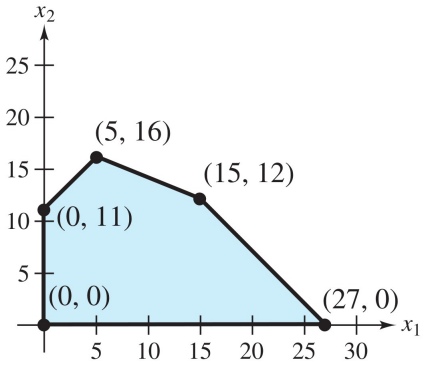
Comenzamos con un ejemplo simple ([Cen, , Capítulo 9.3]. Queremos maximizar la función

$$f(x_1, x_2) = 4x_1 + 6x_2$$

sujeta a las restricciones

$$\begin{aligned} -x_1 + x_2 &\leq 11 \\ x_1 + x_2 &\leq 27 \\ 2x_1 + 5x_2 &\leq 90 \\ x_1, x_2 &\geq 0. \end{aligned}$$

Cada una de las desigualdades en las restricciones define un semiplano en $\mathbb{R}^2$. Podemos graficar el conjunto de todos los puntos de $\mathbb{R}^2$ que cumplen todas las restricciones intersecando los semiplanos correspondientes.



El conjunto de puntos del plano para los cuales la función toma un valor fijo z_0 es una recta

$$l(z_0) = \{(x_1, x_2) : 4x_1 + 6x_2 = z_0\}.$$

Geométricamente, si variamos el valor de z_0, estamos desplazando la recta obteniendo siempre rectas paralelas. En este caso particular, vemos que si desplazamos la recta hacia arriba, el valor de z_0 aumenta, mientras que si desplazamos la recta hacia abajo, el valor de z_0 disminuye.

Por lo tanto, podemos resolver el problema gráficamente, desplazando la recta hacia arriba todo lo que podamos mientras que la intersección de la recta con la figura sea no vacía.

Mediante esta resolución gráfica, podemos observar algunas propiedades del problema de programación lineal. Si la región de puntos que satisfacen las restricciones es acotada, forma un polígono, y el óptimo de la función a optimizar se alcanza en el borde del polígono. Más precisamente en un vértice del polígono (puede suceder que el óptimo se alcance también sobre todo un lado del polígono). En particular, dado que un polígono tiene una cantidad finita de vértices, el problema de programación lineal puede resolverse algorítmicamente evaluando la función objetivo sobre todos los vértices del polígono.

Estudiamos ahora el problema en mayor generalidad. Para escribir una combinación lineal de variables vamos a utilizar la notación de producto interno

$$\boldsymbol{c} \cdot \boldsymbol{x} = c_1 x_1 + \cdots + c_n x_n,$$

para $\boldsymbol{c}, \boldsymbol{x} \in \mathbb{R}^n$. Un problema de programación lineal dado en forma estándar puede plantearse de la siguiente forma:

$$\begin{array}{ll} \text{minimizar:} & \boldsymbol{c} \cdot \boldsymbol{x} \\ \text{sujeto a:} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq 0, \end{array}$$

donde $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, $\boldsymbol{c} \in \mathbb{R}^n$ y la minimización se realiza sobre la variable de decisión $\boldsymbol{x} \in \mathbb{R}^n$. La condición $\boldsymbol{x} \geq 0$ se interpreta componente a componente, es decir $x_i \geq 0$, $1 \leq i \leq n$. Vamos a ver que un problema donde las restricciones vienen dadas por desigualdades puede llevarse a forma estándar agregando variables auxiliares apropiadas.

1.2. Conjuntos convexos y poliedros

Comenzamos con algunas definiciones básicas.

Definición 1.1. *Dado un vector no-nulo $\boldsymbol{a} \in \mathbb{R}^n$ y un escalar b,*

(a) el conjunto $\{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{a} \cdot \boldsymbol{x} = b\}$ es un hiperplano*,*

(b) el conjunto $\{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{a} \cdot \boldsymbol{x} \geq b\}$ es un semiespacio*.*

Dada una matriz $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ y un vector $\boldsymbol{b} \in \mathbb{R}^n$, en la condición $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, cada fila $\boldsymbol{a}_i$ de $\boldsymbol{A}$ impone la restricción $\boldsymbol{a}_i \cdot \boldsymbol{x} = b_i$. El conjunto $\{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ corresponde por lo tanto a una intersección de hiperplanos, que llamamos *espacio afín*. Igualmente, en la condición $\boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}$, cada fila $\boldsymbol{a}_i$ de $\boldsymbol{A}$ impone la restricción $\boldsymbol{a}_i \cdot \boldsymbol{x} \geq b_i$ y el conjunto $\{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}\}$ corresponde a una intersección de semiespacios.

Definición 1.2. *Un conjunto $S \subset \mathbb{R}^n$ es convexo si para todos $\boldsymbol{x}, \boldsymbol{y} \in S$, el segmento con vértices $\boldsymbol{x}$ e $\boldsymbol{y}$ está incluido en S. Es decir, para todo $\lambda \in [0, 1]$,*

$$\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y} \in S.$$

Dado un problema de programación lineal

$$\begin{array}{ll} \text{minimizar:} & \boldsymbol{c} \cdot \boldsymbol{x} \\ \text{sujeto a:} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq 0, \end{array}$$

podemos interpretar geométricamente el conjunto factible (*feasible set* en inglés), es decir la región sobre la que queremos minimizar $\boldsymbol{c}\cdot\boldsymbol{x}$, como la intersección de un espacio afín (definido por la ecuación $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$) y el ortante positivo $\boldsymbol{x} \geq 0$. Como los dos conjuntos son convexos y la intersección de conjuntos convexos es también convexa, el conjunto factible resulta convexo.

Definición 1.3. *Llamamos* poliedro *a un conjunto definido por igualdades y desigualdades lineales:*

$$P = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}_1\boldsymbol{x} = \boldsymbol{b}_1, \boldsymbol{A}_2\boldsymbol{x} = \boldsymbol{b}_2\}.$$

En el caso de que el conjunto resulte acotado, lo llamamos polítopo.

Como las definiciones varían según la literatura, es importante recordar entonces que para nosotros un poliedro no necesariamente es un conjunto acotado.

Ejercicio 1.4. Dado el problema

$$\begin{array}{ll} \text{minimizar:} & 3x_1 + 5x_2 \\ \text{sujeto a:} & x_1 + x_2 = 6 \\ & \boldsymbol{x} \geq 0, \end{array}$$

graficar el conjunto factible, y resolver el problema.

1.3. Variables de holgura

Dado el problema de programación lineal

$$\begin{array}{ll} \text{minimizar:} & \boldsymbol{c}\cdot\boldsymbol{x} \\ \text{sujeto a:} & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0}, \end{array}$$

podemos convertir el problema a un problema con restricciones de igualdad reemplazando cada desigualdad $\boldsymbol{a}_i \cdot \boldsymbol{x} \leq b_i$ por el par de restricciones

$$\boldsymbol{a}_i \cdot \boldsymbol{x} + s_i = b_i, \qquad s_i \geq 0,$$

donde s_i es una nueva variable del problema. Estas nuevas variables se llaman *variables de holgura*.

Obtenemos el problema equivalente con restricciones de igualdad

$$\begin{array}{ll} \text{minimizar:} & \boldsymbol{c}\cdot\boldsymbol{x} \\ \text{sujeto a:} & \boldsymbol{a}_1 \cdot \boldsymbol{x} + s_1 = b_1 \\ & \cdots \\ & \boldsymbol{a}_m \cdot \boldsymbol{x} + s_m = b_m \\ & \boldsymbol{x} \geq \boldsymbol{0}, \boldsymbol{s} \geq \boldsymbol{0}, \end{array}$$

con $\boldsymbol{s} = (s_1, \ldots, s_m)$.

Ejercicio 1.5. Dado el problema

$$\begin{array}{ll} \text{minimizar:} & 3x_1 + 5x_2 \\ \text{sujeto a:} & x_1 \leq 1 \\ & x_2 \leq 1 \\ & \boldsymbol{x} \geq 0, \end{array}$$

(a) graficar el conjunto factible,

(b) convertirlo a un problema con igualdades agregando las variables de holgura necesarias,

(c) calcular las coordenadas de los vértices del poliedro en el nuevo problema.

1.4. Puntos extremales, vértices y soluciones factibles básicas

Referencia: [Bertsimas and Tsitsiklis, 1997, Sección 2.2].

Ya vimos geométricamente que la solución óptima de un problema de programación lineal se encuentra en una *esquina* de la región factible. Veamos ahora diferentes formas de formalizar la idea de esquina.

Definición 1.6. *Dado un poliedro P, un vector $\boldsymbol{x} \in P$ es un* punto extremal *de P si no puede escribirse como combinación convexa de dos puntos de P distintos de $\boldsymbol{x}$. Es decir, si no existen dos vectores $\boldsymbol{y}, \boldsymbol{z} \in P$, ambos diferentes de $\boldsymbol{x}$, y un escalar $\lambda \in [0,1]$ tales que $\boldsymbol{x} = \lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{z}$.*

Otra posibilidad es considerar a los puntos que son solución óptima única de un problema de programación lineal.

Definición 1.7. *Dado un poliedro $P \subset \mathbb{R}^n$, un vector $\boldsymbol{x} \in P$ es un* vértice *de P si existe $\boldsymbol{c} \in \mathbb{R}^n$ tal que*

$$\boldsymbol{c} \cdot \boldsymbol{x} < \boldsymbol{c} \cdot \boldsymbol{y}$$

para todo $\boldsymbol{y} \in P$, $\boldsymbol{y} \neq \boldsymbol{x}$.

Es decir, $\boldsymbol{x}$ es un vértice de P si P está de un lado de un hiperplano que toca a P solo en $\boldsymbol{x}$.

Una desventaja de estas definiciones geométricas es que dado un poliedro P definido como intersección de hiperplanos y semiespacios, y un punto $\boldsymbol{x}$, no es fácil verificar si se cumplen las definiciones. Veremos a continuación una definición alternativa que podemos verificar fácilmente.

Consideramos un poliedro P definido por igualdades y desigualdades lineales,

$$\begin{cases} \boldsymbol{a}_i \cdot \boldsymbol{x} \geq b_i, & i \in M_1 \\ \boldsymbol{a}_i \cdot \boldsymbol{x} \leq b_i, & i \in M_2 \\ \boldsymbol{a}_i \cdot \boldsymbol{x} = b_i, & i \in M_3, \end{cases}$$

donde M_1, M_2, M_3 son conjuntos finitos de índices, $\boldsymbol{a}_i$ son vectores en $\mathbb{R}^n$ y b_i son escalares.

Definición 1.8. *Si un vector $\boldsymbol{x}^\star$ satisface una igualdad $\boldsymbol{a}_i \cdot \boldsymbol{x} = b_i$ para algún $i \in M_1$, M_2 o M_3, decimos que la condición correspondiente está* activa *en $\boldsymbol{x}^\star$.*

Si hay n condiciones activas, $\boldsymbol{x}^\star$ es solución de un sistema de n ecuaciones con n incógnitas. Si las n ecuaciones son linealmente independientes, el sistema tiene solución única. Teniendo esto en cuenta, hacemos la siguiente definición.

Definición 1.9. *Sean un poliedro P definido por restricciones de igualdades y desigualdades lineales y $\boldsymbol{x}^\star \in \mathbb{R}^n$.*

(a) El vector $\boldsymbol{x}^\star$ es una solución básica *si*

i. todas las igualdades están activas,

ii. de todas las restricciones activas, hay n de ellas que son linealmente independientes

(b) Si $\boldsymbol{x}^\star$ es una solución básica que satisface todas las restricciones, decimos que $\boldsymbol{x}^\star$ es una solución básica factible.

Observamos que en el conjunto de restricciones podemos reemplazar una igualdad $\boldsymbol{a} \cdot \boldsymbol{x} = b$ por dos desigualdades $\boldsymbol{a} \cdot \boldsymbol{x} \leq b$ y $\boldsymbol{a} \cdot \boldsymbol{x} \geq b$, obteniendo un problema equivalente. Por lo tanto, la condición de ser solución básica depende de cómo está formulado el problema.

Vimos hasta ahora tres formas de capturar el mismo concepto: punto extremal, vértice y solución básica factible. Veamos ahora que las tres definiciones son equivalentes.

Teorema 1.10. *Dado un poliedro P no vacío y un vector $\boldsymbol{x}^\star \in P$, las siguientes propiedades son equivalentes:*

(a) $\boldsymbol{x}^\star$ es un punto extremal de P,

(b) $\boldsymbol{x}^\star$ es un vértice de P,

(c) $\boldsymbol{x}^\star$ es una solución básica factible de P.

Demostración. Por simplicidad suponemos que el poliedro está definido solo por desigualdades $\boldsymbol{a}_i \cdot \boldsymbol{x} \geq b_i$ e igualdades $\boldsymbol{a}_i \cdot \boldsymbol{x} = b_i$.

Vértice ⇒ Punto extremal. Para un vértice $\boldsymbol{x}^\star \in P$, existe $\boldsymbol{c} \in \mathbb{R}^n$ tal que $\boldsymbol{c} \cdot \boldsymbol{x}^\star < \boldsymbol{c} \cdot \boldsymbol{y}$ para todo $\boldsymbol{y} \in P$, $\boldsymbol{y} \neq \boldsymbol{x}^\star$. Si tomamos $\boldsymbol{y}, \boldsymbol{z} \in P$, $\boldsymbol{y}, \boldsymbol{z} \neq \boldsymbol{x}^\star$, y $0 \leq \lambda \leq 1$, entonces como $\boldsymbol{c} \cdot \boldsymbol{x}^\star < \boldsymbol{c} \cdot \boldsymbol{y}$ y $\boldsymbol{c} \cdot \boldsymbol{x}^\star < \boldsymbol{c} \cdot \boldsymbol{z}$, tenemos que

$$\boldsymbol{c} \cdot \boldsymbol{x}^\star < \boldsymbol{c} \cdot (\lambda \boldsymbol{y} + (1 - \lambda)\boldsymbol{z})$$

y por lo tanto $\boldsymbol{x}^\star \neq \lambda \boldsymbol{y} + (1 - \lambda)\boldsymbol{z}$.

Punto extremal ⇒ Solución básica factible. Supongamos que $\boldsymbol{x}^\star \in P$ no es una solución básica factible. Vamos a ver que $\boldsymbol{x}^\star$ no puede ser un punto extremal de P. Como $\boldsymbol{x}^\star \in P$ es un punto factible, entonces $\boldsymbol{x}^\star \in P$ no es solución básica. Tomamos $I = \{i \mid \boldsymbol{a}_i \cdot \boldsymbol{x}^\star = b_i\}$, el conjunto de índices de las restricciones activas. Como $\boldsymbol{x}^\star$ no es una solución básica, no hay n vectores linealmente independientes en $\{\boldsymbol{a}_i\}_{i \in I}$. Si construimos una matriz $\boldsymbol{A}$ con estos vectores como fila, esta matriz tiene rango menor que n y por lo tanto existe $\boldsymbol{d} \in \mathbb{R}^n$ tal que $\boldsymbol{A}\boldsymbol{d} = \boldsymbol{0}$, es decir, $\boldsymbol{a}_i \cdot \boldsymbol{d}_i = 0$ para todo $i \in I$.

Para $i \notin I$, $\boldsymbol{a}_i \cdot \boldsymbol{x} > b_i$. Tomando ϵ suficientemente pequeño, los vectores $\boldsymbol{y} = \boldsymbol{x}^\star + \epsilon \boldsymbol{d}$ y $\boldsymbol{z} = \boldsymbol{x}^\star - \epsilon \boldsymbol{d}$ también van a cumplir $\boldsymbol{a}_i \cdot \boldsymbol{y} > b_i$, $\boldsymbol{a}_i \cdot \boldsymbol{z} > b_i$. Y por definición de $\boldsymbol{d}$, para $i \in I$, $\boldsymbol{a}_i \cdot \boldsymbol{y} = \boldsymbol{a}_i \cdot \boldsymbol{x}^\star + \epsilon \boldsymbol{d} = b_i$ y $\boldsymbol{a}_i \cdot \boldsymbol{z} = b_i$. Luego $\boldsymbol{y}, \boldsymbol{z} \in P$ y

$$\boldsymbol{x}^\star = \boldsymbol{y} + \boldsymbol{z},$$

por lo tanto $\boldsymbol{x}^\star$ no es un punto extremal.

Solución básica factible ⇒ Vértice. Sea $\boldsymbol{x}^\star$ una solución básica factible y sea $I = \{i \mid \boldsymbol{a}_i \cdot \boldsymbol{x}^\star = b_i\}$. Sea $\boldsymbol{c} = \sum_{i \in I} \boldsymbol{a}_i$. Tenemos

$$\boldsymbol{c} \cdot \boldsymbol{x}^\star = \sum_{i \in I} \boldsymbol{a}_i \cdot \boldsymbol{x}^\star = \sum_{i \in I} b_i.$$

Más aún, para cualquier $\boldsymbol{x} \in P$ y cualquier i, se cumple $\boldsymbol{a}_i \cdot \boldsymbol{x} \geq b_i$ y

$$\boldsymbol{c} \cdot \boldsymbol{x} = \sum_{i \in I} \boldsymbol{a}_i \cdot \boldsymbol{x} \geq \sum_{i \in I} b_i.$$

Por lo tanto $\boldsymbol{x}^\star$ es una solución óptima al problema de minimizar $\boldsymbol{c} \cdot \boldsymbol{x}$ sobre P. Para concluir, observamos que la igualdad en la última fórmula se cumple si y solo si $\boldsymbol{a}_i \cdot \boldsymbol{x} = b_i$ para todo $i \in I$. Como $\boldsymbol{x}^\star$ una solución básica factible, en el conjunto $\{\boldsymbol{a}_i\}_{i \in I}$ hay n vectores linealmente independientes y por lo tanto $\boldsymbol{x}^\star$ es la única solución del sistema de ecuaciones $\{\boldsymbol{a}_i \cdot \boldsymbol{x} = b_i\}_{i \in I}$. Luego $\boldsymbol{x}^\star$ es el único minimizante de $\boldsymbol{c} \cdot \boldsymbol{x}$ en P y por lo tanto es un vértice de P. □

Las dos primeras definiciones son definiciones geométricas y solo dependen del conjunto P y no de las ecuaciones que lo definen. Por la equivalencia, obtenemos que la condición de solución básica factible tampoco depende de las ecuaciones que definen P, a diferencia de las soluciones básicas que sí pueden depender.

Obtenemos el siguiente corolario simple pero muy importante.

Corolario 1.11. *Dado un conjunto finito de restricciones lineales (igualdades o desigualdades), la cantidad de soluciones básicas y soluciones básicas factibles es siempre finita.*

Demostración. Dado un sistema de m igualdades y desigualdades lineales, en una solución básica $\boldsymbol{x}^\star$ se cumplen al menos n de las restricciones. A la vez, las restricciones deben ser linealmente independientes, y por lo tanto $\boldsymbol{x}^\star$ es el único vector que satisface esas restricciones. Por lo tanto soluciones básicas distintas corresponden a distintos conjuntos de n restricciones. Como la cantidad posible de conjuntos de n restricciones es finita, la cantidad de soluciones básicas también. □

Observamos sin embargo que si bien la cantidad de soluciones básicas es siempre finita, puede ser una cantidad muy grande. Por ejemplo, el cubo

$$\{\boldsymbol{x} \in \mathbb{R}^n \mid 0 \leq x_i \leq 1, 1 \leq i \leq n\}$$

está definido por $2n$ ecuaciones y tiene 2^n soluciones básicas factibles. Esto hace que en la práctica, si bien podemos evaluar la función a optimizar en todas las soluciones básicas factibles para encontrar el óptimo, puede ser un método muy ineficiente.

1.5. Poliedros en forma estándar

Referencia: [Bertsimas and Tsitsiklis, 1997, Sección 2.3]

Un poliedro en forma estándar está definido por las siguientes ecuaciones:

$$P = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq 0\},$$

donde $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, es decir el poliedro está definido por m igualdades y n desigualdades $x_i \geq 0$. Eliminando filas redundantes de $\boldsymbol{A}$, podemos suponer que las m filas de $\boldsymbol{A}$ son linealmente independientes, y por lo tanto debe ser $m \leq n$.

Recordemos que en una solución básica debe haber n restricciones linealmente independientes activas, y más aún, todas las restricciones de igualdad se deben cumplir, lo que nos da m restricciones. Como $m \leq n$, para obtener n restricciones activas, debemos elegir $n - m$ variables x_i y darles valor 0, para activar las correspondientes $n - m$ desigualdades $x_i \geq 0$.

Debemos tener cuidado que no cualquier elección de las $n - m$ variables x_i nos va a dar un conjunto de n restricciones linealmente independientes. En el siguiente teorema vemos las condiciones que tenemos que cumplir.

Teorema 1.12. *Consideremos las restricciones $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ y $\boldsymbol{x} \geq 0$, donde suponemos que las m filas de $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ son linealmente independientes. Un vector $\boldsymbol{x}^\star \in \mathbb{R}^n$ es una solución básica si y solo si $\boldsymbol{A}\boldsymbol{x}^\star = \boldsymbol{b}$ y existen índices $B(1), \ldots, B(m)$ tales que*

(a) las columnas $\boldsymbol{A}_{B(1)}, \ldots, \boldsymbol{A}_{B(m)}$ son linealmente independientes,

(b) si $i \neq B(1), \ldots, B(m)$, entonces $x_i = 0$.

Este teorema nos da un procedimiento para construir soluciones básicas de un poliedro en forma estándar.

(a) Elegir m columnas linealmente independientes $\boldsymbol{A}_{B(1)}, \ldots, \boldsymbol{A}_{B(m)}$.

(b) Fijar $x_i = 0$ para todo $i \neq B(1), \ldots, B(m)$.

(c) Resolver el sistema de m ecuaciones $\boldsymbol{Ax} = \boldsymbol{b}$ para las variables $x_{B(1)}, \ldots, x_{B(m)}$.

Recordemos que en una matriz el rango fila y el rango columna coinciden, por lo tanto siempre podemos encontrar m columnas independientes en $\boldsymbol{A}$.

Si una solución básica construida siguiendo el procedimiento cumple que todas sus coordenadas son no-negativas, entonces es una solución básica factible. Recíprocamente, podemos encontrar todas las soluciones básicas factibles de esta forma.

Si $\boldsymbol{x}$ es una solución básica, llamamos *variables básicas* a las variables $x_{B(1)}, \ldots, x_{B(m)}$ y *no-básicas* a las demás variables. Llamamos *columnas básicas* a las columnas $\boldsymbol{A}_{B(1)}, \ldots, \boldsymbol{A}_{B(m)}$. Como son linealmente independientes, forman una base de $\mathbb{R}^m$.

Observamos que dos conjuntos distintos de variables básicas pueden dar la misma solución básica, si algunas de las variables básicas valen también 0. En este caso, decimos que la solución es degenerada. Para simplificar el desarrollo en este apunte, supondremos siempre que todas las soluciones básicas del problema son no-degeneradas.

1.6. Existencia de puntos extremales

En general los poliedros pueden no tener puntos extremales. Por ejemplo, un semiplano en $\mathbb{R}^2$ no tiene puntos extremales. Una condición muy simple para determinar si un poliedro tiene puntos extremales es la existencia o no de líneas rectas incluidas en el poliedro.

Decimos que un poliedro $P \subset \mathbb{R}^n$ contiene una recta si existe un vector $\boldsymbol{x} \in P$ y una dirección $\boldsymbol{d} \in \mathbb{R}^n$ tales que $\boldsymbol{x} + \lambda \boldsymbol{d} \in P$ para todo $\lambda \in \mathbb{R}$.

Teorema 1.13. *Dado un poliedro $P = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{a}_i \cdot \boldsymbol{x} \geq b_i, i = 1, \ldots, m\}$, las siguientes condiciones son equivalentes.*

(a) El poliedro P contiene al menos un punto extremal.

(b) El poliedro P no contiene ninguna recta.

(c) Existen n vectores linealmente independientes entre los vectores $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m$.

En particular, cualquier poliedro acotado tiene puntos extremales y cualquier poliedro en forma estándar tiene puntos extremales, debido a que el ortante positivo $\{x_i \geq 0, 1 \leq i \leq n\}$ no contiene ninguna recta.

Ahora que ya vimos condiciones para que un poliedro tenga puntos extremales, podemos preguntarnos en qué casos un problema de programación lineal tiene soluciones óptimas y verificar que se alcanzan en los puntos extremales, como vimos intuitivamente en los ejemplos.

Teorema 1.14. *Consideremos el problema de programación lineal de minimizar una funcional $\boldsymbol{c} \cdot \boldsymbol{x}$ sobre un poliedro P. Si P tiene al menos un punto extremal y el problema tiene solución óptima, entonces existe una solución óptima que es un punto extremal.*

Demostración. Completar □

Más aún, si el costo óptimo de la función a optimizar es finito, entonces siempre existe solución óptima.

Teorema 1.15. *Consideremos el problema de programación lineal de minimizar una funcional $\boldsymbol{c}\cdot\boldsymbol{x}$ sobre un poliedro P. Si P tiene al menos un punto extremal entonces o bien el costo óptimo es $-\infty$ o existe un punto extremal que es óptimo.*

Demostración. Ver [Bertsimas and Tsitsiklis, 1997, Teorema 2.8] □

1.7. El método Simplex

El método Simplex es uno de los métodos más usados en la práctica para resolver problemas de programación lineal. A partir de las herramientas desarrolladas en las secciones anteriores, podemos derivar el método simplex en forma sencilla.

Muchos métodos de optimización se basan en un principio simple que consiste en comenzar en una solución factible cualquiera e intentar moverse a otra solución factible cercana de forma que se reduzca el costo de la función a minimizar. Si no existe ninguna solución cercana que permita mejorar la función, hemos alcanzado un mínimo local. En general, un mínimo local no tiene por qué ser un mínimo global, una función podría tener varios mínimos locales, y el mínimo global ser solo uno de ellos.

Afortunadamente, en programación lineal un mínimo local es también global, dado que estamos minimizando una función convexa sobre un conjunto convexo. El método simplex se basa en aprovechar una propiedad adicional de la programación lineal. Ya vimos que en caso de existir un mínimo, este se alcanza en un vértice del poliedro de puntos factibles. Veremos que si estamos parados en un vértice, alcanza verificar si la función objetivo decrece al movernos a alguno de los vértices vecinos (que precisaremos más adelante). Si no decrece en ninguno de los vértices vecinos, hemos encontrado un mínimo local y por lo tanto global.

Obtenemos el siguiente algoritmo

(a) Elegir un vértice $\boldsymbol{x}$ del poliedro P.

(b) Para cada uno de los vértices vecinos $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s\}$ de $\boldsymbol{x}$ verificar si la función objetivo mejora en esos puntos.

(c) Si no mejora en ninguno de los puntos, $\boldsymbol{x}$ es un óptimo global y finalizamos el procedimiento.

(d) Si encontramos $\boldsymbol{y}_j$ tal que $\boldsymbol{c}\cdot\boldsymbol{y}_j < \boldsymbol{c}\cdot\boldsymbol{x}$, tomamos $\boldsymbol{x} = \boldsymbol{y}_j$ y volvemos a comenzar.

Veremos ahora como obtener los vértices vecinos de un vértice dado, y completaremos los detalles que faltan para probar la correctitud del algoritmo.

Para todo el desarrollo del método simplex, vamos a considerar el problema en forma estándar

$$\begin{aligned} &\text{minimizar:} && \boldsymbol{c}\cdot\boldsymbol{x} \\ &\text{sujeto a:} && \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & && \boldsymbol{x} \geq 0. \end{aligned}$$

1.7.1. Soluciones adyacentes

Dos soluciones básicas de un conjunto de restricciones lineales en $\mathbb{R}^n$ se dicen *adyacentes* si existen $n-1$ restricciones linealmente independientes que están activas en ambas soluciones. Si las dos soluciones son factibles, llamamos *arista* del conjunto factible al segmento que las une.

Ejercicio 1.16. Sea $P = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{a}_i \cdot \boldsymbol{x} \geq b_i, 1 \leq i \leq m\}$ un poliedro y $\boldsymbol{u}$, $\boldsymbol{v}$ dos soluciones básicas factibles adyacentes, con $\boldsymbol{a}_i \cdot \boldsymbol{u} = \boldsymbol{a}_i \cdot \boldsymbol{v} = b_i$ para $1 \leq i \leq n-1$ y $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{n-1}\}$ vectores linealmente independientes. Probar que el segmento $L = \{\lambda \boldsymbol{u} + (1-\lambda)\boldsymbol{v} \mid 0 \leq \lambda \leq 1\}$ que une $\boldsymbol{u}$ y $\boldsymbol{v}$ verifica

$$L = \{\boldsymbol{z} \in P \mid \boldsymbol{a}_i \cdot \boldsymbol{z} = b_i, 1 \leq i \leq n-1\}.$$

1.7.2. Direcciones básicas

En base a la definición de soluciones adyacentes, para movernos de una solución dada a una adyacente, podríamos reemplazar alguna de las variables básicas por una variable no-básica y calcular la nueva solución básica. Sin embargo, vamos a ver que al elegir cuál es la variable no-básica que vamos a convertir en básica, esto ya determina la dirección en la que debemos movernos para desplazarnos a una solución adyacente.

Sea $\boldsymbol{x}$ una solución básica factible de un problema en forma estándar y sean $B(1), \ldots, B(m)$ los índices de las variables básicas. Sea $\boldsymbol{B} = [\boldsymbol{A}_{B(1)} \cdots \boldsymbol{A}_{B(m)}] \in \mathbb{R}^{m \times m}$ la correspondiente matriz base. En particular $x_i = 0$ para cada variable no-básica y podemos calcular las coordenadas correspondientes a variables básicas por la fórmula

$$\boldsymbol{x}_B = \boldsymbol{B}^{-1}\boldsymbol{b}.$$

Para movernos a una solución adyacente en la que una variable no-básica x_j pase a ser variable básica, debemos desplazarnos a un nuevo vector

$$\boldsymbol{x} + \theta \boldsymbol{d}$$

con $\theta > 0$ y $\boldsymbol{d} = (d_1, \ldots, d_n)$ con $d_j = 1$ y $d_i = 0$ para todos los índices i distintos de j de variables no-básicas. El vector $\boldsymbol{x}_B$ de variables básicas va a cambiar a $\boldsymbol{x}_B + \theta \boldsymbol{d}_B$ con $\boldsymbol{d}_B = (d_{B(1)}, \ldots, d_{B(m)})$.

Como solo nos interesan las soluciones básicas factibles, debe cumplirse

$$\boldsymbol{A}(\boldsymbol{x} + \theta \boldsymbol{d}) = \boldsymbol{b},$$

y como $\boldsymbol{x}$ también es factible, $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Por lo tanto debe cumplirse $\boldsymbol{A}\boldsymbol{d} = \boldsymbol{0}$ y esto nos permite calcular $\boldsymbol{d}_B$. En efecto,

$$\boldsymbol{0} = \boldsymbol{A}\boldsymbol{d} = \sum_{i=1}^{n} \boldsymbol{A}_i d_i = \sum_{i=1}^{m} \boldsymbol{A}_{B(i)} d_{B(i)} + \boldsymbol{A}_j = \boldsymbol{B}\boldsymbol{d}_B + \boldsymbol{A}_j,$$

donde $\boldsymbol{A}_i$ es la i-ésima columna de $\boldsymbol{A}$.

Como la matriz $\boldsymbol{B}$ es inversible, obtenemos

$$\boldsymbol{d}_B = -\boldsymbol{B}^{-1}\boldsymbol{A}_j.$$

Llamamos al vector $\boldsymbol{d}$ que acabamos de construir como la j-ésima dirección básica. Por construcción, las restricciones de igualdad se van a mantener mientras nos vamos en esta dirección. Para las condiciones de no-negatividad de las variables, recordemos que la variable no-básica x_j aumenta y las demás variables no-básicas se mantienen en 0, por lo tanto conservan la no-negatividad. Para las variables básicas, como estamos suponiendo que todas las soluciones son no-degeneradas, las variables básicas son todas positivas y podemos tomar θ suficientemente pequeño para que se sigan cumpliendo.

Antes de calcular el valor de θ apropiado para movernos a una solución adyacente, calculamos cómo varía el costo de la función a optimizar al desplazarnos en la dirección de la j-ésima dirección básica.

Si $\boldsymbol{d}$ es la j-ésima dirección básica, la razón $\boldsymbol{c} \cdot \boldsymbol{d}$ de cambio del costo en la dirección $\boldsymbol{d}$ está dada por

$$\boldsymbol{c} \cdot \boldsymbol{d} = \boldsymbol{c}_B \cdot \boldsymbol{d}_B + c_j,$$

donde $\boldsymbol{c}_B = (c_{B(1)}, \ldots, c_{B(m)})$. Utilizando que $\boldsymbol{d}_B = -\boldsymbol{B}^{-1}\boldsymbol{A}_j$, hacemos la siguiente definición.

Definición 1.17. *Sea $\boldsymbol{x}$ una solución básica, sea $\boldsymbol{B}$ la matriz base asociada, y sea $\boldsymbol{c}_B$ el vector de costos de las variables básicas. Para cada j, definimos el* costo reducido *$\bar{c}_j$ de la variable x_j por la fórmula*

$$\bar{c}_j = c_j - \boldsymbol{c}_B \cdot (\boldsymbol{B}^{-1}\boldsymbol{A}_j).$$

El costo reducido nos dice cuánto varía la función de costo al movernos una unidad en la dirección de la j-ésima dirección básica. El término c_j indica el aumento del costo por unidad de la variable x_j y el último término es el costo de modificar las demás variables para compensar el cambio en la variable x_j.

Ejemplo 1.18. Consideremos el problema de programación lineal

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

con $\boldsymbol{c} = (2, 0, 0, 0) \in \mathbb{R}^4$,

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 0 & 3 & 4 \end{pmatrix} \in \mathbb{R}^{2\times 4} \quad \text{y} \quad \boldsymbol{b} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \in \mathbb{R}^2.$$

Las primeras dos columnas de $\boldsymbol{A}$ son $\boldsymbol{A}_1 = (1, 2)$ y $\boldsymbol{A}_2 = (1, 0)$. Como son linealmente independientes podemos elegir x_1 y x_2 como variables básicas. La matriz base correspondiente es

$$\boldsymbol{B} = \begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix}.$$

Tomamos $x_3 = x_4 = 0$, y calculamos los valores de x_1 y x_2: $x_1 = 1$ y $x_2 = 1$. Luego $\boldsymbol{x} = (1, 1, 0, 0)$ es una solución básica factible no-degenerada.

Construimos una dirección básica $\boldsymbol{d}$ correspondiente a aumentar el valor de la variable x_3. Tenemos $d_3 = 1$ y $d_4 = 0$. Las coordenedas de $\boldsymbol{d}$ correspondientes a las variables básicas las obtenemos por la fórmula

$$\begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} d_{B(1)} \\ d_{B(2)} \end{pmatrix} = \boldsymbol{d}_B = -\boldsymbol{B}^{-1}\boldsymbol{A}_3 = -\begin{pmatrix} 0 & 1/2 \\ 1 & -1/2 \end{pmatrix}\begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} -3/2 \\ 1/2 \end{pmatrix}.$$

Por lo tanto $\boldsymbol{d} = (-3/2, 1/2, 1, 0)$ y el costo de movernos en esta dirección es

$$\bar{c}_3 = \boldsymbol{c} \cdot \boldsymbol{d} = -\frac{3}{2}c_1 + \frac{1}{2}c_2 + c_3 = -3.$$

El siguiente resultado confirma la idea intuitiva que tenemos de los costos reducidos.

Teorema 1.19. *Sea $\boldsymbol{x}$ una solución básica factible asociada a una matriz base $\boldsymbol{B}$, y sea $\bar{\boldsymbol{c}}$ el vector de costos reducidos.*

(a) Si $\bar{\boldsymbol{c}} \geq 0$, entonces $\boldsymbol{x}$ es una solución óptima.

(b) Si $\boldsymbol{x}$ es una solución óptima no-degenerada, entonces $\bar{\boldsymbol{c}} \geq 0$.

Demostración.

(a) Suponemos $\bar{\boldsymbol{c}} \geq 0$, y consideramos una solución básica factible $\boldsymbol{y}$ arbitraria (es decir, $\boldsymbol{y}$ es un punto arbitrario del poliedro, no necesariamente un vértice). Definimos $\boldsymbol{d} = \boldsymbol{y} - \boldsymbol{x}$. Como $\boldsymbol{x}$ e $\boldsymbol{y}$ son factibles, $\boldsymbol{Ax} = \boldsymbol{Ay} = \boldsymbol{b}$ y por lo tanto $\boldsymbol{Ad} = \boldsymbol{0}$.

Separando las variables básicas del resto, reescribimos la última igualdad como

$$\boldsymbol{B}\boldsymbol{d}_B + \sum_{i \in N} \boldsymbol{A}_i d_i = \boldsymbol{0},$$

donde N es el conjunto de índices correspondientes a variables no-básicas. Como $\boldsymbol{B}$ es inversible,

$$\boldsymbol{d}_B = -\sum_{i \in N} \boldsymbol{B}^{-1}\boldsymbol{A}_i d_i,$$

y la variación del costo al movernos de $\boldsymbol{x}$ a $\boldsymbol{y}$ es

$$\begin{aligned}
\boldsymbol{c}\cdot\boldsymbol{y} - \boldsymbol{c}\cdot\boldsymbol{x} = \boldsymbol{c}\cdot\boldsymbol{d} &= \boldsymbol{c}_B \cdot \boldsymbol{d}_B + \sum_{i \in N} c_i d_i \\
&= -\boldsymbol{c}_B \cdot \sum_{i \in N} \boldsymbol{B}^{-1}\boldsymbol{A}_i d_i + \sum_{i \in N} c_i d_i \\
&= \sum_{i \in N} (c_i - \boldsymbol{c}_B \cdot (\boldsymbol{B}^{-1}\boldsymbol{A}_i)) d_i = \sum_{i \in N} \bar{c}_i d_i.
\end{aligned}$$

Para todo índice no-básico $i \in N$, tenemos $x_i = 0$ y como $\boldsymbol{y}$ es factible, $y_i \geq 0$. Por lo tanto, $d_i = y_i - x_i \geq 0$ y $\bar{c}_i d_i \geq 0$. Concluimos que $\boldsymbol{c}\cdot\boldsymbol{y} \geq \boldsymbol{c}\cdot\boldsymbol{x}$ y como $\boldsymbol{y}$ era arbitrario, $\boldsymbol{x}$ es óptimo.

(b) Si $\boldsymbol{x}$ es una solución óptima y existe una coordenada $\bar{c}_j > 0$ del vector de costos reducidos, podemos desplazarnos en la dirección básica correspondiente y reducir el costo, lo que contradice la optimalidad de $\boldsymbol{x}$.

□

En base a este teorema, para determinar si una solución básica factible no-degenerada es óptima, solo necesitamos chequear si todos los costos reducidos son no-negativos, lo que equivale a examinar si aumenta el costo al movernos en cada una de las $n - m$ direcciones básicas.

Por último, si encontramos una dirección básica en la cual movernos para reducir el costo, debemos calcular el mayor valor de θ que podemos tomar, lo que equivale a determinar

$$\theta^\star = \text{máx}\{\theta \geq 0 \mid \boldsymbol{x} + \theta\boldsymbol{d} \in P\}.$$

Vamos a deducir una fórmula para $\theta^\star$. Como $\boldsymbol{Ad} = \boldsymbol{0}$, tenemos que

$$\boldsymbol{A}(\boldsymbol{x} + \theta\boldsymbol{d}) = \boldsymbol{Ax} = \boldsymbol{b}$$

para todo θ, por lo tanto las restricciones de igual se cumplen siempre. El punto $\boldsymbol{x} + \theta\boldsymbol{d}$ solo puede salirse del poliedro si alguna de las coordenadas se vuelve negativa. Distinguimos dos casos:

(a) Si $\boldsymbol{d} \geq 0$ (es decir, $d_i \geq 0$ para todo $1 \leq i \leq n$), entonces $\boldsymbol{x} + \theta\boldsymbol{d} \geq 0$ para todo $\theta \geq 0$. El vector $\boldsymbol{x} + \theta\boldsymbol{d}$ nunca se vuelve no-factible, y tomamos $\theta^\star = +\infty$. El costo óptimo es $-\infty$.

(b) Si $d_i < 0$ para algunos i, la restricción $x_i + \theta d_i > 0$ nos da la condición

$$\theta \leq -\frac{x_i}{d_i}.$$

Estas restricciones se deben cumplir para todo i tal que $d_i < 0$. Por lo tanto el mayor valor de θ que podemos tomar es

$$\theta^\star = \min_{\{i|d_i<0\}} \left(-\frac{x_i}{d_i}\right).$$

Recordemos que para las variables x_i no-básicas, o bien $d_i = 1$ si es la variable que pasa a ser variable básica, o bien $d_i = 0$, y en ambos casos $d_i \geq 0$. Por lo tanto podemos restringirnos a mirar las variables básicas en la fórmula anterior. Obtenemos

$$\theta^\star = \min_{\{i=1,\ldots,m|d_{B(i)}<0\}} \left(-\frac{x_{B(i)}}{d_{B(i)}}\right). \tag{1.1}$$

Más aún, como supusimos que todas las soluciones son no-degeneradas, $x_i > 0$ para todas las variables básicas, y por lo tanto obtenemos siempre $\theta^\star > 0$.

Ejemplo 1.20. Continuamos el Ejemplo 1.18. Obtuvimos $\boldsymbol{x} = (1, 1, 0, 0)$, $\boldsymbol{d} = (-3/2, 1/2, 1, 0)$ y $\bar{c}_3 = -3$. Como $\bar{c}_3$ es negativo, podemos disminuir el costo de la función objetivo desplazándonos en esta dirección. Es decir, consideramos los vectores

$$\boldsymbol{x} + \theta \boldsymbol{d},$$

con $\theta > 0$. La única coordenada que decrece al desplazarnos es x_1, por lo que si tomamos cualquier valor de $\theta > 0$ tal que la primera coordenada se mantenga no-negativa, estaremos dentro del poliedro. El máximo valor de θ que podemos tomar es

$$\theta^\star = -\frac{x_1}{d_1} = \frac{2}{3},$$

y nos desplazamos a

$$\boldsymbol{y} = \boldsymbol{x} + \frac{2}{3}\boldsymbol{d} = \left(0, \frac{4}{3}, \frac{2}{3}, 0\right).$$

Las columnas correspondientes a las coordenadas no-nulas son $\boldsymbol{A}_2 = (1, 0)$ y $\boldsymbol{A}_3 = (1, 3)$, y son linealmente independientes. Por lo tanto $\boldsymbol{y}$ es una nueva solución básica factible. Como queríamos minimizar el valor de x_1, y tenemos la restricción $x_1 \geq 0$, hemos encontrado el costo óptimo. Para verificación, calculamos el costo reducido $\bar{c}_4$ por la fórmula

$$\bar{c}_4 = c_4 - \boldsymbol{c}_B \cdot (\boldsymbol{B}^{-1}\boldsymbol{A}_4) = 1 - \begin{pmatrix}1 & 0\end{pmatrix} \begin{pmatrix}1 & 1\\ 0 & 3\end{pmatrix}^{-1} \begin{pmatrix}1\\ 4\end{pmatrix} = \frac{4}{3}.$$

Vemos que movernos en la dirección básica correspondiente va a incrementar el valor de la función objetivo y por lo tanto no debemos desplazarnos en esa dirección.

Una vez que determinamos $\theta^\star$, suponiendo que es finito, nos movemos a una nueva solución factible

$$\boldsymbol{y} = \boldsymbol{x} + \theta^\star \boldsymbol{d}.$$

Por la construcción que hicimos de $\boldsymbol{d}$ y $\theta^\star$, si ℓ es el índice que minimiza (1.1), se cumple

$$x_{B(\ell)} = 0 \quad \text{y } x_j = \theta^\star.$$

Es decir que x_j reemplaza a $x_{B(\ell)}$ en el conjunto de variables básicas. Obtenemos un nuevo conjunto de índices de las variables básicas $\bar{B} = \{\bar{B}(1), \ldots, \bar{B}(m)\}$ definidos por

$$\bar{B}(i) = \begin{cases} B(i), & i \neq \ell, \\ j, & i = \ell. \end{cases}$$

Llamamos $\bar{\boldsymbol{B}}$ a la nueva matriz base, formada por las columnas de $\boldsymbol{A}$ correspondientes a las nuevas variables básicas.

Teorema 1.21.

(a) Las columnas $\boldsymbol{A}_{B(i)}$, $i \neq \ell$, y $\boldsymbol{A}_j$ son linealmente independientes y, por lo tanto, $\bar{B}$ es una matriz base.

(b) El vector $\boldsymbol{y} = \boldsymbol{x} + \theta^\star \boldsymbol{d}$ es una solución básica factible con matriz base asociada $\bar{\boldsymbol{B}}$.

Demostración. Ver [Bertsimas and Tsitsiklis, 1997, Teorema 3.2]. □

Mediante el procedimiento descripto, pasamos de una solución básica factible a otra solución básica factible con menor costo. Obtenemos la siguiente iteración del método simplex.

(1) Comenzamos con una solución básica factible $\boldsymbol{x}$, correspondiente a variables básicas con índices $I = \{B(1), \ldots, B(m)\}$.

(2) Tomamos la submatriz $\boldsymbol{B}$ de $\boldsymbol{A}$ formada por las columnas correspondientes a las variables básicas, y calculamos los costos reducidos

$$\bar{c}_j = c_j - \boldsymbol{c}_B \cdot (\boldsymbol{B}^{-1} \boldsymbol{A}_j).$$

Si todos los costos son no-negativos, $\boldsymbol{x}$ es una solución óptima y terminamos. Si no, elegimos algún j tal que $\bar{c}_j < 0$.

(3) Calculamos $\boldsymbol{u} = -\boldsymbol{d}_B = \boldsymbol{B}^{-1} \boldsymbol{A}_j$. Si $\boldsymbol{u}$ no tiene ninguna componente positiva, tomamos $\theta^\star = \infty$, el costo óptimo es $-\infty$ y el algoritmo termina.

(4) Si alguna componente de $\boldsymbol{u}$ es positiva, tomamos

$$\theta^\star = \min_{\{i=1,\ldots,m | u_i > 0\}} \frac{x_{B(i)}}{u_i}.$$

(5) Sea ℓ un índice para el cuál se alcanza el mínimo. Construimos una nueva base reemplazando $B(\ell)$ por j. La nueva variable básica $\boldsymbol{y}$ queda definida por

$$\begin{cases} y_j = \theta^\star, \\ y_{B(i)} = x_{B(i)} - \theta^\star u_i, & 1 \leq i \leq m, i \neq l \\ y_i = 0 & \text{para los índices de variables no-básicas.} \end{cases}$$

En base a todo lo que vimos es fácil verificar que el método simplex descripto resuelve correctamente el problema de programación lineal. En particular, esto nos permite demostrar la siguiente propiedad de los problemas de programación lineal.

Teorema 1.22. *Consideremos el problema de programación lineal de minimizar una funcional $\boldsymbol{c} \cdot \boldsymbol{x}$ sobre un poliedro P. Si P tiene al menos un punto extremal entonces o bien el costo óptimo es $-\infty$ o existe un punto extremal que es óptimo.*

1.7.3. El tableau del método Simplex

Veremos ahora una implementación del método simplex utilizando mediante operaciones elementales de filas de una matriz, que suele llamarse *tableau*.

Vamos a verlo en el siguiente ejemplo concreto:

$$\begin{aligned} \text{maximizar:} \quad & 4x_1 + 6x_2 = z \\ \text{sujeto a:} \quad & -x_1 + x_2 \leq 11 \\ & x_1 + x_2 \leq 27 \\ & 2x_1 + 5x_2 \leq 90 \\ & \boldsymbol{x} \geq 0. \end{aligned}$$

Como tenemos todas desigualdades menor o igual, agregamos variables de holgura s_1, s_2, s_3 y reemplazamos las desigualdades por igualdades:

$$\begin{aligned} \text{maximizar:} \quad & 4x_1 + 6x_2 = z \\ \text{sujeto a:} \quad & -x_1 + x_2 + s_1 = 11 \\ & x_1 + x_2 + s_2 = 27 \\ & 2x_1 + 5x_2 + s_3 = 90 \\ & \boldsymbol{x}, \boldsymbol{s} \geq 0. \end{aligned}$$

Como ya vimos, una *solución básica* de un problema en forma estándar es una solución

$$(x_1, x_2, \ldots, x_n, s_1, s_2, \ldots, s_m)$$

en la que *a lo sumo* m variables son nulas. Las variables no-nulas se llaman *variables básicas* de la solución. Si todas las variables son no-negativas, la solución se llama *solución básica factible*.

La matriz inicial del método se construye colocando todas las ecuaciones de restricciones como filas, y una última fila con los coeficientes de la función objetivo escrita de la siguiente forma:

$$-c_1x_1 - c_2x_2 - \cdots - c_nx_n + 0s_1 + \cdots + 0s_m + z = 0.$$

En general omitimos el coeficiente de z que se mantiene siempre en 1.

En el ejemplo, obtenemos el tableau

x_1	x_2	s_1	s_2	s_3	b	
-1	1	1	0	0	11	s_1
1	1	0	1	0	27	s_2
2	5	0	0	1	90	s_3
-4	-6	0	0	0	0	

Las variables básicas al comenzar son s_1, s_2, s_3 y las variables x_1 y x_2 comienzan con valor 0.

La solución básica inicial es

$$x_1 = 0, x_2 = 0, s_1 = 11, s_2 = 27, s_3 = 90$$

que en este caso resulta una solución básica factible.

En este momento podemos verificar si la solución que tenemos es óptima. La condición del método simplex para una solución óptima es que todos los números en la última fila sean no-negativos. En nuestro caso, tenemos dos coeficientes negativos y por lo tanto la solución no es óptima.

Pivoteo

Como la solución no es óptima, realizamos operaciones de pivoteo para obtener en sucesivos pasos mejores soluciones hasta llegar a la solución óptima.

Elegimos una nueva variable básica, que va a reemplazar a alguna de las variables básicas actuales. Definimos

- Variable entrante a la variable con el menor valor (el más negativo) en la última fila, que tomamos como nueva variable básica.
- Variable saliente a la variable con menor cociente b_i/a_{ij} en la columna determinada por la variable entrante.
- El pivot es la casilla de la matriz en la columna de la variable entrante y la fila de la variable saliente.

Finalmente, realizamos eliminación de Gauss-Jordan por filas sobre toda la matriz, tomando el pivot recién definido.

En nuestro ejemplo, la variable entrante es x_2 y la variable saliente es s_1.

Observaciones:

- La operación de pivoteo se corresponde con aumentar el valor de alguna de las variables que tenían valor 0. La elección de la variable entrante se hace de forma tal que un cambio en esta variable produzca un cambio mayor que el de las otras variables.
- Para elegir la variable saliente, buscamos cuál es el máximo valor que podemos darle a la variable entrante sin violar las restricciones. En nuestro ejemplo, el mayor valor es $x_2 = 11$. Este valor nos permite encontrar soluciones factibles (x_1, x_2) para las 3 restricciones. Si le diéramos un valor mayor, se violaría la primera restricción.

Al realizar el pivoteo, obtenemos el tableau

x_1	x_2	s_1	s_2	s_3	b	
-1	1	1	0	0	11	x_2
2	0	-1	1	0	16	s_2
7	0	-5	0	1	35	s_3
-10	0	6	0	0	66	

.

Vemos que obtuvimos una solución mejorada

$$(x_1, x_2, s_1, s_2, s_3) = (0, 11, 0, 16, 35)$$

y el valor de la función objetivo es

$$4x_1 + 6x_2 = 66.$$

Esta solución no es óptima porque x_1 tiene signo negativo en la ultima fila, podemos aumentar el valor de x_1.

Tomamos x_1 como variable saliente y calculando los cocientes, tomamos s_3 como variable saliente.

Al aplicar Gauss-Jordan obtenemos

x_1	x_2	s_1	s_2	s_3	b	
0	1	2/7	0	1/7	16	x_2
0	0	3/7	1	−2/7	6	s_2
1	0	−5/7	0	1/7	5	s_3
0	0	−8/7	0	10/7	116	

.

Esta solución no es óptima porque s_3 tiene signo negativo en la ultima fila, podemos aumentar el valor de s_3. Tomando s_3 como variable entrante y s_2 como variable saliente, obtenemos luego de pivotear:

x_1	x_2	s_1	s_2	s_3	b	
0	1	0	−2/3	1/3	12	x_2
0	0	1	7/3	−2/3	14	s_2
1	0	0	5/3	−1/3	15	x_1
0	0	0	8/3	2/3	132	

.

Encontramos la solución óptima del problema

$$(x_1, x_2, s_1, s_2, s_3) = (15, 12, 0, 14, 0)$$

con valor de la función objetivo

$$4x_1 + 6x_2 = 132.$$

1.8. El problema dual

Referencia principal: [Bertsimas and Tsitsiklis, 1997, Capítulo 4].

Una de las propiedades más interesantes de la programación lineal es que a cada problema primal en forma estándar

$$\begin{aligned} \text{maximizar:} \quad & \boldsymbol{c} \cdot \boldsymbol{x} \\ \text{sujeto a:} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq 0 \end{aligned}$$

podemos asociarle un problema dual de la siguiente forma:

$$\begin{aligned} \text{maximizar:} \quad & \boldsymbol{y} \cdot \boldsymbol{b} \\ \text{sujeto a:} \quad & \boldsymbol{A}^T \boldsymbol{y} \leq \boldsymbol{c}, \end{aligned}$$

que guarda relaciones muy fuertes con el problema primal.

1.8.1. Motivación

Consideremos el problema en forma estándar

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

que llamamos *problema primal*, y notamos $\boldsymbol{x}^\star$ a una solución óptima que suponemos que existe.

Podemos relajar este problema reemplazando la restricción $\boldsymbol{Ax} = \boldsymbol{b}$ por una penalidad $\boldsymbol{y} \cdot (\boldsymbol{b} - \boldsymbol{Ax})$, donde $\boldsymbol{y}$ es un vector de precios del mismo tamaño que $\boldsymbol{b}$. Obtenemos entonces el siguiente problema:

$$\begin{aligned} &\text{minimizar:} && \boldsymbol{c} \cdot \boldsymbol{x} + \boldsymbol{y} \cdot (\boldsymbol{b} - \boldsymbol{Ax}), \\ &\text{sujeto a:} && \boldsymbol{x} \geq \boldsymbol{0}. \end{aligned}$$

Llamamos $g(\boldsymbol{y})$ al costo óptimo en función de $\boldsymbol{y}$ del problema relajado. Como tenemos más libertad en la elección de variables en el problema relajado, esperamos que $g(\boldsymbol{y})$ sea menor o igual que el costo óptimo $\boldsymbol{c} \cdot \boldsymbol{x}^\star$ del problema primal. En efecto,

$$g(\boldsymbol{y}) = \min_{\boldsymbol{x} \geq 0} \{\boldsymbol{c} \cdot \boldsymbol{x} + \boldsymbol{y} \cdot (\boldsymbol{b} - \boldsymbol{Ax})\} \leq \boldsymbol{c} \cdot \boldsymbol{x}^\star + \boldsymbol{y} \cdot (\boldsymbol{b} - \boldsymbol{Ax}^\star) = \boldsymbol{c} \cdot \boldsymbol{x}^\star,$$

donde usamos para la última desigualdad que $\boldsymbol{x}^\star$ es una solución óptima del problema primal y por lo tanto $\boldsymbol{Ax}^\star = \boldsymbol{b}$.

Por lo tanto, para cualquier vector $\boldsymbol{y}$, $g(\boldsymbol{y}) \leq \boldsymbol{c} \cdot \boldsymbol{x}^\star$ y el valor de $g(\boldsymbol{y})$ es una cota inferior del costo óptimo $\boldsymbol{c} \cdot \boldsymbol{x}^\star$ del problema primal. Queremos hallar la mejor cota posible, es decir, queremos hallar el vector $\boldsymbol{y}$ que maximice $g(\boldsymbol{y})$. Este problema se conoce como el *problema dual*, que podemos plantear en la forma:

$$\begin{aligned} &\text{maximizar:} && g(\boldsymbol{y}), \\ &\text{sujeto a:} && \text{ninguna restricción.} \end{aligned}$$

El resultado principal de la teoría de dualidad afirma que el costo óptimo del problema dual es igual al costo óptimo $\boldsymbol{c} \cdot \boldsymbol{x}^\star$ del problema primal. En otras palabras, si elegimos el valor de $\boldsymbol{y}$ para el cual $g(\boldsymbol{y})$ es máximo, la posibilidad de violar las restricciones $\boldsymbol{Ax} = \boldsymbol{b}$ no es de ninguna ayuda.

Por la definición de $g(\boldsymbol{y})$,

$$\begin{aligned} g(\boldsymbol{y}) = \min_{\boldsymbol{x} \geq 0} \{\boldsymbol{c} \cdot \boldsymbol{x} + \boldsymbol{y} \cdot (\boldsymbol{b} - \boldsymbol{Ax})\} &= \boldsymbol{y} \cdot \boldsymbol{b} + \min_{\boldsymbol{x} \geq 0} \{\boldsymbol{c} \cdot \boldsymbol{x} - \boldsymbol{y} \cdot (\boldsymbol{Ax})\} \\ &= \boldsymbol{y} \cdot \boldsymbol{b} + \min_{\boldsymbol{x} \geq 0} \left\{\boldsymbol{c} \cdot \boldsymbol{x} - (\boldsymbol{A}^T \boldsymbol{y}) \cdot \boldsymbol{x}\right\} \\ &= \boldsymbol{y} \cdot \boldsymbol{b} + \min_{\boldsymbol{x} \geq 0} \left\{(\boldsymbol{c} - \boldsymbol{A}^T \boldsymbol{y}) \cdot \boldsymbol{x}\right\}, \end{aligned}$$

recordando que $\boldsymbol{x} \cdot (\boldsymbol{Ay}) = \boldsymbol{x}^T \boldsymbol{Ay} = (\boldsymbol{A}^T \boldsymbol{x})^T \boldsymbol{y} = (\boldsymbol{A}^T \boldsymbol{x}) \cdot \boldsymbol{y}$.

Ahora bien, notemos que

$$\min_{\boldsymbol{x} \geq 0} \left\{(\boldsymbol{c} - \boldsymbol{A}^T \boldsymbol{y}) \cdot \boldsymbol{x}\right\} = \begin{cases} 0 & \text{si } \boldsymbol{c} - \boldsymbol{A}^T \boldsymbol{y} \geq \boldsymbol{0}, \\ -\infty, & \text{en otro caso.} \end{cases}$$

Como queremos maximizar $g(\boldsymbol{y})$, podemos quedarnos solo con los casos para los que $g(\boldsymbol{y}) \neq -\infty$. Concluimos que el problema dual es equivalente al problema

$$\begin{aligned} &\text{maximizar:} && \boldsymbol{y} \cdot \boldsymbol{b}, \\ &\text{sujeto a:} && \boldsymbol{A}^T \boldsymbol{y} \leq \boldsymbol{c}. \end{aligned}$$

En resumen, definimos un vector $\boldsymbol{y}$ de parámetros o variables duales, y para cada elección de $\boldsymbol{y}$ podemos encontrar una cota inferior para costo óptimo del problema primal. El problema dual consiste en maximizar esa cota, es decir, obtener la cota mas precisa posible. Para algunos vectores $\boldsymbol{y}$, la cota que obtenemos es $-\infty$, que no tiene utilidad. Por lo tanto, nos restringimos a las elecciones de $\boldsymbol{y}$ que nos dan cotas no triviales, y eso nos da las restricciones del problema dual.

1.8.2. Teoremas de dualidad débil y fuerte

A partir del razonamiento anterior, deducimos el siguiente resultado

Teorema 1.23 (Dualidad débil)**.** *Si $\boldsymbol{x}$ es una solución factible del problema primal y $\boldsymbol{y}$ es una solución factible del problema dual, entonces*

$$\boldsymbol{y} \cdot \boldsymbol{b} \leq \boldsymbol{c} \cdot \boldsymbol{x}.$$

Si bien este resultado es muy simple, tiene consecuencias interesantes.

Corolario 1.24.

- *Si el costo óptimo del problema primal es $-\infty$, entonces el problema dual no es factible.*
- *Si el costo óptimo del problema dual es $+\infty$, entonces el problema primal no es factible.*

Corolario 1.25. *Sean $\boldsymbol{x}^\star \in \mathbb{R}^n$ e $\boldsymbol{y}^\star \in \mathbb{R}^m$ soluciones factibles de los problemas primal y dual respectivamente, y supongamos que $\boldsymbol{y}^\star \cdot \boldsymbol{b} = \boldsymbol{c} \cdot \boldsymbol{x}^\star$. Entonces $\boldsymbol{x}^\star$ y $\boldsymbol{y}^\star$ son soluciones óptimas de los problemas primal y dual respectivamente.*

Enunciamos ahora el resultado principal de la teoría dual de programación lineal.

Teorema 1.26 (Dualidad fuerte)**.** *Si un problema primal de programación lineal tiene solución óptima, entonces también el problema dual tiene solución óptima, y los respectivos costos óptimos son iguales.*

Demostración. Si aplicamos el método Simplex al problema primal

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

vamos a obtener una solución óptima $\boldsymbol{x}$ con base $\boldsymbol{B}$. Tomamos $\boldsymbol{x}_B = \boldsymbol{B}^{-1}\boldsymbol{b}$ el vector de los valores óptimos de las variables básicas. El vector de costos reducidos es

$$\boldsymbol{c}^T - \boldsymbol{c}_B^T \boldsymbol{B}^{-1}\boldsymbol{A},$$

donde $\boldsymbol{c}_B^T$ es el vector de costos (coeficientes) de las variables básicas. Como $\boldsymbol{x}$ es una solución óptima,

$$\boldsymbol{c}^T - \boldsymbol{c}_B^T \boldsymbol{B}^{-1}\boldsymbol{A} \geq 0.$$

Definimos ahora $\boldsymbol{y} \in \mathbb{R}^m$ tal que $\boldsymbol{y}^T = \boldsymbol{c}_B^T \boldsymbol{B}^{-1}$. Vemos que $\boldsymbol{y}^T\boldsymbol{A} \leq \boldsymbol{c}^T$, o equivalentemente $\boldsymbol{A}^T\boldsymbol{y} \leq \boldsymbol{c}$, y por lo tanto $\boldsymbol{y}$ es una solución factible del problema dual

$$\begin{aligned} &\text{maximizar:} && \boldsymbol{y} \cdot \boldsymbol{b} \\ &\text{sujeto a:} && \boldsymbol{A}^T\boldsymbol{y} \leq \boldsymbol{c}. \end{aligned}$$

Más aún,

$$\boldsymbol{y} \cdot \boldsymbol{b} = (\boldsymbol{B}^{-1})^T \boldsymbol{c}_B \cdot \boldsymbol{b} = \boldsymbol{c}_B \cdot \boldsymbol{B}^{-1}\boldsymbol{b} = \boldsymbol{c}_B \cdot \boldsymbol{x}_B = \boldsymbol{c} \cdot \boldsymbol{x}.$$

Por lo tanto, por dualidad débil, $\boldsymbol{y}$ es una solución optima del problema dual y el costo óptimo dual coincide con el costo óptimo primal. □

Capítulo 2

Introducción a la programación semidefinida

2.1. Motivación

Consideremos una sucesión de vectores definida recursivamente por

$$x[k+1] = Ax[k], \quad x[0] = x_0$$

Se puede ver que este sistema tiende a 0 para todo vector inicial si y solo todos los autovalores de A tienen módulo menor que 1. En este caso entonces, para estudiar el comportamiento en el largo plazo del sistema, nos interesa conocer el máximo de los módulos de los autovectores de A, que llamamos *radio espectral* de A.

Dada una matriz A con autovalores $\lambda_1, \ldots, \lambda_n$, los autovalores de $A - \alpha I$ son $\lambda_1 - \alpha, \ldots, \lambda_n - \alpha$. SI A es simétrica, todos sus autovalores son reales, y podemos calcular su radio espectral calculando el mayor y menor autovalor.

En este caso, podemos plantear el problema de hallar el menor autovalor como un problema de optimización:

$$\text{mín}\{\lambda : \lambda \text{ autovalor de } A\} = \sup_{\alpha \in \mathbb{R}} \{\alpha : A - \alpha I \succeq 0\}$$

y análogamente, podemos plantear el problema de hallar el mayor autovalor como

$$\text{máx}\{\lambda : \lambda \text{ autovalor de } A\} = \inf_{\alpha \in \mathbb{R}} \{\alpha : \alpha I - A \succeq 0\}.$$

Estudiaremos cómo resolver este tipo de problemas.

2.2. Preliminares

2.2.1. Matrices simétricas

Notamos $\mathcal{S}^n$ al espacio de matrices simétricas

$$\mathcal{S}^n = \{A \in \mathbb{R}^{n \times n} | a_{ij} = a_{ji}, \forall 1 \leq i, j \leq n\}.$$

Es un subespacio vectorial de $\mathbb{R}^{n\times n}$ de dimensión $\frac{n(n+1)}{2}$.

Una propiedad fundamental de las matrices simétricas es que todos los autovalores son reales, y poseen base ortonormal de autovectores. Este resultado se conoce como el teorema espectral.

Proposición 2.1. *Si $\boldsymbol{A}$ es simétrica, entonces*

(a) todos los autovalores de $\boldsymbol{A}$ son reales,

(b) autovectores correspondientes a autovalores distintos son perpendiculares,

(c) $\boldsymbol{S}^T\boldsymbol{A}\boldsymbol{S}$ es simétrica para toda $\boldsymbol{S} \in \mathbb{R}^{n\times n}$.

Demostración. Ejercicio. □

Teorema 2.2. *Cualquier matriz simétrica $X \in \mathcal{S}^n$ admite una descomposición*

$$X = PDP^T,$$

con $P \in \mathbb{R}^{n\times n}$ matriz ortogonal y $D \in \mathbb{R}^{n\times n}$ matriz diagonal, con los autovalores $\{\lambda_1, \ldots, \lambda_n\} \subset \mathbb{R}$ de X en la diagonal. Equivalentemente,

$$X = \sum_{i=1}^{n} \lambda_i u_i u_i^T,$$

con $u_i \in \mathbb{R}_n$, $1 \leq i \leq n$, los autovectores de X, que se corresponden con las columnas de P.

Demostración. Ver [Horn and Johnson, 1985, Teorema 4.1.5]. □

2.2.2. Matrices simétricas y ley de inercia

En el caso de matrices simétricas existe una diagonalización más simple que la diagonalización por matrices semejantes, que es especialmente útil en problemas de programación semidefinida.

Definición 2.3. *Dos matrices $A, B \in \mathbb{R}^{n\times n}$ son* congruentes *si existe una matrix $S \in \mathbb{R}^{n\times n}$ inversible tal que*

$$A = SBS^T.$$

La relación de congruencia es una relación de equivalencia.

Dada una matriz simétrica $\boldsymbol{A} \in \mathcal{S}^n$ definimos la *inercia* de $\boldsymbol{A}$ como la terna (n_+, n_-, n_0), donde n_+ la cantidad de autovalores positivos, n_- la cantidad de autovalores negativos y n_0 es la cantidad de autovalores nulos de $\boldsymbol{A}$.

Si $\boldsymbol{A} \in \mathcal{S}^n$, podemos escribir $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ con $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, la matriz con los autovalores de $\boldsymbol{A}$ en la diagonal, y $\boldsymbol{U}$ unitaria. Suponemos que los autovalores positivos son los primeros, luego los autovalores negativos y finalmente los autovalores cero. Definiendo la matriz diagonal no-singular real

$$\boldsymbol{D} = \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_{n_+}}, \sqrt{-\lambda_{n_++1}}, \ldots, \sqrt{-\lambda_{n_++n_-}}, 1, \ldots, 1),$$

obtenemos

$$\Lambda = D \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & -1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & -1 & & & \\ & & & & & & 0 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 0 \end{pmatrix} D.$$

Podemos escribir a la matriz $\boldsymbol{A}$ como

$$\boldsymbol{A} = \boldsymbol{U}\Lambda\boldsymbol{U}^T = \boldsymbol{S} \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & -1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & -1 & & & \\ & & & & & & 0 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 0 \end{pmatrix} \boldsymbol{S}^T = SI(A)S^T,$$

donde $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D}$ es una matriz no-singular y $I(A)$ es la matriz de inercia de $\boldsymbol{A}$.

Utilizando esta escritura, la ley de inercia nos permite determinar fácilmente si dos matrices simétricas son congruentes.

Teorema 2.4. *Dos matrices simétricas $A, B \in \mathcal{S}^n$ son congruentes si y solo si tienen la misma signatura.*

Demostración. Si dos matrices $\boldsymbol{A}, \boldsymbol{B}$ tienen la misma inercia, utilizando la construcción anterior, obtenemos que $\boldsymbol{A}$ y $\boldsymbol{B}$ son congruentes a la misma matriz de inercia. Como la relación de congruencia es una relación de equivalencia, las matrices $\boldsymbol{A}$ y $\boldsymbol{B}$ son congruentes.

Para la otra implicación, suponemos $\boldsymbol{A}$ y $\boldsymbol{B}$ congruentes, con $\boldsymbol{A} = \boldsymbol{S}\boldsymbol{B}\boldsymbol{S}^T$, para alguna matriz no-singular $\boldsymbol{S} \in \mathbb{R}^{n\times n}$. Como las matrices congruentes tienen el mismo rango, $n_0(\boldsymbol{A}) = n_0(\boldsymbol{B})$ y alcanza ver que $n_+(\boldsymbol{A}) = n_+(\boldsymbol{B})$. Notamos $k = n_+(\boldsymbol{A})$.

Sean $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ autovectores ortonormales de $\boldsymbol{A}$ correspondientes a los autovalores positivos $\lambda_1(\boldsymbol{A}), \ldots, \lambda_k$ y sea $S_+(\boldsymbol{A}) = \langle \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \rangle$.

La dimensión de $S_+(\boldsymbol{A})$ es k, y si

$$\boldsymbol{x} = \alpha_1\boldsymbol{v}_1 + \cdots + \alpha_k\boldsymbol{v}_k \neq 0,$$

entonces

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \lambda_1(A)|\alpha_1|^2 + \cdots + \lambda_k(A)|\alpha_k|^2 > 0.$$

Por lo tanto,

$$\boldsymbol{x}^T(\boldsymbol{S}\boldsymbol{B}\boldsymbol{S}^T)\boldsymbol{x} = (\boldsymbol{S}^T\boldsymbol{x})^T\boldsymbol{B}(\boldsymbol{S}^T\boldsymbol{x}) > 0,$$

luego $\boldsymbol{y}^T\boldsymbol{B}\boldsymbol{y} > 0$ para todo vector no nulo $\boldsymbol{y} \in \langle \boldsymbol{S}^T\boldsymbol{v}_1, \ldots, \boldsymbol{S}^T\boldsymbol{v}_k \rangle$, que es un espacio de dimensión k. Por lo tanto, $n_+(\boldsymbol{B}) \geq k = n_+(\boldsymbol{A})$ y realizando el mismo razonamiento intercambiando $\boldsymbol{A}$ por $\boldsymbol{B}$, obtenemos $n_+(\boldsymbol{A}) = n_+(\boldsymbol{B})$. □

En particular, cualquier matriz simétrica es congruente a una matriz diagonal con valores 0, +1 o −1 en la diagonal, y la cantidad de cada uno de estos valores en la diagonal es invariante, no depende de la matriz $\boldsymbol{S}$.

Finalmente, notamos que dada una matriz simétrica $\boldsymbol{A} \in \mathcal{S}^n$ podemos obtener una matriz diagonal $\boldsymbol{D}$ congruente a $\boldsymbol{A}$ por eliminación gaussiana (realizando simultáneamente eliminación en filas y columnas). Por lo tanto, podemos calcular eficientemente tanto la diagonalización por congruencia como la signatura de la matriz.

2.2.3. Matrices definidas positivas

Una matriz $A \in \mathbb{R}^{n\times n}$ se llama *semidefinida positiva* si es simétrica y

$$x^T Ax \geq 0$$

para todo $x \in \mathbb{R}^n$. Si además $x^T Ax = 0$ solo para $x = 0$, decimos que la matriz es *definida positiva*.

Si $A = M^T M$ para $M \in \mathbb{R}^{m\times n}$ cualquiera, entonces

$$x^T Ax = x^T M^T Mx = \|Mx\|^2 \geq 0,$$

lo que nos da un amplio stock de matrices positivas semidefinidas.

Para el siguiente resultado, dado un conjunto $S \subset \{1, \ldots, n\}$, definimos el menores principales asociado a S como el determinante de la submatriz cuadrada $A_{S,S}$ formada por las filas y columnas de A con índices en S. SI $S = \{1, 2, \ldots, k\}$, decimos que el menor asociado es un *menor principal de cabeza*.

Teorema 2.5. *Si $A \in \mathbb{R}^{n\times n}$ es un matriz simétrica, las siguientes propiedades son equivalentes.*

(a) A es semidefinida positiva ($x^T Ax \geq 0$ para todo $x \in \mathbb{R}^n$),

(b) $A = M^T M$ para alguna matriz $M \in \mathbb{R}^{n\times n}$,

(c) todos los autovalores de A son no-negativos,

(d) todos *los menores principales de A son no-negativos.*

Para matrices definidas positivas, obtenemos equivalencias similares.

Teorema 2.6. *Si $A \in \mathbb{R}^{n\times n}$ es un matriz simétrica, las siguientes propiedades son equivalentes.*

(a) A es definida positiva ($x^T Ax > 0$ para todo $x \in \mathbb{R}^n$),

(b) $A = M^T M$ para alguna matriz $M \in \mathbb{R}^{n\times n}$ no singular,

(c) todos los autovalores de A son positivos,

(d) todos los menores principales de A son positivos,

(e) el polinomio característico de $\boldsymbol{A}$ tiene signos alternados. Si $\chi_A(x) = x^n + a_{n-1}x^{n-1} + a_1 x + a_0$, entonces $a_i a_{i+1} < 0$ para todo $0 \leq i \leq n-1$ (definiendo $a_n = 1$).

En el caso de matrices definidas positivas, podemos restringir la condición (d) a considerar solo los menores principales de cabeza.

En ambos casos, la condición (b) podemos restringirla a matrices triangulares inferiores. Recordemos que cualquier matriz $C \in \mathbb{R}^{n\times n}$ admite una descomposición $\boldsymbol{QR}$ y puede escribirse como $\boldsymbol{C} = \boldsymbol{QR}$ con $\boldsymbol{Q}$ unitaria y $\boldsymbol{R}$ triangular superior del mismo rango que $\boldsymbol{C}$. Luego

$$\boldsymbol{A} = \boldsymbol{C}^T\boldsymbol{C} = (\boldsymbol{QR})^T\boldsymbol{QR} = \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{QR} = \boldsymbol{R}^T\boldsymbol{R}$$

y tomando $\boldsymbol{L} = \boldsymbol{R}^T$, obtenemos $\boldsymbol{A} = \boldsymbol{LL}^T$.

Si $\boldsymbol{C}$ es no singular, podemos elegir $\boldsymbol{R}$ con todos los valores en la diagonal positivos (y la descomposición de esta forma es única). Esto prueba el siguiente corolario, que nos da la *descomposición de Cholesky* de una matriz definida positiva.

Corolario 2.7. *Una matriz $\boldsymbol{A}$ es definida positiva si y solo si existe una matriz trinagular inferior $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ con valores positivos en la diagonal tal que $\boldsymbol{A} = \boldsymbol{LL}^T$.*

La siguiente propiedad será clave para estudiar la dualidad en programación semidefinida.

Teorema 2.8. *Una matriz $\boldsymbol{A}$ es semidefinida positiva si y solo si $\boldsymbol{A} \bullet X \geq 0$ para toda matriz $\boldsymbol{X} \in \mathcal{S}_+$.*

Demostración. Si $\boldsymbol{A} \succeq 0$, entonces $\boldsymbol{A} = \boldsymbol{B}^T\boldsymbol{B}$ para alguna matriz $\boldsymbol{B}$ y por lo tanto,

$$\boldsymbol{A} \bullet \boldsymbol{X} = \mathrm{Tr}(\boldsymbol{AX}) = \mathrm{Tr}(\boldsymbol{B}^T\boldsymbol{BX}) = \mathrm{Tr}(\boldsymbol{BXB}^T).$$

Como $\boldsymbol{X} \succeq 0$, también $\boldsymbol{BXB}^T \succeq 0$ y por lo tanto $\mathrm{Tr}(\boldsymbol{BXB}^T) \geq 0$.

Para la otra dirección, si $\boldsymbol{x} \in \mathbb{R}^n$, entonces $\boldsymbol{xx}^T \in \mathbb{R}^{n\times n}$ es una matriz semidefinida positiva. Si $\boldsymbol{A} \bullet \boldsymbol{X} \geq 0$ para toda $\boldsymbol{X}$, entonces $\boldsymbol{A} \bullet \boldsymbol{xx}^T \geq 0$ para todo $\boldsymbol{x} \in \mathbb{R}^n$ y

$$\boldsymbol{A} \bullet \boldsymbol{xx}^T = \mathrm{Tr}(\boldsymbol{Axx}^T) = \mathrm{Tr}(\boldsymbol{x}^T\boldsymbol{Ax}) = \boldsymbol{x}^T\boldsymbol{Ax}$$

y por lo tanto $\boldsymbol{x}^T\boldsymbol{Ax} \geq 0$ para todo $\boldsymbol{x}$. □

2.2.4. Conos

Referencia principal: [Sección 1.5, Laurent-Vallentin].

Un subconjunto $\mathcal{C}$ de un espacio vectorial real V es un *cono convexo* si es cerrado por combinaciones lineales positivas. Es decir,

(a) Si $\boldsymbol{x} \in \mathcal{C}$ y $a \geq 0$, entonces $a\boldsymbol{x} \in \mathcal{C}$.

(b) Si $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$, entonces $\boldsymbol{x} + \boldsymbol{y} \in \mathcal{C}$.

En este apunte siempre que hablemos de *conos* vamos a referirnos a conos convexos.

Decimos que un cono es *puntiagudo* si

$$\boldsymbol{x} \in \mathcal{C} \text{ y } -\boldsymbol{x} \in \mathcal{C} \Rightarrow \boldsymbol{x} = 0.$$

El *cono dual* de un cono $\mathcal{C}$ es

$$\mathcal{C}^* = \{\boldsymbol{y} \in \mathbb{R}^n \mid \boldsymbol{x}^T\boldsymbol{y} \geq 0 \ \ \forall \boldsymbol{x} \in \mathcal{C}\}.$$

Dado un cono puntiagudo $\mathcal{C}$ en $\mathbb{R}^n$, podemos definir un orden parcial en $\mathbb{R}^n$ por

$$\boldsymbol{x} \succeq \boldsymbol{y} \iff \boldsymbol{x} - \boldsymbol{y} \in \mathcal{C}$$

para $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Este orden satisface las siguientes propiedades:

- **reflexividad:** $\forall \boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} \succeq \boldsymbol{x}$
- **antisimetría:** $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{x} \succeq \boldsymbol{y}, \boldsymbol{y} \succeq \boldsymbol{x} \Rightarrow \boldsymbol{x} = \boldsymbol{y}$
- **transitividad:** $\forall \boldsymbol{x}, \boldsymbol{y}, z \in \mathbb{R}^n : \boldsymbol{x} \succeq \boldsymbol{y}, \boldsymbol{y} \succeq \boldsymbol{z} \Rightarrow \boldsymbol{x} \succeq \boldsymbol{z}$
- **homogeneidad:** $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \forall \alpha \in \mathbb{R}_{\geq 0} : \boldsymbol{x} \succeq \boldsymbol{y} \Rightarrow \alpha \boldsymbol{x} \succeq \alpha \boldsymbol{y}$
- **aditividad:** $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x}', \boldsymbol{y}' \in \mathbb{R}^n : \boldsymbol{x} \succeq \boldsymbol{y}, \boldsymbol{x}' \succeq \boldsymbol{y}' \Rightarrow \boldsymbol{x} + \boldsymbol{x}' \succeq \boldsymbol{y} + \boldsymbol{y}'$

En general, nos van a interesar conos convexos de dimensión completa, es decir conos con interior no vacío. En ese caso, podemos definir la desigualdad estricta

$$\boldsymbol{x} \succ \boldsymbol{y} \iff \boldsymbol{x} - \boldsymbol{y} \in \operatorname{int} \mathcal{C}.$$

Tenemos el siguiente resultado de separación.

Lema 2.9. *Sean $\mathcal{C} \subset \mathbb{R}^n$ un cono convexo cerrado y $\boldsymbol{x} \in \mathbb{R}^n \setminus \mathcal{C}$ un punto fuera de $\mathcal{C}$. Existe un hiperplano que separa a $\{\boldsymbol{x}\}$ de $\mathcal{C}$. Más aún, existe un vector no nulo $c \in \mathbb{R}^n$ tal que*

$$\forall \boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{c} \cdot \boldsymbol{y} \geq 0 > c \cdot \boldsymbol{x}.$$

Ejemplos

Cono generado por un conjunto de puntos. El cono generado por un conjunto de puntos $A = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subset \mathbb{R}^n$ es el menor cono que contiene a A. Es el cono

$$\operatorname{cono} A = \left\{ \sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i \mid \alpha_1, \ldots, \alpha_N \in \mathbb{R}_{\geq 0} \right\}.$$

El ortante no-negativo. El ortante no-negativo que utilizamos en programación lineal, definido por

$$\mathbb{R}^n_{\geq 0} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1, \ldots, x_n \geq 0\}$$

es un cono puntiagudo convexo cerrado de dimensión completa.

El cono de matrices semidefinidas positivas.

Proposición 2.10. *El conjunto $\mathcal{S}^n_+$ de matrices positivas semidefinidas es un cono convexo puntiagudo.*

Demostración. Podemos escribir a $\mathcal{S}^n_+$ como intersección de infinitos semiespacios:

$$\mathcal{S}^n_+ = \{\boldsymbol{A} \in \mathcal{S}^n \mid \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0 \ \ \forall \boldsymbol{x} \in \mathbb{R}^n\} = \bigcap_{\boldsymbol{x} \in \mathbb{R}^n} \{\boldsymbol{A} \in \mathcal{S}^n \mid \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0\},$$

y por lo tanto es un conjunto convexo cerrado. Para ver que forman un cono, si $\boldsymbol{A}, \boldsymbol{B} \succeq 0$ y $a \geq 0$, es claro que $a\boldsymbol{A} \succeq 0$ y también que

$$\boldsymbol{x}^T (\boldsymbol{A} + \boldsymbol{B}) \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{B} \boldsymbol{x} \geq 0$$

para todo $\boldsymbol{x} \in \mathbb{R}^n$.

Finalmente, para ver que es un cono puntiagudo, si $\boldsymbol{A} \in \mathcal{S}^n_+$ y $-\boldsymbol{A} \in \mathcal{S}^n_+$, todos los autovalores de $\boldsymbol{A}$ son 0. Como $\boldsymbol{A}$ es simétrica, el teorema espectral para matrices simétricas implica que $\boldsymbol{A}$ es la matriz nula. □

Proposición 2.11. *El interior de cono $\mathcal{S}^n_+$ de matrices semidefinidas positivas es $\mathcal{S}^n_{++}$, el conjunto de matrices definidas positivas. En particular, $\mathcal{S}^n_+$ es un cono de dimensión completa.*

Demostración. Ver [Fawzi, 2018, Teorema 3.1]. □

2.2.5. Espectrahedros.

Generalizamos la noción de poliedros permitiendo desigualdades matriciales.

Definición 2.12. *Una desigualdad lineal matricial (LMI por su nombre en inglés) tiene la forma*

$$\boldsymbol{A}_0 + \sum_{i=1}^{m} \boldsymbol{A}_i y_i \succeq 0,$$

con $\boldsymbol{A}_i \in \mathcal{S}^n$ matrices simétricas dadas.

Un conjunto de desigualdades matriciales definen un espectrahedro.

Definición 2.13. *Un conjunto $S \subset \mathbb{R}^m$ es un espectrahedro si tiene la forma*

$$S = \left\{ (y_1, \ldots, y_m) \in \mathbb{R}^m : \boldsymbol{A}_0 + \sum_{i=1}^{m} \boldsymbol{A}_i y_i \succeq 0 \right\},$$

para matrices simétricas $\boldsymbol{A}_0, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_m \in \mathcal{S}^n$.

Ejemplo 2.14. Todos los poliedros son espectrahedros. En efecto, para una matriz $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ diagonal, $\boldsymbol{X} \succeq 0$ si y solo $x_{ii} \geq 0$ para todo $1 \leq i \leq n$. Por lo tanto, podemos describir al poliedro $S = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$ como

$$S = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n : \begin{pmatrix} \boldsymbol{a}_1 \cdot \boldsymbol{x} - b_1 & & \\ & \ddots & \\ & & \boldsymbol{a}_m \cdot \boldsymbol{x} - b_m \end{pmatrix} \succeq 0 \right\}.$$

Sin embargo, existen espectrahedros que no son poliedros.

Ejemplo 2.15. El disco unitario $D = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$ es un espectrahedro. En efecto,

$$D = \left\{ (x, y) \in \mathbb{R}^2 \mid \begin{pmatrix} 1 - x & y \\ y & 1 + x \end{pmatrix} \succeq 0 \right\}.$$

Utilizando el criterio de los menores principales, vemos que $\begin{pmatrix} 1 - x & y \\ y & 1 + x \end{pmatrix} \succeq 0$ si y solo si:

$$\begin{aligned} 1 - x &\geq 0, \\ 1 + x &\geq 0, \\ (1 - x)(1 + x) - y^2 = 1 - x^2 - y^2 &\geq 0. \end{aligned}$$

Geométricamente, un espectrahedro está definido por la intersección del cono $\mathcal{S}_+$ de matrices positivas semidefinidas con un espacio afín: el espacio generado por las matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m$ trasladado a $\boldsymbol{A}_0$. Por lo tanto, un espectrahedro es siempre un conjunto cerrado y convexo.

Determinar qué conjuntos convexos son espectrahedros y cuáles no mediante un criterio sencillo es un problema abierto de investigación.

Ejercicio 2.16. Demostrar que el conjunto $D = \{(x, y) \in \mathbb{R}^2 \mid x^4 + y^4 \leq 1\}$ es un espectrahedro.

Sugerencia:

$$x^4 + y^4 \leq 1 \iff \exists u, v \text{ tales que } x^2 \leq u, y^2 \leq v, u^2 + v^2 \leq 1.$$

Ejemplo 2.17. Consideramos el espectrahedro en $\mathbb{R}^2$ definido por

$$\{(x,y) \in \mathbb{R}^2 : \boldsymbol{A}(x,y) := \begin{pmatrix} x+1 & 0 & y \\ 0 & 2 & -x-1 \\ y & -x-1 & 2 \end{pmatrix} \succeq 0\}.$$

El polinomio caracterísitico de $\boldsymbol{A}$ es $\xi_{\boldsymbol{A}}(t) = t^3 - p_2 t^2 + p_1 t + p_0$, con $p_2 = -x-5$, $p_1 = -x^2+2x-y^2+7$ y $p_0 = -3-x+x^3+3x^2+2y^2$. Por lo tanto, la condición $\boldsymbol{A} \succeq 0$ es equivalente a

$$\begin{aligned} -p_2 &= x+5 \geq 0 \\ p_1 &= -x^2+2x-y^2+7 \geq 0 \\ -p_0 &= 3+x-x^3-3x^2-2y^2 \geq 0. \end{aligned}$$

Este conjunto corresponde al óvalo de la curva elíptica $3+x-x^3-3x^2-2y^2=0$.

2.2.6. Espectrahedros proyectados

Resulta interesante también estudiar las proyecciones lineales de espectrahedros.

Definición 2.18. *Un conjunto $S \subset \mathbb{R}^m$ es un espectrahedro proyectado si es de la forma*

$$S = \{(x_1,\ldots,x_m) \in \mathbb{R}^m \mid \exists (y_1,\ldots,y_p) \in \mathbb{R}^p, \boldsymbol{A}_0 + \sum_{i=1}^m \boldsymbol{A}_i x_i + \sum_{j=1}^p \boldsymbol{B}_j y_j \succeq 0\},$$

donde $\boldsymbol{A}_0,\ldots,\boldsymbol{A}_m,\boldsymbol{B}_1,\ldots,\boldsymbol{B}_p$ son matrices simétricas dadas.

Ejemplo 2.19. Ver [Blekherman et al., 2013, Ejemplo 2.9].

Veremos más adelante que este es un ejemplo de espectrahedro proyectado que no es un espectrahedro. Es decir, que trabajando con espectrahedros proyectados podemos ampliar el conjunto de conjuntos convexos sobre los que podemos resolver problemas.

2.2.7. Formulación primal del problema de programación semidefinida

Para plantear este tipo de problemas en forma análoga a un problema de programación lineal, introducimos la siguiente notación.

Para matrices $\boldsymbol{X},\boldsymbol{Y} \in \mathbb{R}^{n\times n}$, llamamos *producto interno traza* al producto

$$\boldsymbol{X} \bullet \boldsymbol{Y} = \mathrm{Tr}(\boldsymbol{X}^T\boldsymbol{Y}) = \sum_{i=1}^n \sum_{j=1}^n x_{ij}y_{ij}, \quad \text{donde } \mathrm{Tr}(\boldsymbol{X}) = \sum_{i=1}^n x_{ii},$$

que coincide con el producto interno usual pensado a $\boldsymbol{X}$ e $\boldsymbol{Y}$ como vectores de $n \times n$ coordenadas.

Con este producto interno, $\mathbb{R}^{n\times n}$ resulta un espacio euclídeo.

Se cumplen las siguientes propiedades:

(a) $\boldsymbol{X} \bullet \boldsymbol{Y} = \boldsymbol{Y} \bullet \boldsymbol{X}$,

(b) $\boldsymbol{X} \bullet \boldsymbol{I} = \boldsymbol{I} \bullet \boldsymbol{X} = \mathrm{Tr}(\boldsymbol{X})$.

Recordemos que $\mathcal{S}^n \subset \mathbb{R}^{n\times n}$ denota al conjunto de matrices simétricas de $n \times n$.

Con la notación introducida, escribimos un problema de programación semidefinida como

$$\begin{aligned} \text{minimizar:} \quad & \boldsymbol{C} \bullet \boldsymbol{X} \\ \text{sujeto a:} \quad & \boldsymbol{A}_i \bullet \boldsymbol{X} = b_i, \quad i = 1, \ldots, m, \\ & \boldsymbol{X} \succeq 0, \end{aligned}$$

donde $\boldsymbol{C}, \boldsymbol{A}_i \in \mathcal{S}^n$. La matrix $\boldsymbol{X} \in \mathcal{S}^n$ es la variable sobre la cual realizamos la minimización.

Podemos ver en seguida la similitud con un problema de programación lineal. La función a minimizar es una funcional lineal exactamente igual que en programación lineal. El conjunto factible, es decir el conjunto sobre el que se minimiza la función es un espectrahedro.

En efecto, para describir el conjunto $\boldsymbol{A}_i \bullet \boldsymbol{X} = b_i$, $i = 1, \ldots, m$, $\boldsymbol{X} \succeq 0$ mediante una desigualdad matricial lineal alcanza escribir cada coordenada de $\boldsymbol{X}$ como una expresión lineal en nuevas variables $\{y_1, \ldots, y_s\}$.

Ejemplo 2.20. Podemos describir conjunto de matrices

$$S = \{\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} \\ x_{12} & x_{22} \end{pmatrix} \mid x_{11} + x_{22} = 1, \boldsymbol{X} \succeq 0\}$$

mediante la condición

$$\begin{pmatrix} x_{11} & x_{12} \\ x_{12} & 1 - x_{11} \end{pmatrix} \succeq 0$$

que a la vez es equivalente a la desigualdad matricial

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} + x_{11} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + x_{12} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \succeq 0.$$

2.3. Dualidad

Estudiamos ahora la teoría de dualidad en programación semidefinida.

Definimos un problema en forma primal estándar como

$$\begin{aligned} \text{minimizar:} \quad & \boldsymbol{C} \bullet \boldsymbol{X} \\ \text{sujeto a:} \quad & \boldsymbol{A}_i \bullet \boldsymbol{X} = b_i, \quad i = 1, \ldots, m, \\ & \boldsymbol{X} \succeq 0, \end{aligned} \tag{SDP-P}$$

y asociamos a un problema de esta forma otro problema, llamado *problema dual*, que puede plantearse en la forma

$$\begin{aligned} \text{maximizar:} \quad & \boldsymbol{y} \cdot \boldsymbol{b} \\ \text{sujeto a:} \quad & \sum_{i=1}^{m} y_i \boldsymbol{A}_i \preceq \boldsymbol{C} \end{aligned} \tag{SDP-D}$$

Utilizando el Teorema 2.8, podemos seguir casi al pie de la letra la motivación que vimos para el problema de programación lineal para motivar la formulación de este problema dual.

Podemos relajar este problema reemplazando las restricciones $\boldsymbol{A}_i \bullet \boldsymbol{X} = b_i, \quad i = 1, \ldots, m$, por una penalidad

$$\sum_{i=1}^{m} y_i(b_i - \boldsymbol{A}_i \bullet \boldsymbol{X}),$$

donde $\boldsymbol{y} \in \mathbb{R}^m$ es un vector de costos del mismo tamaño que $\boldsymbol{b}$. Obtenemos entonces el siguiente problema relajado:

$$\begin{aligned} \text{minimizar:} \quad & \boldsymbol{C} \bullet \boldsymbol{X} + \sum_{i=1}^{m} y_i(b_i - \boldsymbol{A}_i \bullet \boldsymbol{X}), \\ \text{sujeto a:} \quad & \boldsymbol{X} \succeq 0. \end{aligned}$$

Llamamos $g(\boldsymbol{y})$ al costo óptimo en función de $\boldsymbol{y}$ del problema relajado. Como tenemos más libertad en la elección de variables en el problema relajado, esperamos que $g(\boldsymbol{y})$ sea menor o igual que el costo óptimo $\boldsymbol{C} \bullet \boldsymbol{X}$ del problema primal. En efecto, si $\boldsymbol{X}^\star$ es una solución óptima del problema primal (suponiendo que existe tal matriz),

$$g(\boldsymbol{y}) = \min_{\boldsymbol{X} \succeq 0} \left\{ \boldsymbol{C} \bullet \boldsymbol{X} + \sum_{i=1}^{m} y_i(b_i - \boldsymbol{A}_i \bullet \boldsymbol{X}) \right\} \leq \boldsymbol{C} \cdot \boldsymbol{X}^\star + \sum_{i=1}^{m} y_i(b_i - \boldsymbol{A}_i \bullet \boldsymbol{X}^\star) = \boldsymbol{C} \cdot \boldsymbol{X}^\star,$$

donde usamos para la última desigualdad que $\boldsymbol{X}^\star$ es una solución óptima del problema primal y por lo tanto $\boldsymbol{A}_i \bullet \boldsymbol{X}^\star = b_i$ para todo $1 \leq i \leq m$.

Por lo tanto, para cualquier vector $\boldsymbol{y}$, $g(\boldsymbol{y}) \leq \boldsymbol{C} \cdot \boldsymbol{X}^\star$ y el valor de $g(\boldsymbol{y})$ es una cota inferior del costo óptimo $\boldsymbol{C} \cdot \boldsymbol{X}^\star$ del problema primal. Queremos hallar la mejor cota posible, es decir, queremos hallar el vector $\boldsymbol{p}$ que maximice $g(\boldsymbol{y})$. Este problema se conoce como el *problema dual*, que podemos plantear en la forma:

$$\begin{aligned} \text{maximizar:} \quad & g(\boldsymbol{y}), \\ \text{sujeto a:} \quad & \text{ninguna restricción.} \end{aligned}$$

El resultado principal de la teoría de dualidad afirma que el costo óptimo del problema dual es igual al costo óptimo $\boldsymbol{C} \bullet \boldsymbol{X}^\star$ del problema primal. En otras palabras, si elegimos el valor de $\boldsymbol{y}$ para el cual $g(\boldsymbol{y})$ es máximo, la posibilidad de violar las restricciones $\boldsymbol{A}_i \bullet \boldsymbol{X}^\star = b_i$, $1 \leq i \leq m$, no es de ninguna ayuda.

Por la definición de $g(\boldsymbol{y})$,

$$\begin{aligned} g(\boldsymbol{y}) = \min_{\boldsymbol{X} \succeq 0} \left\{ \boldsymbol{C} \bullet \boldsymbol{X} + \sum_{i=1}^{m} y_i(b_i - \boldsymbol{A}_i \bullet \boldsymbol{X}) \right\} &= \sum_{i=1}^{m} y_i b_i + \min_{\boldsymbol{X} \succeq 0} \left\{ \boldsymbol{C} \bullet \boldsymbol{X} - \sum_{i=1}^{m} y_i(\boldsymbol{A}_i \bullet \boldsymbol{X}) \right\} \\ &= \boldsymbol{y} \cdot \boldsymbol{b} + \min_{\boldsymbol{X} \succeq 0} \left\{ \left(\boldsymbol{C} - \sum_{i=1}^{m} y_i \boldsymbol{A}_i \right) \bullet \boldsymbol{X} \right\}. \end{aligned}$$

Ahora bien, por el Teorema 2.8, si $\boldsymbol{C} - \sum_{i=1}^{m} y_i \boldsymbol{A}_i \succeq 0$,

$$\boldsymbol{C} - \sum_{i=1}^{m} y_i \boldsymbol{A}_i \bullet \boldsymbol{X} \geq 0$$

para toda $\boldsymbol{X} \succeq 0$ y vale 0 si $\boldsymbol{X} = \boldsymbol{0}$. Recíprocamente, si $\boldsymbol{C} - \sum_{i=1}^{m} y_i \boldsymbol{A}_i \not\succeq 0$, existe $\boldsymbol{X} \succeq 0$ tal que $\boldsymbol{C} - \sum_{i=1}^{m} y_i \boldsymbol{A}_i \succeq 0 \bullet \boldsymbol{X} < 0$, y tomando múltiplos de $\boldsymbol{X}$ podemos hacer este valor tan chico como querramos. Obtenemos

$$\min_{\boldsymbol{X} \succeq 0} \left\{ \left(\boldsymbol{C} - \sum_{i=1}^{m} y_i \boldsymbol{A}_i \right) \bullet \boldsymbol{X} \right\} = \begin{cases} 0 & \text{si } \boldsymbol{C} - \sum_{i=1}^{m} y_i \boldsymbol{A}_i \succeq 0, \\ -\infty, & \text{en otro caso.} \end{cases}$$

Como queremos maximizar $g(\boldsymbol{y})$, podemos quedarnos solo con los casos para los que $g(\boldsymbol{y}) \neq -\infty$, que corresponden a $\boldsymbol{C} - \sum_{i=1}^{m} y_i \succeq 0$. Concluimos que el problema dual es equivalente al problema

$$\begin{aligned} \text{maximizar:} \quad & \boldsymbol{y} \cdot \boldsymbol{b}, \\ \text{sujeto a:} \quad & \sum_{i=1}^{m} y_i \boldsymbol{A}_i \preceq \boldsymbol{C}, \end{aligned}$$

como queríamos.

En resumen, definimos un vector $\boldsymbol{y}$ de parámetros o variables duales, y para cada elección de $\boldsymbol{y}$ podemos encontrar una cota inferior para costo óptimo del problema primal. El problema dual consiste en maximizar esa cota, es decir, obtener la cota mas precisa posible. Para algunos vectores $\boldsymbol{y}$, la cota que obtenemos es $-\infty$, que no tiene utilidad. Por lo tanto, nos restringimos a las elecciones de $\boldsymbol{y}$ que nos dan cotas no triviales, y eso nos da las restricciones del problema dual.

Podemos sintetizar lo que acabamos de ver en el siguiente teorema.

Teorema 2.21 (dualidad débil)**.** *Dadas $\boldsymbol{X}$ e $\boldsymbol{y}$ dos soluciones factibles del problema primal y el problema dual respectivamente,*

$$\boldsymbol{C} \bullet \boldsymbol{X} - \boldsymbol{y} \cdot \boldsymbol{b} \geq 0.$$

Demostración. Tenemos

$$\boldsymbol{C} \bullet \boldsymbol{X} - \boldsymbol{y} \cdot \boldsymbol{b} = \boldsymbol{C} \bullet \boldsymbol{X} - \sum_{i=1}^{m} y_i \left(\boldsymbol{A}_i \bullet \boldsymbol{X}\right) = \left(\boldsymbol{C} - \sum_{i=1}^{m} y_i \boldsymbol{A}_i\right) \bullet \boldsymbol{X} \geq 0,$$

por el Teorema 2.8. □

Como en el caso de programación lineal, tenemos el siguiente importante corolario inmediato.

Corolario 2.22. *Sean $\boldsymbol{X}$ e $\boldsymbol{y}$ soluciones factibles de los problemas primal y dual respectivamente, y supongamos que $\boldsymbol{y} \cdot \boldsymbol{b} = \boldsymbol{C} \bullet \boldsymbol{X}$. Entonces $\boldsymbol{X}$ e $\boldsymbol{y}$ son soluciones óptimas de los problemas primal y dual respectivamente.*

La condición $\boldsymbol{y} \cdot \boldsymbol{b} = \boldsymbol{C} \bullet \boldsymbol{X}$ es equivalente a la condición

$$\left(\boldsymbol{C} - \sum_{i=1}^{m} y_i \boldsymbol{A}_i\right) \bullet \boldsymbol{X} = 0$$

que llamamos la *condición de holgura complementaria*.

Veamos algunos ejemplos de la relación que puede darse entre el problema primal y dual.

Ejemplo 2.23. Para el problema primal para $\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{21} \\ x_{21} & x_{22} \end{pmatrix}$:

$$\begin{aligned} \text{minimizar:} \quad & 2x_{11} + 2x_{12}, \\ \text{sujeto a:} \quad & x_{11} + x_{22} = 1, \\ & \boldsymbol{X} \succeq 0, \end{aligned}$$

despejando $x_{22} = 1 - x_{11}$, las restricciones que deben cumplirse son: $x_{11} \geq 0$, $x_{11} \leq 1$ y

$$x_{11}(1 - x_{11}) \geq x_{12}^2,$$

cuyo gráfico es un disco de radio $1/2$ centrado en el punto $(1/2, 0)$.

La solución óptima es

$$\boldsymbol{X}^\star = \begin{pmatrix} \frac{2-\sqrt{2}}{4} & -\frac{\sqrt{2}}{4} \\ -\frac{\sqrt{2}}{4} & \frac{2+\sqrt{2}}{4} \end{pmatrix}$$

y el costo óptimo es $1 - \sqrt{2}$.

Para obtener el problema dual definimos $\boldsymbol{C} = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$, $\boldsymbol{A}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ y $b_1 = 1$ ($m = 1$).

El problema dual es

$$\begin{array}{ll} \text{maximizar:} & y_1 \\ \text{sujeto a:} & y_1 \boldsymbol{A}_1 \preceq \boldsymbol{C}. \end{array}$$

o equivalentemente

$$\begin{array}{ll} \text{maximizar:} & y \\ \text{sujeto a:} & \begin{pmatrix} 2-y & 1 \\ 1 & -y \end{pmatrix} \succeq 0. \end{array}$$

En este caso tenemos las restricciones $y \leq 2$, $y \leq 0$ y $-(2-y)y - 1 \geq 0$. Completando cuadrado obtenemos

$$(y-1)^2 \geq 2,$$

y el máximo valor posible de $y < 0$ es $y = 1 - \sqrt{2}$.

Vemos que los valores óptimos de ambos problemas coinciden y se cumple la condición de holgura complementaria

$$(\boldsymbol{C} - y^\star \boldsymbol{A}_1) \bullet \boldsymbol{X}^\star = \begin{pmatrix} 1+\sqrt{2} & 1 \\ 1 & \sqrt{2}-1 \end{pmatrix} \bullet \begin{pmatrix} \frac{2-\sqrt{2}}{4} & -\frac{\sqrt{2}}{4} \\ -\frac{\sqrt{2}}{4} & \frac{2+\sqrt{2}}{4} \end{pmatrix} = 0.$$

Veamos a continuación otro ejemplo en el cual ambos problemas son factibles y sin embargo los valores óptimos son diferentes.

Ejemplo 2.24. Para $\boldsymbol{X} \in \mathbb{R}^{3\times 3}$ y $\alpha \geq 0$ dado, consideramos el siguiente par de problemas primal-dual

$$\begin{array}{ll} \text{minimizar:} & \alpha x_{11} \\ \text{sujeto a:} & x_{22} = 0, \\ & x_{11} + 2x_{23} = 1, \\ & \boldsymbol{X} \succeq 0. \end{array} \qquad \begin{array}{ll} \text{maximizar:} & y_2 \\ \text{sujeto a:} & \begin{pmatrix} \alpha - y_2 & 0 & 0 \\ 0 & -y_1 & -y_2 \\ 0 & -y_2 & 0 \end{pmatrix} \succeq 0. \end{array}$$

Para el problema primal, como $x_{22} = 0$ y $\boldsymbol{X} \succeq 0$, todas las coordenadas de $\boldsymbol{X}$ en la segunda fila y segunda columna deben ser 0. Por lo tanto, $0 = x_{23} = \frac{1-x_{11}}{2}$ y obtenemos que $x_{11} = 1$. La única solución factible del problema es

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

y el costo óptimo es α.

Para el problema dual, debe ser $y_2 = 0$ y tomando cualquier $y_1 \geq 0$ obtenemos una solución factible. El costo óptimo es por lo tanto 0, el salto de dualidad es $p^\star - d^\star = \alpha - 0 = \alpha$.

Analizamos ahora otra situación que puede darse en la que si bien los costos óptimos coinciden, los problemas pueden no tener solución óptima.

Ejemplo 2.25. Para $\boldsymbol{X} \in \mathbb{R}^{2\times 2}$, consideramos el siguiente par de problemas primal-dual

$$\begin{array}{lll} \text{minimizar:} & x_{11} & \\ \text{sujeto a:} & x_{21} = 1, & \\ & \boldsymbol{X} \succeq 0. & \end{array} \qquad \begin{array}{lll} \text{maximizar:} & y & \\ \text{sujeto a:} & \begin{pmatrix} 1 & -\frac{y}{2} \\ -\frac{y}{2} & 0 \end{pmatrix} \succeq 0. & \end{array}$$

Para el problema primal, la restricción $\begin{pmatrix} x_{11} & 1 \\ 1 & x_{22} \end{pmatrix} \succeq 0$ se cumple si

$$x_{11} \geq 0, \quad x_{22} \geq 0, \quad x_{11}x_{22} \geq 1.$$

El costo óptimo es 0 pero no existe ningún valor de (x_{11}, x_{22}) para el cuál se alcance el ínfimo.

Para el problema dual, debe ser $\frac{y}{2} = 0$. Por lo tanto, el costo óptimo es $d^\star = 0$ que se alcanza para la solución factible $y = 0$.

Vemos que si bien el salto de dualidad $p^\star - d^\star = 0$, no existen $\boldsymbol{X}$ e $\boldsymbol{y}$ que cumplan la condición de holgura complementaria.

En cierto sentido, estos últimos ejemplos son patológicos. Bajo condiciones blandas la dualidad fuerte también se cumple en programación semidefinida. Una de tales condiciones es requerir que los problemas sean *estrictamente factibles*. En el problema primal, esto significa que existe $\boldsymbol{X} \succ 0$ que verifica las restricciones. En el problema dual, esto significa que existe $\boldsymbol{y} \in \mathbb{R}^m$ tal que $\boldsymbol{C} - \sum_{i=1}^m y_i \boldsymbol{A}_i \succ 0$. En este caso la situación es tan buena como en programación lineal.

Teorema 2.26 (dualidad fuerte)**.** *Suponemos que tanto el problema primal* (SDP-P) *como el problema dual* (SDP-D) *son estrictamente factibles. Entonces ambos problemas tienen solución óptima, y el costo óptimo coincide. Es decir, no hay salto de dualidad.*

Demostración. Ver [Todd, 2001, Teorema 4.1]. □

Capítulo 3

Aplicaciones

3.1. Minimización del máximo autovalor

Referencia principal: [Todd, 2001, Sección 3, Ejemplo 1].

Este problema aparece por ejemplo en la estabilización de ecuaciones diferenciales.

Suponemos que tenemos una matriz simétrica $M(z)$ que depende linealmente (afín) de un vector z. Queremos elegir z que minimice el máximo autovalor de $M(z)$. Observemos que

$$\lambda_{\text{máx}}(M(z)) \leq \eta \text{ si y solo si } \lambda_{\text{máx}}(M(z) - \eta I) \leq 0,$$

o equivalentemente, $\lambda_{\text{mín}}(\eta I - M(z)) \geq 0$. Esto se cumple si y solo si $\eta I - M(z) \succeq 0$. Obtenemos entonces el problema SDP en forma dual:

$$\begin{array}{ll} \text{maximizar:} & -\eta \\ \text{sujeto a:} & \eta I - M(z) \succeq 0. \end{array}$$

El problema contrario, maximizar el mínimo autovalor puede interpretarse como obtener la matriz "lo más positiva posible". Veremos más adelante cómo utilizar este problema SDP para calcular descomposiciones en suma de cuadrados de polinomios.

3.2. Optimización combinatoria: maxcut

Referencias principales: [Todd, 2001, Sección 3, Ejemplo 9], [O'Donell, 2008, Lección 14].

Comenzamos con un grafo no-dirigido $G = (V, E)$, donde V es el conjunto $\{v_1, \ldots, v_n\}$ de n nodos y $E \subset V \times V$ es el conjunto de aristas. Definimos una matriz de costos no-negativos $W = (w_{ij}) \subset \mathbb{R}^{n \times n}_{\geq 0}$. Podemos suponer que el grafo es completo, es decir que todos los nodos son adyacentes entre sí, tomando $w_{ij} = 0$ para todas las no-aristas ij; definimos también $w_{ii} = 0$ para todo i. W resulta una matriz simétrica.

Para un subconjunto $K \subset N$ de nodos, definimos el corte $\delta(K)$ como el conjunto de todos las aristas desde K al complemento de K:

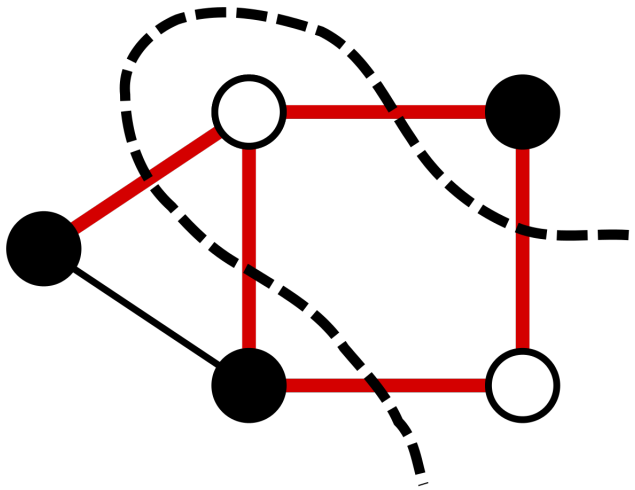
$$\delta(K) = \{(i, j) \in E : i \in K, j \notin K\}.$$

Dado un corte K definimos el costo del corte

$$w(\delta(K)) := \sum_{(i,j)\in\delta(K)} w_{ij}.$$

Queremos hallar el corte con el mayor costo posible.

Ejemplo 3.1. En el siguiente grafo de 5 nodos, asignamos costo 1 a todas las aristas dibujadas y costo 0 a las aristas no dibujadas. Tomando K el conjunto de nodos blancos, obtenemos el corte de mayor costo posible.



Ejemplo 3.2. Si consideramos un grafo completo de 4 vértices, con costo 1 en todas las aristas, el costo de un corte queda determinado por la cantidad de nodos en el corte. Obtenemos los siguientes valores:

$$\delta(K) = \begin{cases} 0 & \text{si } \#K = 0, \\ 3 & \text{si } \#K = 1, \\ 4 & \text{si } \#K = 2, \\ 3 & \text{si } \#K = 3, \\ 0 & \text{si } \#K = 4. \end{cases}$$

Por lo tanto, para obtener el corte de mayor costo tomamos un conjunto K de 2 elementos.

Comenzamos formulando el problema en forma matricial. Queremos calcular el costo de un corte mediante un producto de matrices. Fijamos un corte K y definimos el vector $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^n$ por

$$x_i = \begin{cases} 1 & \text{si } v_i \in K \\ -1 & \text{si } v_i \notin K \end{cases}.$$

De esta forma, $x_i x_j = -1$ si $(i,j) \in \delta(K)$ y $x_i x_j = 1$ si $(i,j) \notin \delta(K)$, y por lo tanto

$$1 - x_i x_j = \begin{cases} 0 & \text{si } (i,j) \in \delta(K) \\ 2 & \text{si } (i,j) \notin \delta(K). \end{cases}$$

Obtenemos

$$w(\delta(K)) = \frac{1}{2}\sum_{i<j} w_{ij}(1 - x_i x_j) = \frac{1}{4}\sum_{i\neq j} w_{ij}(1 - x_i x_j).$$

Como además $w_{ii} = 0$ para todo i,

$$w(\delta(K)) = \frac{1}{4}\sum_i \sum_j w_{ij}(1 - x_i x_j) = \left(\frac{1}{4}\sum_i \sum_j w_{ij}\right) - \frac{1}{4}x^T W x.$$

Para poder juntar los últimos dos términos en uno, reescribimos:

$$\left(\frac{1}{4}\sum_i \sum_j w_{ij}\right) - \frac{1}{4}x^T W x = \left(\frac{1}{4}\sum_i \left(\sum_j w_{ij}\right) x_i x_i\right) - \frac{1}{4}x^T W x.$$

y definiendo $C \in \mathbb{R}^{n\times n}$ con $c_{ij} = -w_{ij}/4$ para $i \neq j$ y $c_{ii} = \sum_j w_{ij}/4$ para todo i obtenemos

$$w(\delta(K)) = x^T C x.$$

Finalmente, como cualquier vector $x \in \{-1, 1\}^n$ define un corte, podemos escribir el problema max-cut como el problema de programación cuadrática entera:

$$\text{(IQP):} \quad \text{máx}\, x^T C x, x_i \in \{+1, -1\}, i \in N,$$

o como un problema cuadrático con restricciones cuadráticas no-convexas

$$\text{(NQCQP):} \quad \text{máx}\, x^T C x, x_i^2 = 1, i \in N.$$

Ninguna de estas dos formulaciones corresponde a un problema de programación semidefinida. A continuación veremos como relajar las condiciones para obtener un problema SDP.

3.2.1. Relajación 1

Observamos que (NQCQP) es linear en los productos $x_i x_j$ y que estos productos son las coordenadas de la matriz $X = xx^T \in \mathbb{R}^{n\times n}$ de rango 1. Más aún, X es simétrica, $X_{ii} = 1$ para todo i y $X \succeq 0$. Recíprocamente, cualquier matriz de rango 1 con esas propiedades puede escribirse como xx^T para algún vector $x \in \{-1, 1\}^n$. (COMPLETAR DEMOSTRACION: $X = yx^T$, y podemos tomar $y = x$ y como $x_i^2 = 1$ es de la forma buscada.)

Finalmente, como $x^T C x = C \bullet (xx^T)$, obtenemos que (IQP) es equivalente al problema

$$\text{máx}\, C \bullet X, \quad X_{ii} = 1, i \in N, X \succeq 0, \text{rank}(X) = 1.$$

Eliminando la restricción del rango, obtenemos el problema SDP

$$\text{máx}\, C \bullet X, \quad X_{ii} = 1, i \in N, X \succeq 0.$$

3.2.2. Relajación 2

Observemos que en (IQP) asociamos a cada nodo v_i un valor $x_i \in \{-1, 1\}$, que podemos considerar como un vector unitario de dimensión 1. Ahora, en cambio, asociamos a cada nodo v_i un vector unitario

$p_i \in \mathbb{R}^n$, y consideramos la matrix $\boldsymbol{P}$ con estos vectores como filas. Reemplazamos entonces la función objetivo $C \bullet (xx^T)$ por $C \bullet (PP^T)$ y las restricciones $x_i \in \{+1, -1\}$ por restricciones de los elementos de la diagonal de PP^T: $(PP^T)_{ii} = 1$. Como PP^T es semidefinida positiva, y cualquier matrix semidefinida positiva se puede factorizar de esta forma, vemos que mediante esta construcción obtenemos el mismo problema SDP de antes. Considerando una matriz $\boldsymbol{P}$ donde todas las filas son de la forma v o $-v$ para un vector unitario v, vemos que este problema es efectivamente una relajación del problema (IQP).

3.2.3. Relación entre la relajación y el problema original

Como las formulaciones SDP son relajaciones del problema max-cut, el valor óptimo del problema SDP es una cota superior del valor óptimo del problema original. Pero vamos a ver que podemos usar la solución del problema SDP para obtener un corte razonablemente bueno.

Usamos la segunda relajación, y tomamos $X = PP^T$ una solución óptima de ese problema. Si todas las filas de $\boldsymbol{P}$ fueran $+v$ o $-v$ para un vector v, definimos el corte tomando en K todos los nodos para los cuales la fila correspondiente es $+v$.

En la situación general, tomamos un hiperplano que divide a la esfera unitaria en dos mitades y tomamos en K a todos los nodos que quedan en una de estas dos mitades.

Más concretamente, para un vector aleatorio g de norma 1, definimos

$$K = \{v_i \in V : g^T p_i \geq 0\}.$$

Esto nos da un corte aleatorio, y podemos calcular la esperanza del valor del corte (veremos más adelante cómo tomar un vector aleatorio).

La esperanza del valor del corte es

$$E[w(\delta(K))] = E\left[\sum_{i,j} w_{ij} \mathbb{1}_{(i,j)\in\delta(K)}\right] = \sum_{i,j} w_{ij} \Pr[(i,j) \in \delta(K)].$$

(observamos que las probabilidades no son independientes, pero no es necesaria independencia para distribuir la esperanza con respecto a la suma).

Calculamos ahora $\Pr[(i,j) \in \delta(K)]$ para (i,j) fijo. Consideramos el plano que pasa por el origen y contiene a los dos vectores p_i y p_j. La probabilidad de que un hiperplano separe a los dos vectores es la misma que la probabilidad de que un diámetro en este plano los separe. Esta probabilidad es

$$GW(K) = \Pr[(i,j) \in \delta(K)] = \frac{\angle(p_i, p_j)}{\pi} = \frac{\arccos(p_i \cdot p_j)}{\pi}.$$

Por lo tanto, el valor esperado del corte es

$$E[w(\delta(K))] = \sum_{i,j} w_{ij} \frac{\arccos(p_i \cdot p_j)}{\pi}$$

y queremos comparar este valor

$$SDP(K) = \sum_{i,j} w_{ij} \left(\frac{1}{2} - \frac{1}{2} p_i \cdot p_j\right) \geq OPT(K).$$

Para comparar ambas expresiones término a término, consideramos $\rho = p_i \cdot p_j$, $-1 \leq \rho \leq 1$ y graficamos las funciones $f_1(\rho) = \frac{\arccos(p_i \cdot p_j)}{\pi}$ y $f_2(\rho) = \frac{1}{2} - \frac{1}{2} p_i \cdot p_j$.

Calculando la mayor diferencia entre las dos funciones, obtenemos que $f_1(\rho) \geq 0.87584 f_2(rho)$ y por lo tanto:

$$GW(K) \geq 0.87584 SDP(K),$$

y esto implica que existe algún corte $K^\star$ tal que $w(\delta(K^\star)) \geq 0.87584$ del valor óptimo del problema SDP.

Para el pentágono con todos los pesos de las aristas iguales a 1, la razón entre el valor óptimo del problema max-cut y la relajación SDP es aproximadamente 0.884, por lo tanto la cota obtenida anteriormente esta cerca de la mejor cota que podemos obtener.

Ejercicio: ¿qué cota podemos obtener en la otra dirección? Es decir, ¿puede ser que el valor obtenido por la relajación SDP coincida con el valor óptimo del problema max-cut? ¿O cuál es la mejor cota que podemos obtener?

3.3. Teoría de control en sistemas dinámicos

Referencia principal: [Blekherman et al., 2013, Sección 2.2.1].

Una de las primeras y más importantes aplicaciones de optimización semidefinida es en la teoría de control. En este caso la programación semidefinida nos permite caracterizar propiedades dinámicas (por ejemplo, estabilidad) en términos de relaciones algebraicas, específicamente como la factibilidad de sistemas de inecuaciones.

Comenzamos con un ejemplo muy sencillo que nos permite darnos una idea de las características de problemas más complicados.

3.3.1. Estabilidad de sistemas lineales

Consideremos una relación de recurrencia lineal,

$$\boldsymbol{x}[k+1] = \boldsymbol{A}\boldsymbol{x}[k], \quad \boldsymbol{x}[0] = \boldsymbol{x}_0,$$

con $\boldsymbol{x}[k] \in \mathbb{R}^n$ para todo $k \in \mathbb{N}_0$ y $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.

Esta relación es un ejemplo simple de un sistema dinámico discreto, el estado $\boldsymbol{x}[k]$ evoluciona con el tiempo a partir de un estado inicial $\boldsymbol{x}[0]$. El análogo continuo está dado por la ecuación diferencial

$$\frac{d}{dt}\boldsymbol{x}(u) = \boldsymbol{A}\boldsymbol{x}(t).$$

Estos modelos son utilizados para modelar la evolución en el tiempo de cantidades tales como temperatura, tamaño de la población, etc.

Una pregunta natural e importante es el comportamiento en el largo plazo del vector de estados. En particular, queremos determinar condiciones sobre la matriz $\boldsymbol{A}$ que permitan garantizar que el vector de estados se mantenga acotado o converja a 0.

Calculando autovalores y autovectores, sabemos que $\boldsymbol{x}[k]$ converge a 0 para todo estado inicial si y solo si el radio espectral $\rho(\boldsymbol{A}) < 1$, es decir si los autovalores λ_i de $\boldsymbol{A}$ cumplen $|\lambda_i| < 1$ para todo $1 \leq i \leq n$. En este caso decimos que el sistema (o la matriz $\boldsymbol{A}$ es estable).

Veremos ahora una forma alternativa de estudiar el problema, que resulta en ciertas ocasiones más conveniente.

Para una matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ positiva semidefinida, definimos la función V,

$$V(\boldsymbol{x}[k]) = \boldsymbol{x}[k]^T \boldsymbol{P}\boldsymbol{x}[T].$$

Vamos a ver que el sistema es asintóticamente estable si existe $\boldsymbol{P} \succ 0$ tal que V es no-decreciente sobre las trayectorias del sistema. Es decir, si $V(\boldsymbol{x}[k+1]) < V(\boldsymbol{x}[k])$ para todos los estados $\boldsymbol{x}[k]$. Observamos primero que esto es equivalente a la desigualdad de matrices

$$\boldsymbol{A}^T\boldsymbol{P}\boldsymbol{A} - \boldsymbol{P} \prec 0.$$

Podemos probar ahora el siguiente resultado.

Teorema 3.3. *Dada una matriz $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, las siguientes condiciones son equivalentes:*

(a) $\rho(\boldsymbol{A}) < 1$

(b) Existe una matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ simétrica tal que

$$\boldsymbol{P} \succ 0, \qquad \boldsymbol{A}^T\boldsymbol{P}\boldsymbol{A} - \boldsymbol{P} \prec 0.$$

Demostración. (b) ⇒ (a): Dado $\boldsymbol{v} \in \mathbb{C}^n$, $\boldsymbol{v} \neq 0$, autovector de $\boldsymbol{A}$, $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$,

$$0 > \boldsymbol{v}^*(\boldsymbol{A}^T\boldsymbol{P}\boldsymbol{A} - \boldsymbol{P})\boldsymbol{v} = (|\lambda|^2 - 1)\boldsymbol{v}^*\boldsymbol{P}\boldsymbol{v},$$

donde $\boldsymbol{v}^*\boldsymbol{P}\boldsymbol{v} > 0$ y por lo tanto $|\lambda| < 1$.

(a) ⇒ (b): Tomamos $\boldsymbol{P} = \sum_{k=0}^{\infty}(\boldsymbol{A}^k)^T\boldsymbol{A}^k$ (como $\rho(\boldsymbol{A}) < 1$, la suma converge). Luego

$$\boldsymbol{A}^T\boldsymbol{P}\boldsymbol{A} - \boldsymbol{P} = \sum_{k=1}^{\infty}(\boldsymbol{A}^k)^T\boldsymbol{A}^k - \sum_{k=0}^{\infty}(\boldsymbol{A}^k)^T\boldsymbol{A}^k = -\boldsymbol{I} \prec 0.$$

□

De esta forma convertimos el problema en un problema de factibilidad de SDP. Observamos que las coordenadas de la matriz $\boldsymbol{A}^T\boldsymbol{P}\boldsymbol{A} - \boldsymbol{P}$ son lineales afines en los coeficientes de $\boldsymbol{P}$, por lo tanto el problema es efectivamente un problema SDP.

3.3.2. Diseño de control

Consideramos ahora una extensión del problema anterior en la que agregamos una señal de control $\boldsymbol{u}[k] \in \mathbb{R}^m$:

$$\boldsymbol{x}[k+1] = \boldsymbol{A}\boldsymbol{x}[k] + \boldsymbol{B}\boldsymbol{u}[k], \quad \boldsymbol{x}[0] = \boldsymbol{x}_0,$$

con $\boldsymbol{B} \in \mathbb{R}^{n\times m}$. La función del término de control es poder ajustar el comportamiento de $\boldsymbol{x}[k]$ para lograr un cierto objetivo. Analizamos en particular el caso en que $\boldsymbol{A}$ no es estable, pero podemos usar un control lineal $\boldsymbol{u}[k] = \boldsymbol{K}\boldsymbol{x}[k]$ para una matriz fija $\boldsymbol{K}$ (a elegir apropiadamente). En este caso, podemos escribir el sistema de la forma

$$\boldsymbol{x}[k+1] = (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K})\boldsymbol{x}[k], \quad \boldsymbol{x}[0] = \boldsymbol{x}_0,$$

que es equivalente al problema original reemplazando a la matriz $\boldsymbol{A}$ por $\boldsymbol{A}+\boldsymbol{B}\boldsymbol{K}$. Este es un problema de una dificultad mayor al anterior, debido a que los autovalores de $\boldsymbol{A}+\boldsymbol{B}\boldsymbol{K}$ dependen en forma no-lineal de los autovalores de la matriz a calcular $\boldsymbol{K}$. Sin embargo, vamos a ver que podemos resolver este problema por optimización semidefinida usando la caracterización alternativa de Lyapunov.

Utilizando complementos de Schur, podemos reescribir las condiciones

$$(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K})^T\boldsymbol{P}(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}) - \boldsymbol{P} \prec 0, \quad \boldsymbol{P} \succ 0,$$

como

$$\begin{pmatrix} \boldsymbol{P} & (\boldsymbol{A}+\boldsymbol{BK})^T\boldsymbol{P} \\ \boldsymbol{P}(\boldsymbol{A}+\boldsymbol{BK}) & \boldsymbol{P} \end{pmatrix} \succ 0.$$

Esta formulación no un problema SDP debido a los términos no-lineales en las entradas de $\boldsymbol{K}$ y $\boldsymbol{P}$, las matrices a calcular.

Sin embargo, definiendo $\boldsymbol{Q} = \boldsymbol{P}^{-1}$ y multiplicando a izquierda y derecha por la matriz

$$\begin{pmatrix} \boldsymbol{Q} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Q} \end{pmatrix},$$

obtenemos la restricción equivalente

$$\begin{aligned} \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{Q}(\boldsymbol{A}+\boldsymbol{BK})^T \\ (\boldsymbol{A}+\boldsymbol{BK})\boldsymbol{Q} & \boldsymbol{Q} \end{pmatrix} &= \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{QA}^T + \boldsymbol{QK}^T\boldsymbol{B}^T \\ \boldsymbol{AQ}+\boldsymbol{BKQ} & \boldsymbol{Q} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{QA}^T + (\boldsymbol{KQ})^T\boldsymbol{B}^T \\ \boldsymbol{AQ}+\boldsymbol{BKQ} & \boldsymbol{Q} \end{pmatrix} \succ 0 \end{aligned}$$

(en la última igualdad utilizamos que $\boldsymbol{Q}$ es simétrica).

Si bien parece que no ganamos mucho con esta transformación, observamos ahora que la matriz $\boldsymbol{K}$ siempre aparece multiplicada por $\boldsymbol{Q}$ a derecha. Por lo tanto, definiendo $\boldsymbol{Y} = \boldsymbol{KQ}$, obtenemos la condición

$$\begin{pmatrix} \boldsymbol{Q} & \boldsymbol{QA}^T + \boldsymbol{Y}^T\boldsymbol{B}^T \\ \boldsymbol{AQ}+\boldsymbol{BY} & \boldsymbol{Q} \end{pmatrix} \succ 0. \tag{3.1}$$

En este caso el problema es lineal en las variables $(\boldsymbol{Q}, \boldsymbol{Y})$, y por lo tanto es un problema SDP. Luego de resolver este problema, podemos recuperar la matriz $\boldsymbol{K}$ por la fórmula $\boldsymbol{K} = \boldsymbol{Q}^{-1}\boldsymbol{Y}$.

Resumimos lo obtenido en el siguiente resultado.

Teorema 3.4. *Dadas matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ y $\boldsymbol{B} \in \mathbb{R}^{n\times m}$, existe una matriz $\boldsymbol{K} \in \mathbb{R}^{m\times n}$ tal que $\boldsymbol{A}+\boldsymbol{BK}$ es estable si y solo si el espectrahedro definido por la ecuación 3.1 es no vacío, es decir, existen matrices $(\boldsymbol{Q}, \boldsymbol{Y})$ tales que se satisface la desigualdad matricial estricta.*

Concluimos entonces que el problema de control planteado es equivalente a un problema de programación semidefinida.

3.3.3. Caso continuo

Consideramos un sistema dinámico autónomo

$$\dot{\boldsymbol{x}}(t) = \frac{d}{dt}\boldsymbol{x}(t) = f(\boldsymbol{x}(t)), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0,$$

y suponemos que se cumple

$$\dot{\boldsymbol{x}}(t) \in \operatorname{conv}\{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m\}\boldsymbol{x}(t),$$

donde $\operatorname{conv}\{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m\} = \{\alpha_1\boldsymbol{A}_1 + \cdots + \alpha_m\boldsymbol{A}_m : 0 \le \alpha_i \le 1, \sum \alpha_i = 1\}$.

Ejemplo. Si $m = 1$, obtenemos el sistema $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t)$.

La condición más general nos permite indicar que las derivadas se mueven en ciertos rangos.

Queremos determinar si $\boldsymbol{x}(t)$ se mantiene acotada, es decir, si el equilibrio $\boldsymbol{x} = \boldsymbol{0}$ es estable.

Observamos que $\boldsymbol{x}(t)$ se mantiene acotada si y solo si existe $\boldsymbol{P} \succ 0$ tal que

$$v : \mathbb{R}^n \to \mathbb{R}, v(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x}$$

se mantiene acotada sobre las trayectorias.

Una condición para asegurar esto es pedir que v sea decreciente sobre las trayectorias.

Estas funciones se conocen como *funciones de Lyapunov*.

Utilizando la regla de derivación de un producto interno para $f, g : \mathbb{R} \to \mathbb{R}^n$:

$$\frac{d}{dt}(f(t) \cdot g(t)) = \dot{f}(t) \cdot g(t) + f(t) \cdot \dot{g}(t),$$

obtenemos que $\boldsymbol{x}(t)$ se mantiene acotada si

$$\dot{v}(\boldsymbol{x}) = \frac{d}{dt}\boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x} = \dot{\boldsymbol{x}}^T \boldsymbol{P} \boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{P} \dot{\boldsymbol{x}} \leq 0.$$

Si $\boldsymbol{x}(0)$ es arbitrario y $\dot{\boldsymbol{x}}(0)$ puede estar en cualquier punto del conjunto convexo, necesitamos

$$\boldsymbol{A}_i^T \boldsymbol{P} + \boldsymbol{P} \boldsymbol{A}_i \preceq 0, \quad \text{para todo } 1 \leq i \leq m.$$

3.3.4. Problema SDP

Las restricciones obtenidas corresponden a un problema SDP.

Si buscamos $\boldsymbol{P} \succ 0$ bien condicionada, definimos el siguiente problema SDP:

$$\begin{array}{ll} \text{minimizar:} & \eta \\ \text{sujeto a:} & \boldsymbol{A}_i^T \boldsymbol{P} + \boldsymbol{P} \boldsymbol{A}_i \preceq 0, \quad \text{para todo } 1 \leq i \leq m \\ & \eta \boldsymbol{I} \succeq \boldsymbol{P} \succeq \boldsymbol{I} \end{array}$$

donde las variables del problema son η y las entradas de $\boldsymbol{P}$.

3.4. Conjuntos estables en grafos

Referencia principal: [Blekherman et al., 2013, Sección 2.2.3].

El código genético está compuesto por secuencias de bases, que podemos representar por las letras A, C, D y G. En algunos problemas de diseño de código genético es importante evitar las repeticiones en el código.

Si pensamos a distintas secuencias de código genético como nodos en un grafo, y unimos con aristas los pares de secuencias que presentan alguna repetición, el problema de evitar repeticiones se traduce en seleccionar la mayor cantidad de nodos tales que no haya dos nodos conectados entre sí.

En la teoría de grafos, esto se traduce como un subgrafo estable.

Dado un grafo no dirigido $G = (V, E)$, un conjunto estable (o conjunto independiente) es un subconjunto de $S \subset V$ tal que el grafo inducido por S no tiene ninguna arista, es decir, no hay dos vértices de S conectados por una arista en E.

El número de estabilidad de un grafo, que notamos $\alpha(G)$, es el cardinal del mayor conjunto estable. Calcular el número de estabilidad de un grafo es en general NP-hard. Veremos ahora como obtener una relajación SDP que nos permite dar una cota superior de $\alpha(G)$.

Dado un conjunto $S \subset V$, definimos el vector indicador $\chi(S)$,

$$\chi(S) = \begin{cases} 1 & \text{si } v_i \in S \\ 0 & \text{si no} \end{cases}.$$

Observamos que la matriz $\boldsymbol{Y} = \chi(S)\chi(S)^T$ cumple las siguientes propiedades:

- $\boldsymbol{Y} \succeq 0$.
- $y_{ij} = 1$ si y solo si $v_i \in S$ y $v_j \in S$. Por lo tanto, $\sum_{i,j} y_{ij} = |S|^2$.
- En particular $y_{ii} = 1$ si y solo si $v_i \in S$. Por lo tanto, $\mathrm{Tr}(\boldsymbol{Y}) = \sum_i y_{ii} = |S|$.

Tomando $X = \frac{1}{|S|}\boldsymbol{Y}$, vemos que $\sum_{i,j} \boldsymbol{x}_{ij} = |S|$ y $\mathrm{Tr}(\boldsymbol{X}) = 1$. Luego $\boldsymbol{X}$ es una solución factible del siguiente problema de programación semidefinida:

$$\begin{array}{ll} \text{maximizar:} & \mathrm{Tr}\, JX \\ \text{sujeto a:} & \mathrm{Tr}\, X = 1 \\ & X_{ij} = 0 \text{ para } (i,j) \in E, \\ & X \succeq 0, \end{array}$$

donde $J \in \mathbb{R}^{n \times n}$ es la matriz con 1 en todas las coordenadas.

Para un grafo G definimos la función theta de Lovász $\vartheta(G)$ como el valor óptima el problema SDP dado. Como la matriz X construida a partir de un subconjunto $S \subset V$ es una solución factible, obtenemos que

$$\alpha(G) \leq \vartheta(G),$$

es decir $\vartheta(G)$ es una cota superior de $\alpha(G)$, el número de estabilidad de G.

3.5. Distancia euclídea

Consideramos un grafo $G = (N, E)$ con nodos $N = \{1, \dots, n\}$, aristas $E \subset N \times N$, y pesos no negativos $D = \{d_{ij}\} \in \mathbb{R}^E_+$, que representan distancias entre los nodos.

Decimos que (G, D) es k-realizable si podemos ubicar los nodos de G en puntos $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n \in \mathbb{R}^k$ de forma tal que las distancias euclídeas entre los nodos respeten las longitudes dadas:

$$\exists \boldsymbol{v}_1, \dots, \boldsymbol{v}_n \in \mathbb{R}^k : \|\boldsymbol{v}_i - \boldsymbol{v}_j\| = d_{ij} \quad \forall \{i, j\} \in E.$$

Decimos que (G, D) es realizable (a secas) si existe k tal que (G, D) es k-realizable.

3.5.1. Grafo completo

Para el caso de un grafo completo tenemos la siguiente caracterización.

Teorema 3.5. *Sea $G = K_n$ un grafo completo, con pesos D, y sea $\boldsymbol{A} \in \mathcal{S}^n$ la matriz*

$$\boldsymbol{A} = \begin{pmatrix} 0 & d_{21}^2 & d_{31}^2 & \dots & d_{n1}^2 \\ d_{21}^2 & 0 & d_{32}^2 & \dots & d_{n2}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1}^2 & d_{n2}^2 & d_{n3}^2 & \dots & 0 \end{pmatrix}.$$

El grafo G es realizable si y solo si $\boldsymbol{A}$ es semidefinida negativa en el espacio ortogonal al vector $\boldsymbol{e} = (1, 1, \dots, 1)$. Es decir, si

$$\boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y} \leq 0 \quad \text{para todo } \boldsymbol{y} \in \mathbb{R}^n \text{ tal que } \sum_{i=1}^n y_i = 0.$$

Demostración. Demostramos solo $\Rightarrow$. Suponemos que para algún $k \geq 1$ existen vectores $\boldsymbol{v}_i \in \mathbb{R}^k$, $1 \leq i \leq n$, tales que $d_{ij} = \|\boldsymbol{v}_i - \boldsymbol{v}_j\|$. Consideramos ahora la matriz $\boldsymbol{X}$ de productos internos

$$\boldsymbol{X} = \begin{pmatrix} \langle \boldsymbol{v}_1, \boldsymbol{v}_1 \rangle & \langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle & \dots & \langle \boldsymbol{v}_1, \boldsymbol{v}_n \rangle \\ \langle \boldsymbol{v}_2, \boldsymbol{v}_1 \rangle & \langle \boldsymbol{v}_2, \boldsymbol{v}_2 \rangle & \dots & \langle \boldsymbol{v}_2, \boldsymbol{v}_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{v}_n, \boldsymbol{v}_1 \rangle & \langle \boldsymbol{v}_n, \boldsymbol{v}_2 \rangle & \dots & \langle \boldsymbol{v}_n, \boldsymbol{v}_n \rangle \end{pmatrix} = (\boldsymbol{v}_1 \dots \boldsymbol{v}_n)^T (\boldsymbol{v}_1 \dots \boldsymbol{v}_n),$$

que es semidefinida positiva. Como $a_{ij} = \|\boldsymbol{v}_i - \boldsymbol{v}_j\|^2 = \langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle + \langle \boldsymbol{v}_j, \boldsymbol{v}_j \rangle - 2\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$, tenemos

$$\boldsymbol{A} = \operatorname{diag}(\boldsymbol{X}) \cdot \boldsymbol{e}^T + \boldsymbol{e} \cdot \operatorname{diag}(\boldsymbol{X}) - 2\boldsymbol{X}.$$

Por lo tanto, para $\boldsymbol{y} \perp \boldsymbol{e}$, $\boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y} = -2\boldsymbol{y}^T \boldsymbol{X} \boldsymbol{y} \leq 0$.

□

3.5.2. Grafos no completos

En general, para grafos no completos determinar si el grafo es realizable equivale a buscar una solución factible de un problema SDP.

Teorema 3.6. *Un grafo con pesos (G, D) es realizable si y solo si el siguiente problema SDP tiene solución*

$$\begin{aligned} &\textit{existe:} && \boldsymbol{X} \in \mathcal{S}^n \\ &\textit{sujeto a:} && x_{ii} + x_{jj} - 2x_{ij} = d_{ij}^2 \quad \forall \{i,j\} \in E, \\ & && \boldsymbol{X} \succeq 0 \end{aligned}$$

Más aún, (G, D) es k-realizable si existe una solución $\boldsymbol{X}$ de rango a lo sumo k.

Demostración. Si $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^k$ es una realización de (G, D), entonces la matriz de Gram

$$\boldsymbol{X} = (\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle) = (\boldsymbol{v}_1 \ldots \boldsymbol{v}_n)^T (\boldsymbol{v}_1 \ldots \boldsymbol{v}_n)$$

es una solución del problema, de rango a lo sumo k.

Recíprocamente, si $\boldsymbol{X}$ es una solución del problema, podemos encontrar una descomposición

$$\boldsymbol{X} = \boldsymbol{V}^T \boldsymbol{V}, \quad \boldsymbol{V} \in \mathbb{R}^{k \times n}$$

y tomamos como $\boldsymbol{v}_i$ las columnas de $\boldsymbol{V}$. □

3.5.3. Grafos completos revisitados

Utilizando el último teorema obtenemos otra caracterización para grafos completos.

Teorema 3.7. *Sea $G = K_n$ grafo completo, con pesos D, y sea $\boldsymbol{X} \in \mathcal{S}^{n-1}$ la matriz definida por*

$$\begin{aligned} x_{ii} &= d_{in}^2, \quad \forall 1 \le i \le n-1 \\ x_{ij} &= \frac{d_{in}^2 + d_{jn}^2 - d_{ij}^2}{2}, \quad \forall 1 \le i \ne j \le n-1. \end{aligned}$$

Entonces K_n es k-realizable si y solo si $\boldsymbol{X} \succeq 0$ y $\operatorname{rank} \boldsymbol{X} \le k$.

Idea de la demostración. Si el problema SDP tiene solución factible, mediante una traslación, podemos suponer $\boldsymbol{v}_n = \boldsymbol{0}$.

3.5.4. El problema de la partición

Como vimos, determinar si un grafo (G, D) es realizable, puede resolverse eficientemente por optimización semi-definida. Sin embargo, determinar si es k-realizable para un k dado es un problema mucho más difícil.

Veremos un caso simple en el que el problema es NP-completo.

Consideramos el siguiente problema para el cual se sabe que es NP-completo.

Problema 3.8 (El problema de la partición)**.** Dada una secuencia de números naturales $a_1, \ldots, a_n \in \mathbb{N}$, determinar si los números pueden separarse en dos conjuntos con la misma suma. Es decir, si existe $\boldsymbol{\epsilon} \in \{\pm 1\}^n$ tal que

$$\epsilon_1 a_1 + \cdots + \epsilon_n a_n = 0.$$

3.5.5. Grafo ciclo

Un grafo (G, E) se llama ciclo o cíclico si las aristas forma un ciclo de longitud n. Es decir, podemos suponer que las aristas son $(i, i+1)$ para $1 \leq i \leq n-1$ y $(n, 1)$.

Teorema 3.9. *Dado un grafo ciclo (G, E) con pesos naturales $d \in \mathbb{N}^E$, decidir si (G, D) es 1-realizable es un problema NP-completo.*

Demostración. Para una instancia $a_1, \ldots, a_n \in \mathbb{N}$ del problema de la partición, consideramos el grafo ciclo (G, E) con pesos $d_{i(i+1)} = a_i$.

Si (G, D) es 1-realizable, con $v_i \in \mathbb{R}$, definimos

$$\begin{cases} \epsilon_i = 1 & \text{si } v_{i+1} > v_i \\ \epsilon_i = -1 & \text{si } v_{i+1} < v_i \end{cases}$$

y obtenemos una partición de los a_i. □

3.5.6. Rango 1 y rango 2

Puede demostrarse la siguiente propiedad:

Un grafo ciclo es realizable si y solo si es 2-realizable.

Ejercicio 3.10. Ver geométricamente que puedo llevar una realización 3D de un grafo ciclo a una realización 2D.

Concluimos que es posible contestar eficientemente si un grafo ciclo es realizable en un plano o una recta, pero determinar en cuál de los dos es un problema NP-completo.

3.6. Minimización de rango

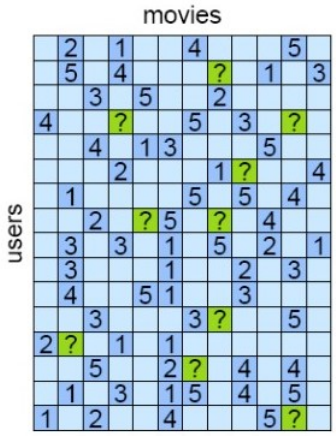
Como vimos, un problema interesante en optimización es el problema de la minimización de rango, que podemos plantear en la forma

$$\begin{array}{ll} \text{minimizar:} & \operatorname{rank} \boldsymbol{X} \\ \text{sujeto a:} & \boldsymbol{X} \in \mathcal{C}, \end{array}$$

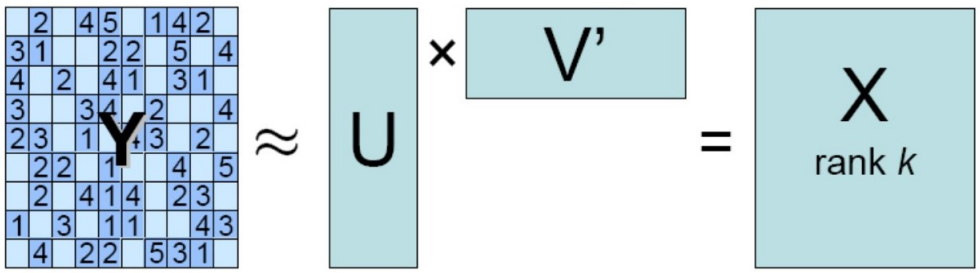
donde la matriz $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ es la variable de decisión, y $\mathcal{C}$ es un conjunto convexo. Como la función a optimizar tiene valores enteros, este problema en general no es un problema convexo. En el ejemplo anterior vimos que en general es un problema NP-completo.

3.6.1. Ejemplo - El problema de Netflix

- Algunos usuarios califican algunas de las películas que vieron.
- Los puntajes son números enteros entre 1 y 5.
- Queremos predecir los puntajes para cada par usuario / película.
- Es decir, queremos completar una matriz incompleta como la de la figura.



Si pensamos que hay unos pocos perfiles de usuarios arquetípicos (amante de las películas de terror, románticas, comedias, etc.) y cada usuario es una combinación lineal de esos perfiles, podemos factorizar la matriz:



3.6.2. Rango y valores singulares

- Para una matriz $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ cuadrada, el rango de $\boldsymbol{A}$ es igual a la cantidad de autovalores no-nulos.

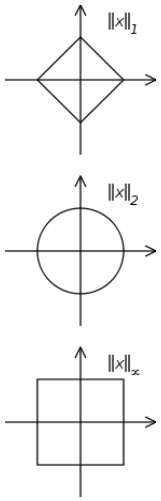
- Para una matriz $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, el rango de $\boldsymbol{A}$ es igual a la cantidad de valores singulares no-nulos.

Heurística

- Como no podemos minimizar el rango eficientemente, minimizamos la suma de los valores singulares.

- Para $\boldsymbol{A} \succeq 0$, la suma de los valores singulares es igual a $\mathrm{Tr}(\boldsymbol{A})$, que es una función lineal en los coeficientes de $\boldsymbol{A}$.

3.6.3. Norma nuclear



En $\mathbb{R}^{m\times n}$ definimos la norma nuclear

$$\|\boldsymbol{X}\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i.$$

Como los valores singulares son todos no-negativos,

$$\|\boldsymbol{X}\|_* = \|\boldsymbol{\sigma}\|_1.$$

Comparando las curvas de nivel distintas normas, vemos que minimizando la norma-1 obtenemos en general vectores esparsos con gran cantidad de 0's. Por este motivo es una buena elección para obtener matrices de rango bajo.

Para analizar las propiedades de la norma nuclear, recordemos las siguientes propiedades:

- Para $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$,
$$\boldsymbol{A} \bullet \boldsymbol{B} = \operatorname{Tr}(\boldsymbol{A}^T\boldsymbol{B}) = \operatorname{Tr}(\boldsymbol{A}\boldsymbol{B}^T)$$
es un producto interno en $\mathbb{R}^{n\times n}$, que define la norma
$$\|\boldsymbol{A}\|_F = \sqrt{\boldsymbol{A}\bullet\boldsymbol{A}} = \sqrt{\operatorname{Tr}(\boldsymbol{A}\boldsymbol{A}^T)}.$$
Si $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{A}^T$, $\operatorname{Tr}(\boldsymbol{X}) = \|\boldsymbol{A}\|_F^2$.
- Desigualdad de Cauchy-Schwarz: $|\langle \boldsymbol{u}, \boldsymbol{v}\rangle| \le \|\boldsymbol{u}\|\|\boldsymbol{v}\|$
- Desigualdad MA-MG: $\|\boldsymbol{u}\|\|\boldsymbol{v}\| \le \frac{1}{2}(\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2)$

Lema 3.11. *Para $\boldsymbol{X} \in \mathbb{R}^{n\times n}$,*

$$\|\boldsymbol{X}\|_* = \min_{\boldsymbol{X}=\boldsymbol{U}\boldsymbol{V}^T} \|\boldsymbol{U}\|_F\|\boldsymbol{V}\|_F = \min_{\boldsymbol{X}=\boldsymbol{U}\boldsymbol{V}^T} \frac{1}{2}(\|\boldsymbol{U}\|^2 + \|\boldsymbol{V}\|^2).$$

Este y los siguientes resultados valen también para $\boldsymbol{X}$ rectangular.

Demostración. Tomamos una descomposición SVD de $\boldsymbol{X}$,

$$\boldsymbol{X} = \boldsymbol{P}\boldsymbol{S}\boldsymbol{Q}^T,$$

con $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ unitaria, $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ diagonal, con los valores singulares $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n \ge 0$ en la diagonal y $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ unitaria.

Si $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^T$, con $\boldsymbol{U} \in \mathbb{R}^{n\times k}$, $\boldsymbol{V} \in \mathbb{R}^{n\times k}$, entonces

$$\begin{aligned}\|\boldsymbol{X}\|_* &= \operatorname{Tr}(\boldsymbol{S}) = \operatorname{Tr}(\boldsymbol{P}^T\boldsymbol{U}\boldsymbol{V}^T\boldsymbol{Q}) \\ &= \boldsymbol{P}^T\boldsymbol{U} \bullet \boldsymbol{Q}^T\boldsymbol{V} \le \|\boldsymbol{P}^T\boldsymbol{U}\|_F\|\boldsymbol{Q}^T\boldsymbol{V}\|_F \\ &= \|\boldsymbol{U}\|_F\|\boldsymbol{V}\|_F \le \frac{1}{2}(\|\boldsymbol{U}\|_F + \|\boldsymbol{V}\|_F),\end{aligned}$$

donde $\boldsymbol{P}^T\boldsymbol{U}, \boldsymbol{Q}^T\boldsymbol{V} \in \mathbb{R}^{n\times k}$.

Tomando $\boldsymbol{U} = \boldsymbol{P}\boldsymbol{S}^{\frac{1}{2}}$ y $\boldsymbol{V} = \boldsymbol{Q}\boldsymbol{S}^{\frac{1}{2}}$, alcanzamos el mínimo. □

Utilizando el lema, obtenemos una caracterización de la norma nuclear por un problema SDP.

Lema 3.12. *Para cualquier matriz* $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ *y* $t \in \mathbb{R}$, $\|\boldsymbol{X}\|_* \leq t$ *si y solo si existen* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *y* $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ *tales que*

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{B} \end{pmatrix} \succeq 0 \quad y \quad \operatorname{Tr}(\boldsymbol{A}) + \operatorname{Tr}(\boldsymbol{B}) \leq 2t.$$

Demostración. Si $\begin{pmatrix} \boldsymbol{A} & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{B} \end{pmatrix} \succeq 0$, podemos factorizarla

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{B} \end{pmatrix} = \begin{pmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{pmatrix} \begin{pmatrix} \boldsymbol{U}^T & \boldsymbol{V}^T \end{pmatrix},$$

con $\boldsymbol{U} \in \mathbb{R}^{n\times(n+n)}$ y $\boldsymbol{V} \in \mathbb{R}^{n\times(n+n)}$.

Tenemos $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{U}^T$, $\boldsymbol{B} = \boldsymbol{V}\boldsymbol{V}^T$ y $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^T$.

Por lo tanto, $\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2 = \operatorname{Tr}\boldsymbol{A} + \operatorname{Tr}\boldsymbol{B} \leq 2t$, y obtenemos

$$\|\boldsymbol{X}\|_* \leq \frac{1}{2}(\|\boldsymbol{U}\|^2 + \|\boldsymbol{V}\|^2) \leq t.$$

Recíprocamente, si $\|\boldsymbol{X}\|_* \leq t$, y $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{S}\boldsymbol{Q}^T$ es una descomposición SVD, tomando $\boldsymbol{U} = \boldsymbol{P}\boldsymbol{S}^{\frac{1}{2}}$ y $\boldsymbol{V}^T = \boldsymbol{S}^{\frac{1}{2}}\boldsymbol{Q}^T$, obtenemos

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{pmatrix} \begin{pmatrix} \boldsymbol{U}^T & \boldsymbol{V}^T \end{pmatrix} = \begin{pmatrix} \boldsymbol{U}\boldsymbol{U}^T & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{V}\boldsymbol{V}^T \end{pmatrix} \succeq 0$$

y $\operatorname{Tr}(\boldsymbol{Z}) = \operatorname{Tr}(\boldsymbol{U}\boldsymbol{U}^T) + \operatorname{Tr}(\boldsymbol{V}\boldsymbol{V}^T) \leq 2t$.

□

3.6.4. Norma nuclear como problema SDP

Extendiendo los resultados anteriores a matrices rectangulares (ejercicio), obtenemos que la norma nuclear $\|\boldsymbol{X}\|_*$, $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, se corresponde con el valor óptimo del problema SDP

$$\begin{aligned} \text{minimizar:} \quad & \frac{1}{2}\operatorname{Tr}\begin{pmatrix} \boldsymbol{A} & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{B} \end{pmatrix} \\ \text{sujeto a:} \quad & \begin{pmatrix} \boldsymbol{A} & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{B} \end{pmatrix} \succeq 0 \end{aligned}$$

para matrices $\boldsymbol{A} \in \mathbb{R}^{m\times m}$, $\boldsymbol{B} \in \mathbb{R}^{n\times n}$.

Para el problema de Netflix, obtenemos que el problema de completar una matriz $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ para la cual solo se conocen algunas casillas m_{ij}, $(i,j) \in I$ de forma tal que la norma nuclear sea mínima es equivalente a resolver el problema

$$\begin{aligned} \text{minimizar:} \quad & \frac{1}{2}\operatorname{Tr}\begin{pmatrix} \boldsymbol{A} & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{B} \end{pmatrix} \\ \text{sujeto a:} \quad & x_{ij} = m_{ij}, (i,j) \in I, \\ & \begin{pmatrix} \boldsymbol{A} & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{B} \end{pmatrix} \succeq 0 \end{aligned}$$

para matrices $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, $\boldsymbol{A} \in \mathbb{R}^{m\times m}$, $\boldsymbol{B} \in \mathbb{R}^{n\times n}$.

Más generalmente, podemos reemplazar la restricción $x_{ij} = m_{ij}, (i,j) \in I$, por $\boldsymbol{X} \in \mathcal{C}$, para $\mathcal{C}$ un conjunto convexo (definido por restricciones lineales).

3.6.5. Ejemplo: sumas de cuadrados

Determinar si un polinomio $f \in \mathbb{R}[x_1, x_2, x_3]$ homogéneo de grado 4 es suma de cuadrados, es equivalente a determinar si existe una matrix $\boldsymbol{A}$ semidefinida positiva tal que

$$f = \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v},$$

para $\boldsymbol{v} = \big(x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3\big)$.

La escritura como suma de cuadrados se puede obtener factorizando $\boldsymbol{A} = \boldsymbol{X}^T \boldsymbol{X}$ y el rango de $\boldsymbol{A}$ nos dice la cantidad de polinomios linealmente independientes en la descomposición.

Si queremos estudiar el problema de hallar la menor cantidad de polinomios que aparecen en una descomposición, debemos minimizar el rango de la matriz $\boldsymbol{A}$.

3.6.6. Norma nuclear de matrices simétricas

Si trabajamos con matrices simétricas, minimizar la norma nuclear equivale a resolver el problema SDP:

$$\begin{array}{ll} \text{minimizar:} & \operatorname{Tr} \boldsymbol{X} \\ \text{sujeto a:} & \boldsymbol{X} \in \mathcal{C}, \\ & \boldsymbol{X} \succeq 0. \end{array}$$

Capítulo 4

Polinomios positivos y sumas de cuadrados

4.1. Introducción

Notamos $\mathbb{R}[\boldsymbol{x}] = \mathbb{R}[x_1, \ldots, x_n]$ al anillo de polinomios sobre $\mathbb{R}$ en n variables.

Dado un polinomio $p(\boldsymbol{x}) \in \mathbb{R}[\boldsymbol{x}]$, decimos que

- p es *positivo* ($p \geq 0$) si $p(\boldsymbol{x}) \geq 0$ para todo $\boldsymbol{x} \in \mathbb{R}^n$ (y *estrictamente positivo* si $p > 0$ para todo $\boldsymbol{x} \in \mathbb{R}^n$).
- p es una *suma de cuadrados* (SOS) si existen $q_1, \ldots, q_s \in \mathbb{R}[\boldsymbol{x}]$ tales que

$$p = q_1^2 + \cdots + q_s^2.$$

Proposición 4.1. *Si p es SOS entonces $p \geq 0$.*

Problema 4.2. Dado un polinomio $p(\boldsymbol{x}) \in \mathbb{R}[\boldsymbol{x}]$, determinar si se puede escribir como suma de cuadrados (SOS) $p = p_1^2 + \cdots + p_s^2$ y construir la descomposición (aproximada o exacta).

Algunas aplicaciones de sumas de cuadrados son:

- certificados de positividad,
- La ecuación $p(\boldsymbol{x}) + 1 = 0$ no tiene soluciones reales si p es una suma de cuadrados,
- Dado un polinomio p, si queremos hallar el mínimo de p en $S = \mathbb{R}^n$ o en una región $S = \{\boldsymbol{x} \in \mathbb{R}^n : g(\boldsymbol{x}) \geq 0\}$, para polinomios $\{g_1, \ldots, g_s\}$, tenemos

$$\text{mín}\{f(x) : \boldsymbol{x} \in S\} = \sup\{a \in \mathbb{R} | f - a \geq 0 \text{ en } S\}.$$

4.2. Sumas de cuadrados en una variable

Proposición 4.3. *Si $p \in \mathbb{R}[x]$ (polinomios en una variable) entonces*

$$p \text{ es suma de cuadrados} \iff p \text{ es positivo}$$

Demostración. Tomamos $p \geq 0$. Por el teorema fundamental del álgebra, podemos factorizar

$$p(x) = \prod_{i=1}^{r}(x - a_i) \prod_{j=1}^{t}(x - b_j)(x - \bar{b}_j),$$

donde $a_i \in \mathbb{R}$ son las raíces reales y $b_j, \bar{b}_j \in \mathbb{C} \setminus \mathbb{R}$ son las raíces complejas conjugadas.

Si la multiplicidad de alguna raíz real a es impar, entonces p atraviesa transversalmente al eje X en a y por lo tanto no puede ser $p \geq 0$.

Por lo tanto, todas las raíces reales aparecen con multiplicidad par y podemos factorizar

$$p(x) = \prod_{i=1}^{s}(x - a_i)^{2k_i} \prod_{j=1}^{t}(x - b_j)(x - \bar{b}_j).$$

Para las raíces complejas tenemos

$$\begin{aligned}(x - b_j)(x - \bar{b}_j) &= (x - (\alpha_i + I\beta_i))(x - (\alpha_i - I\beta_i)) \\ &= ((x - \alpha_i) - I\beta_i))((x - \alpha_i) + I\beta_i)) \\ &= (x - \alpha_i)^2 + \beta_i^2,\end{aligned}$$

que es una suma de cuadrados.

Concluimos que $p(x)$ es un producto de sumas de cuadrados, y por lo tanto, distribuyendo los productos, $p(x)$ es una suma de cuadrados.

Más aún, utilizando la identidad

$$(a^2 + b^2)(c^2 + d^2) = (ac + bd)^2 + (ad - bc)^2$$

podemos escribir a cualquier polinomio $p(x) \geq 0$ como suma de 2 cuadrados.

□

4.3. Sumas de cuadrados en varias variables

Proposición 4.4. *El polinomio*

$$f(x, y) = x^4y^2 + x^2y^4 - 3x^2y^2 + 1$$

es no-negativo en $\mathbb{R}^2$ *pero no puede escribirse como suma de cuadrados. (Motzkin, 1967)*

Demostración. Veamos primero $f \geq 0$. Por la desigualdad aritmética-geométrica,

$$\frac{x^4y^2 + x^2y^4 + 1}{3} \geq \sqrt[3]{(x^4y^2)(x^2y^4)1} = \sqrt[3]{x^6y^6} = x^2y^2,$$

y despejando obtenemos $x^4y^2 + x^2y^4 - 3x^2y^2 + 1 \geq 0$

Para ver que el polinomio de Motzkin no es una suma de cuadrados, escribimos $f = p_1^2 + \cdots + p_s^2$, $p_i \in \mathbb{R}[x, y, z_1, ..., z_m]$. Vamos obteniendo condiciones sobre los posibles polinomios que pueden aparecer en la descomposición.

(a) Podemos evaluar $z_i = 0$ para todo i y obtenemos una descomposición en $\mathbb{R}[x, y]$.

(b) Si tomamos un monomio m en los p_i con el mayor grado d en x, el monomio m^2 no se va a cancelar en la suma, y por lo tanto debe ser un monomio de f.

(c) Por lo tanto, debe ser $d \leq 2$.

(d) Luego, siguiendo el mismo razonamiento, no puede aparecer x^2y^2 en ningún p_i.

(e) Luego, tampoco pueden aparecer x^2 ni y^2 en ningún p_i.

(f) Finalmente, no puede aparecer x ni y en ningún p_i.

(g) Concluimos que los polinomios p_i son de la forma

$$ax^2y + bxy^2 + cxy + d,$$

con $a, b, c, d \in \mathbb{R}$.

(h) Al elevar al cuadrado un polinomio de esta forma, el coeficiente de x^2y^2 es siempre no-negativo. ¡Absurdo!

□

Podemos extender el razonamiento anterior a casos más generales. Comenzamos realizando algunas definiciones.

Definición 4.5. *Dado un vector $\boldsymbol{a} \in \mathbb{N}_0^n$, $\boldsymbol{a} = (a_1, \ldots, a_n)$, definimos el monomio $m = \boldsymbol{x}^{\boldsymbol{a}}$ como*

$$\boldsymbol{x}^{\boldsymbol{a}} = x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n},$$

y análogamente, para un monomio m de esa forma, llamamos a $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ su vector de exponentes*.*

Definición 4.6. *Dado un polinomio $p \in \mathbb{K}[\boldsymbol{x}]$, definimos el* soporte *de p,* $\operatorname{supp}(p)$*, como el conjunto de todos los vectores de exponentes de los monomios que aparecen en p.*

Definimos su polítopo de Newton $\mathcal{N}(p)$ *como la cápsula convexa de los vectores de exponentes de los monomios que aparecen en p,*

$$\mathcal{N}(p) = \operatorname{conv}(\operatorname{supp}(p)).$$

Por ejemplo, si $p = x_1x_2^2 + x_2^2 + x_1x_2x_3$ entonces

$$\mathcal{N}(p) = \operatorname{conv}\left(\{(1,2,0), (0,2,0), (1,1,1)\}\right),$$

que es un triángulo en $\mathbb{R}^3$.

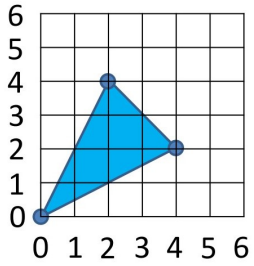
Utilizando el polítopo de Newton podemos obtener información sobre los monomios que pueden aparecer en una suma de cuadrados.

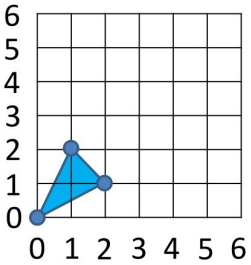
Teorema 4.7. *Si $p = \sum_{i=1}^s q_i^2$ es una suma de cuadrados, entonces*

$$\mathcal{N}(q_i) \subset \frac{1}{2}\mathcal{N}(p).$$

Demostración. Consideramos la cápsula convexa de la unión de todos los polítopos de Newton de los q_i, $1 \leq i \leq s$,

$$K = \operatorname{conv}(\cup_{i=1}^s \mathcal{N}(q_i)).$$





Recordamos que un polítopo está generado por las combinaciones convexas de sus vértices.

Tomamos un *vértice* $\boldsymbol{v}$ de K y suponemos por contradicción que $2\boldsymbol{v} \notin \operatorname{supp}(p)$.

Si $\alpha \boldsymbol{x}^{\boldsymbol{v}}$ aparece en q_i, entonces $\alpha^2 \boldsymbol{x}^{2\boldsymbol{v}}$ aparece en q_i^2 con coeficiente $\alpha^2 > 0$. Para que estos términos se cancelen, debe aparecer también $\boldsymbol{x}^{2\boldsymbol{v}}$ como producto cruzado de términos de algunos q_i.

Es decir, existen $\boldsymbol{u}, \boldsymbol{w} \in K$ tales que $2\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{w}$.

Pero luego, $\boldsymbol{v} = \frac{\boldsymbol{u}+\boldsymbol{w}}{2} \in K$ no es un vértice, lo que contradice la hipótesis.

Concluimos que para cualquier vértice $\boldsymbol{v}$ de K, $2\boldsymbol{v} \in \operatorname{supp}(p) \subset \mathcal{N}(p)$.

Como $2K$ es un polítopo, es la cápsula convexa del conjunto de todos sus vértices.

Como todos los vértices de $2K$ están contenidos en el conjunto convexo $\mathcal{N}(p)$, obtenemos $2K \subset \mathcal{N}(p)$.

Por lo tanto, $2\mathcal{N}(q_i) \subset \mathcal{N}(p)$ para todo $1 \leq i \leq n$.

□

Motzkin revisitado

Utilizando este resultado podemos simplificar la demostración de que el polinomio de Motzkin $f(x, y) = x^4y^2 + x^2y^4 - 3x^2y^2 + 1$ no es una suma de cuadrados.

A la izquierda representamos el polítopo de Newton $\mathcal{N}(f)$ y a la derecha $\frac{1}{2}\mathcal{N}(f)$.

Concluimos inmediatamente que $f = \sum_{i=1}^{s} q_i$, los q_i son de la forma

$$q_i(x, y) = ax^2y + bxy^2 + cxy + d.$$

Grado de un polinomio

Definimos el grado de un monomio $x_1^{a_1} \cdots x_n^{a_n}$ como

$$d = |\boldsymbol{a}| = a_1 + \cdots + a_n,$$

y el grado de un polinomio $p \in \mathbb{R}[\boldsymbol{x}]$ como el mayor de los grados de sus monomios.

Polinomios homogéneos

Decimos que un polinomios es homogéneo si todos sus monomios tienen el mismo grado.

Dado un polinomio no-homogéneo $f(x_1, \ldots, x_n)$ de grado d, definimos su homogeneización

$$F(x_0, x_1, \ldots, x_n) = x_0^d f\left(\frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\right)$$

que equivale a multiplicar cada monomio por la potencia de x_0 apropiada para que todos los términos tengan grado d.

Por ejemplo, homegeneizamos $f(x_1, x_2) = x_1^4x_2^2 + x_1^2x_2^4 - 3x_1^2x_2^2 + 1$ a

$$F(x_0, x_1, x_2) = x_1^4x_2^2 + x_1^2x_2^4 - 3x_0^2x_1^2x_2^2 + x_0^6$$

multiplicando cada término por la potencia de x_0 apropiada.

Como corolario del teorema anterior sobre los polítopos de Newton, obtenemos el siguiente resultado.

Proposición 4.8. *Si $p(\boldsymbol{x})$ es SOS homogéneo, entonces $p(\boldsymbol{x})$ tiene grado par $2d$ y es suma de cuadrados de polinomios homogéneos de grado d.*

Proposición 4.9. *Las condiciones de no-negatividad y suma de cuadrados se mantienen al homogeneizar, por lo tanto no perdemos generalidad al asumir polinomios homogéneos.*

Vamos a trabajar a partir de ahora con polinomios homogéneos.

Los conos de polinomios positivos y sumas de cuadrados

Llamamos

- $H_{n,2d}$ al espacio vectorial de polinomios homogéneos de n variables y grado $2d$.
- $P_{n,2d} \subset H_{n,2d}$ al conjunto de polinomios homogéneos positivos de n variables y grado $2d$.
- $\Sigma_{n,2d} \subset P_{n,2d}$ al subconjunto de sumas de cuadrados.

Proposición 4.10. *$P_{n,2d}$ y $\Sigma_{n,2d}$ son conos convexos cerrados de dimensión máxima.*

Ejercicio 4.11. Demostrar que $P_{n,2d}$ es cerrado escribiéndolo como intersección de infinitos semi-espacios cerrados.

4.4. El teorema de Hilbert

Teorema 4.12 (Teorema de Hilbert (1888))**.** *Los conjuntos $P_{n,2d}$ y $\Sigma_{n,2d}$ son iguales solo en los siguientes casos:*

(a) $n = 2$

(b) $2d = 2$

(c) $(n, 2d) = (3, 4)$

4.5. Problema de programación semidefinida

Dado $p \in \mathbb{R}[x_1, \ldots, x_n]$, homogéneo de grado $2d$, podemos escribirlo como un producto

$$p(\boldsymbol{x}) = \boldsymbol{v}(\boldsymbol{x})^T \boldsymbol{Q} \boldsymbol{v}(\boldsymbol{x}),$$

con $\boldsymbol{v}$ el vector de monomios de grado d y $\boldsymbol{Q} \in \mathbb{R}^{M(d)\times M(d)}$ simétrica, con $M(d) = \binom{n+d-1}{d}$, la cantidad de monomios de grado d en n variables

Esta ecuación nos una ecuación lineal para cada coeficiente de p, en total $\binom{n+2d-1}{2d}$ ecuaciones.

Ejemplo 4.13. Para el polinomio $p(x,y) = 10x^4 + 2x^3y + 27x^2y^2 - 24xy^3 + 5y^4$, planteamos la ecuación matricial $10x^4 + 2x^3y + 27x^2y^2 - 24xy^3 + 5y^4 =$

$$\begin{pmatrix} x^2 & xy & y^2 \end{pmatrix} \begin{pmatrix} q_{00} & q_{10} & q_{20} \\ q_{10} & q_{11} & q_{21} \\ q_{20} & q_{21} & q_{22} \end{pmatrix} \begin{pmatrix} x^2 \\ xy \\ y^2 \end{pmatrix}$$

y obtenemos que se debe cumplir la igualdad

$$\begin{aligned} &10x^4 + 2x^3y + 27x^2y^2 - 24xy^3 + 5y^4 = \\ &= q_{00}x^4 + 2q_{10}x^3y + (2q_{20} + q_{11})x^2y^2 + 2q_{21}xy^3 + q_{22}y^4. \end{aligned}$$

Igualando coeficiente a coeficiente

$$\begin{aligned} &10x^4 + 2x^3y + 27x^2y^2 - 24xy^3 + 5y^4 = \\ &= q_{00}x^4 + 2q_{10}x^3y + (2q_{20} + q_{11})x^2y^2 + 2q_{21}xy^3 + q_{22}y^4 \end{aligned}$$

obtenemos

$$\begin{aligned} q_{00} &= 10 \\ 2q_{10} &= 2 \\ 2q_{20} + q_{11} &= 27 \\ 2q_{21} &= -24 \\ q_{22} &= 5 \end{aligned}$$

Despejando, obtenemos

$10x^4 + 2x^3y + 27x^2y^2 - 24xy^3 + 5y^4 =$

$$\begin{pmatrix} x^2 & xy & y^2 \end{pmatrix} \begin{pmatrix} 10 & 1 & a \\ 1 & -2a+27 & -12 \\ a & -12 & 5 \end{pmatrix} \begin{pmatrix} x^2 \\ xy \\ y^2 \end{pmatrix}$$

para cualquier $a \in \mathbb{R}$, y todas las matrices que cumplen la igualdad son de esta forma.

Descomposición como combinación lineal de cuadrados

Tomamos por ejemplo $a = 1$ y diagonalizamos (descomposición LDL^t por eliminación gaussiana):

$$\begin{pmatrix} 10 & 1 & 1 \\ 1 & 25 & -12 \\ 1 & -12 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{10} & 1 & 0 \\ \frac{1}{10} & -\frac{121}{249} & 1 \end{pmatrix} \begin{pmatrix} 10 & 0 & 0 \\ 0 & \frac{249}{10} & 0 \\ 0 & 0 & -\frac{244}{249} \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{10} & \frac{1}{10} \\ 0 & 1 & -\frac{121}{249} \\ 0 & 0 & 1 \end{pmatrix}$$

Obtenemos la descomposición

$$f = 10\left(x^2 + \frac{1}{10}xy + \frac{1}{10}y^2\right)^2 + \frac{249}{10}\left(xy - \frac{121}{249}y^2\right)^2 - \frac{244}{249}\left(y^2\right)^2$$

No es una suma de cuadrados porque el último coeficiente es negativo.

Signatura y suma de cuadrados

La cantidad de valores positivos en la diagonal es igual a la cantidad de autovalores positivos. Por lo tanto, debemos hallar $a \in \mathbb{R}$ tal que $\boldsymbol{Q}$ sea semidefinida positiva $\longrightarrow$ *Problema de programación semidefinida.*

Concretamente, tenemos el siguiente resultado.

Proposición 4.14. *Si $p(\boldsymbol{x}) \in \mathbb{R}[\boldsymbol{x}]$ es un polinomio homogéneo de grado $2d$, las siguientes propiedades son equivalentes:*

(a) $p(\boldsymbol{x})$ es una suma de cuadrados,

(b) existe $\boldsymbol{Q} \succeq 0$ que satisface la fórmula $p(\boldsymbol{x}) = \boldsymbol{v}(\boldsymbol{x})^T \boldsymbol{Q} \boldsymbol{v}(\boldsymbol{x})$,

para $\boldsymbol{v}(\boldsymbol{x})$ el vector de monomios de grado d en $\mathbb{R}[\boldsymbol{x}]$.

Demostración. Para probar (2) $\Rightarrow$ (1), dada una matriz $\boldsymbol{Q} \succeq 0$, podemos factorizarla $\boldsymbol{Q} = \boldsymbol{L}^T \boldsymbol{L}$, con $\boldsymbol{L}$ triangular inferior y obtenemos

$$p(\boldsymbol{x}) = \boldsymbol{v}(\boldsymbol{x})^T \boldsymbol{Q} \boldsymbol{v}(\boldsymbol{x}) = \boldsymbol{v}(\boldsymbol{x})^T \boldsymbol{L}^T \boldsymbol{L} \boldsymbol{v}(\boldsymbol{x}) = \sum (\boldsymbol{L}_i \cdot \boldsymbol{v}(\boldsymbol{x}))^2 = \sum q_i(\boldsymbol{x})^2.$$

Recíprocamente, si $p(\boldsymbol{x})$ es SOS, $p(\boldsymbol{x}) = \sum q_i(\boldsymbol{x})^2$, construimos la matriz $\boldsymbol{X}$ tomando en la fila i los coeficientes de $q_i(\boldsymbol{x})$ y tomamos $\boldsymbol{Q} = \boldsymbol{X}^T \boldsymbol{X} \succeq 0$.

□

Representación núcleo y representación imagen

Vemos que podemos plantear el problema mediante ecuaciones sobre los coeficientes de $\boldsymbol{Q}$, que se obtienen igualando coeficiente a coeficiente la expresión

$$p(\boldsymbol{x}) = \boldsymbol{v}(\boldsymbol{x})^T \boldsymbol{Q} \boldsymbol{v}(\boldsymbol{x}).$$

Llamamos representación núcleo o implícita a esta representación.

Resolviendo las ecuaciones, podemos plantear el problema mediante una desigualdad lineal matricial (LMI)

$$\boldsymbol{Q} = \boldsymbol{A}_0 + \sum_i y_i \boldsymbol{A}_i$$

que llamamos representación imagen o explícita.

Pregunta: ¿para el ejemplo anterior cuál es la representación explícita y cuál es la representación implícita?

4.6. Programas Sumas de Cuadrados

4.6.1. Motivación

Vimos la siguiente aplicación de polinomios no-negativos:

Dado un polinomio p, si queremos hallar el mínimo de p en $S = \mathbb{R}^n$, planteamos

$$\text{mín}\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^n\} = \sup\{\gamma \in \mathbb{R} | f(\boldsymbol{x}) - \gamma \geq 0 \ \forall \boldsymbol{x} \in \mathbb{R}^n\}.$$

Para $\gamma \in \mathbb{R}$ dado, podemos reemplazar la condición $f(\boldsymbol{x}) - \gamma \geq 0$ por $f(\boldsymbol{x})$ es SOS, lo que nos da un problema de factibilidad de un SDP.

Resolviendo estos problemas para distintos valores de γ, podemos obtener cotas para el mínimo de p.

Pregunta: ¿podemos obtener la mejor cota resolviendo un solo problema SDP?

Vimos cómo verificar si un polinomio dado es una suma de cuadrados.

Podemos extender los resultados para definir una clase de problemas de optimización convexa que llamamos *programas sumas de cuadrados* (SOS).

Definición 4.15. *Un problema de optimización por sumas de cuadrados o programa SOS es un problema de optimización convexa de la forma*

$$\begin{array}{ll} \textit{maximizar:} & b_1 y_1 + \cdots + b_m y_m \\ \textit{sujeto a:} & p_i(\boldsymbol{x}; \boldsymbol{y}) \textit{ es SOS }, 1 \leq i \leq k, \end{array}$$

donde $\boldsymbol{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m$ *es la variable de optimización,* $b_i \in \mathbb{R}$, $1 \leq i \leq m$, $p_i(\boldsymbol{x}; \boldsymbol{y}) = a_{i0}(\boldsymbol{x}) + a_{i1}(\boldsymbol{x})y_1 + \cdots + a_{im}(\boldsymbol{x})y_m$, $1 \leq i \leq k$ *y* $a_{ij}(\boldsymbol{x}) \in \mathbb{R}[\boldsymbol{x}]$ *son polinomios dados.*

Observaciones:

- Los polinomios $p_i(\boldsymbol{x}, \boldsymbol{y})$ son polinomios arbitrarios que son combinaciones afines en los parámetros $y_1, \ldots, y_m$.
- Las variables $\boldsymbol{x}$ son variables "dummy", no optimizamos sobre ellas sino que son las indeterminadas de los polinomios p_i.

Ejemplo. Consideramos el problema

$$\begin{array}{ll} \text{maximizar:} & y_1 + y_2 \\ \text{sujeto a:} & x^4 + y_1 x + (2 + y_2) \quad \text{es SOS,} \\ & (y_1 - y_2 + 1)x^2 + y_2 x + 1 \quad \text{es SOS.} \end{array}$$

Aunque a primera vista, los programas SOS parecen más generales que los problemas SDP, cada restricción del problema podemos plantearla como la existencia de una matriz $\boldsymbol{Q}_i \succeq 0$ tal que

$$p_i(\boldsymbol{x}; \boldsymbol{y}) = \boldsymbol{v}^T \boldsymbol{Q}_i \boldsymbol{v},$$

donde las coordenadas de $\boldsymbol{Q}_i$ dependen linealmente de las variables y_i, $1 \leq i \leq m$, y por lo tanto un programa SOS es un problema SDP.

Ejemplo 4.16. Para la restricción

$$p_1(x; y_1, y_2) = x^4 + y_1 x + (2 + y_2) \quad \text{es sos}$$

planteamos

$$x^4 + y_1 x + (2 + y_2) = \begin{pmatrix} 1 & x & x^2 \end{pmatrix} \begin{pmatrix} q_{00} & q_{10} & q_{20} \\ q_{10} & q_{11} & q_{21} \\ q_{20} & q_{21} & q_{22} \end{pmatrix} \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix}$$

y obtenemos la restricción $\boldsymbol{Q} = \begin{pmatrix} 2 + y_2 & \frac{y_1}{2} & -\frac{a}{2} \\ \frac{y_1}{2} & a & 0 \\ -\frac{a}{2} & 0 & 1 \end{pmatrix} \succeq 0$, donde a, y_1, y_2 son las variables del problema SDP.

4.6.2. Aplicación: optimización polinomial sin restricciones

Caso polinomios univariados

Para encontrar el mínimo de un polinomio en una variable, utilizamos la equivalencia $p(x) \geq \gamma \ \forall x \in \mathbb{R} \iff p(x) - \gamma \geq 0 \ \forall x \in \mathbb{R}$.

Obtenemos el siguiente problema de optimización:

$$\begin{array}{ll} \text{maximizar:} & \gamma \\ \text{sujeto a:} & p(x) - \gamma \geq 0 \ \forall x \in \mathbb{R}. \end{array}$$

Como en una variable un polinomio es no-negativo si y solo si es SOS, obtenemos el problema SOS equivalente

$$\begin{array}{ll} \text{maximizar:} & \gamma \\ \text{sujeto a:} & p(x) - \gamma \quad \text{es SOS.} \end{array}$$

Caso polinomios multivariados

Análogamente, si queremos encontrar el mínimo de un polinomio multivariado planteamos el problema de optimización:

$$\begin{array}{ll} \text{maximizar:} & \gamma \\ \text{sujeto a:} & p(\boldsymbol{x}) - \gamma \geq 0 \ \forall \boldsymbol{x} \in \mathbb{R}^n. \end{array}$$

En el caso general este problema no se puede plantear eficientemente, pero podemos plantear el problema alternativo

$$\begin{array}{ll} \text{maximizar:} & \gamma \\ \text{sujeto a:} & p(\boldsymbol{x}) - \gamma \quad \text{es SOS.} \end{array}$$

Llamamos $p_\star$ al ínfimo de p (que coincide con el óptimo del primer problema) y p_{SOS} al óptimo del segundo problema.

Como el conjunto factible del problema SOS está incluido en el conjunto factible del primero, obtenemos la desigualdad

$$p_{SOS} \leq p_\star.$$

Si bien en muchos casos (especialmente en dimensión baja) ambos óptimos coinciden, el primer problema es NP-hard y por lo tanto no podemos esperar que los óptimos coincidan siempre.

Capítulo 5

Momentos

Referencia principal: [Lasserre, 2010, Capítulo 3].

5.1. El problema de los momentos en una variable

Ya vimos los siguientes resultados para polinomios en una variable:

(a) Si $f \in \mathbb{R}[x]$ es positivo sobre todo $\mathbb{R}$, entonces f es una suma de cuadrados, $f = q_1^2 + \cdots + q_s^2$.

(b) Si $f \in \mathbb{R}[x]$ es positivo en $[0, +\infty)$, entonces $f = p_0 + xp_1$, para dos polinomios SOS $p_0, p_1 \in \Sigma[x]$.

Veamos cómo se traducen estos dos resultados al problema de momentos. Nos interesa calcular la integral de un polinomio sobre un subconjunto $K \subset \mathbb{R}$.

$$\int_K f(x)dx.$$

Observamos que si $f(x) = \sum_{i=0}^n f_i x^i$, entonces su integral es

$$\int_K f(x)dx = \int_K \sum_{i=0}^n f_i x^i dx = \sum_{i=0}^n f_i \int_K x^i dx.$$

Es decir, que conociendo $\int_K x^i dx$ para todo $i \in \mathbb{N}_0$ podemos calcular fácilmente la integral de cualquier polinomio sobre K. Siguiendo esta idea, consideramos una secuencia infinita $y = (y_i)_{i \in \mathbb{N}_0} \subset \mathbb{R}$ y definimos la funcional lineal $L_y : \mathbb{R}[x] \to \mathbb{R}$,

$$f(x) = \sum_{i \in \mathbb{N}_0} f_i x^i \mapsto L_y(f) = \sum_{i \in \mathbb{N}_0} f_i y_i.$$

Si tomamos $y_i = \int_K x^i dx$, para un conjunto $K \subset \mathbb{R}$, entonces

$$L_y(f) = \sum_{i \in \mathbb{N}_0} f_i \int_K x^i dx = \int_K \sum_{i \in \mathbb{N}_0} f_i x^i dx = \int_K f dx.$$

El problema de momentos consiste en determinar para qué secuencias $y = (y_i)_{i\in\mathbb{N}_0}$ existe un conjunto K tal que $y_i = \int_K x^i dx$, o más generalmente, una medida μ tal que

$$y_i = \int_K x^i d\mu.$$

Comenzamos por el teorema principal, que veremos sin demostración.

Teorema 5.1 (Riesz-Haviland)**.** *Sea $y = (y_i)_{i\in\mathbb{N}_0} \subset \mathbb{R}$ y sea $K \subset \mathbb{R}$ un conjunto cerrado. Existe una medida de Borel finita μ en K tal que*

$$\int_K x^i d\mu = y_i, \quad \forall i \in \mathbb{N}_0,$$

si y solo si $L_y(f) \geq 0$ para todos los polinomios $f \in \mathbb{R}[x]$ no-negativos en K.

En las condiciones del teorema, decimos que la medida de Borel μ representa a y en K.

Utilizando este teorema, podemos dar las caracterizaciones simples que buscamos.

Dada una sucesión $y = (y_i) \subset \mathbb{R}$, definimos las matrices de Hankel $H_n(y)$ y $B_n(y) \in \mathbb{R}^{(n+1)\times(n+1)}$ por

$$H_n(y)(i,j) := y_{i+j-2},$$
$$B_n(y)(i,j) := y_{i+j-1},$$

para todo $i, j \in \mathbb{N}$, $1 \leq i, j \leq n+1$.

Ejemplo 5.2. Para $y = (1, 2, 3, 4, 5, 6, 7, 8, 0, 0, 0, \dots)$,

$$H_3 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{pmatrix} \quad y \quad B_3 = \begin{pmatrix} 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \\ 5 & 6 & 7 & 8 \end{pmatrix}.$$

Teorema 5.3. *Sea $y = (y_j)_{j\in\mathbb{N}_0} \subset \mathbb{R}$. Entonces,*

(a) existe una medida de Borel μ que representa a y en $\mathbb{R}$ si y solo si la forma cuadrática

$$x \mapsto s_n(x) := \sum_{i,j=0}^{n} y_{i+j} x_i x_j \tag{5.1}$$

es positiva semidefinida para todo $n \in \mathbb{N}$. Equivalentemente, $H_n(y) \succeq 0$ para todo $n \in \mathbb{N}$.

(b) existe una medida de Borel μ que representa a y en $\mathbb{R}_+$ si y solo si las formas cuadráticas 5.1 y

$$x \mapsto u_n(x) := \sum_{i,j=0}^{n} y_{i+j+1} x_i x_j \tag{5.2}$$

son ambas positivas semidefinidas para todo $n \in \mathbb{N}$. Equivalentemente, $H_n(y) \succeq 0$ y $B_n(y) \succeq 0$ para todo $n \in \mathbb{N}$.

Demostración. (a) Si $y_n = \int_{\mathbb{R}} z^n d\mu(z)$, entonces

$$s_n(x) = \sum_{i,j=0}^{n} x_i x_j \int_{\mathbb{R}} z^{i+j} d\mu(z) = \int_{\mathbb{R}} (\sum_{i=0}^{n} x_i z^i)^2 d\mu(z) \geq 0.$$

Recíprocamente, si $H_n(y) \succeq 0$ para todo $n \in \mathbb{N}$, para todo $q \in \mathbb{R}^{n+1}$ tenemos

$$q^t H_n(y) q \geq 0.$$

Dado $p \in \mathbb{R}[x]$ un polinomio no-negativo en $\mathbb{R}$, p se puede escribir como una suma de cuadrados $p = \sum_{j=1}^{r} q_j^2$. Luego,

$$\sum_{k=0}^{2n} p_k y_k = L_y(p) = L_y(\sum_{j=1}^{r} q_j^2) = \sum_{j=1}^{r} q_j^t H_n(y) q_j \geq 0,$$

donde q_j es el vector de coeficientes de $q_j \in \mathbb{R}[x]$. Como $p \geq 0$ es arbitrario, obtenemos por el Teorema 5.1 que $y_i = \int_{\mathbb{R}} x^i d\mu$ para alguna medida μ en $\mathbb{R}$.

(b) La demostración es similar al ítem anterior, utilizando que si p es no-negativo en $\mathbb{R}_+$, entonces p es de la forma

$$p(x) = p_0(x) + x p_1(x),$$

con $p_0, p_1 \in \Sigma[x]$. □

5.2. El problema de los momentos en varias variables

5.3. Algoritmos. Relajación semidefinida

Bibliografía

[Cen,] Linear programming.

[Bertsimas and Tsitsiklis, 1997] Bertsimas, D. and Tsitsiklis, J. (1997). *Introduction to Linear Optimization*. Athena Scientific.

[Blekherman et al., 2013] Blekherman, G., Parrilo, P., and Thomas, R., editors (2013). *Semidefinite Optimization and Convex Algebraic Geometry*. SIAM.

[Fawzi, 2018] Fawzi, H. (2018). Topics in convex optimisation (lecture notes).

[Horn and Johnson, 1985] Horn, R. and Johnson, C. (1985). *Matrix Analysis*. Cambridge University Press.

[Lasserre, 2010] Lasserre, J. (2010). *Moments, Positive Polynomials and Their Applications*. Imperial College Press.

[O'Donell, 2008] O'Donell, R. (2008). Lecture notes in advanced approximation algorithms (cmu 18-854b).

[Todd, 2001] Todd, M. (2001). Semidefinite optimization. *Acta Numerica*, 10:515–560.