Notas de Métodos Numéricos

Guido Tagliavini Ponce
Universidad de Buenos Aires
guido.tag@gmail.com

v1.2 (16/05/2015)

Índice

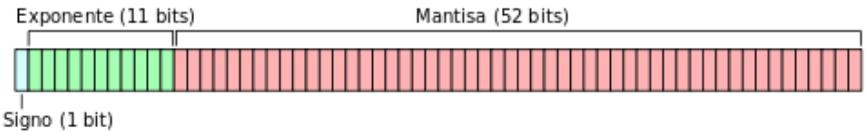
1. Aritmética de la computadora

Dado que en una computadora todos los números son representados mediante una cantidad de dígitos finita y fija, valores como π o $\sqrt{2}$ no se pueden manipular con completa exactitud, pues al ser irracionales tienen infinitos decimales no periódicos. Los irracionales no son los únicos números no representables correctamente en una computadora: aquellos racionales con una cola decimal no periódica suficientemente grande tampoco lo serán. En una computadora sólo se pueden representar con precisión un subconjunto de los números racionales. Esto hace que al hacer cómputos con números reales se genere un error numérico.

1.1. Representación estándar IEEE

El estándar fijado por la IEEE contempla varias representaciones que se distinguen por su precisión. Las dos más frecuentemente utilizadas son *single* (32 bits) y *double* (64 bits). Las dos restantes son *half* (16 bits) y *quadruple* (128 bits). Todas estas representaciones son binarias y de punto flotante.

La precisión double tiene la siguiente estructura:



- **Signo** (s) - 1 bit. El número representado es positivo si $s = 0$ y negativo si no.
- **Exponente** (e) - 11 bits. La base del exponente es 2. Para poder representar números con valor absoluto menor a 1 es necesario admitir exponentes negativos. Por esto, el exponente está representado en exceso a $2^{10} - 1$, es decir, el exponente del número representado será $(e)_2 - (2^{10} - 1)$. Como e tiene 11 bits entonces $0 \leq (e)_2 \leq 2^{11} - 1 \Rightarrow -(2^{10} - 1) \leq (e)_2 - (2^{10} - 1) \leq 2^{11} - 1 - (2^{10} - 1) \Rightarrow -2^{10} + 1 \leq (e)_2 - (2^{10} - 1) \leq 2^{10}$.
- **Mantisa** (m) - 52 bits. Se considera una mantisa normalizada, i. e., el número representado tiene mantisa $(1, m)_2$.

En definitiva, el número representado es

$$(-1)^s \cdot (1, m)_2 \cdot 2^{(e)_2 - (2^{10}-1)}$$

Definición 1.1. Decimos que un cómputo genera *underflow* si su resultado es menor que el mínimo positivo representable, en módulo. Análogamente, decimos que genera *overflow* si su resultado es mayor que el máximo positivo representable, en módulo.

1.2. Aproximación de los reales mediante números de máquina

En esta sección estudiaremos cuán eficaz es la aproximación de un número real mediante un sistema con las características del presentado previamente. El tipo de sistemas a los que nos referimos son representaciones de punto flotante con una longitud de mantisa fija y exponente acotado.

Supongamos que nuestro conjunto de números de máquina es

$$\mathcal{M} = \{x \in \mathbb{R} : x = \pm(0, d_1 d_2 \cdots d_k) \cdot 10^e, 0 \leq d_i \leq 9, d_1 \neq 0, e_1 \leq e \leq e_2\}$$

dadas ciertas constantes k, e_1 y e_2. Esta es una representación normalizada con mantisa de k dígitos y exponente entre e_1 y e_2. Es un sistema decimal y no binario como el de la IEEE, porque así será más fácil razonar sobre él. Todos los resultados que veremos en esta sección aplican, mutatis mutandi, a un sistema similar que utilice una base $b > 1$ cualquiera.

También por simplicidad, asumiremos que $e_1 = -\infty$ y $e_2 = +\infty$. Esto ahorra hablar del rango en el que corre el exponente, lo cual, como veremos, no hace a la esencia de los resultados.

Definición 1.2. Si $x \in \mathbb{R}$ es cualquiera, no necesariamente de máquina, llamamos $fl(x)$ a la aproximación de x vía números de máquina.

Es definición depende del modo en que realicemos la aproximación. Consideremos la escritura

$$x = (0, d_1 \cdots d_k d_{k+1} \cdots) \cdot 10^e$$

con $d_1 \neq 0$ (esta escritura es única dado que $d_1 \neq 0$). Entonces dos formas de aproximar x son:

- **Truncamiento.** Simplemente descartamos los dígitos $d_{k+1}, d_{k+2}, \cdots$, para obtener

$$fl(x) = (0, d_1 \cdots d_k) \cdot 10^e$$

- **Redondeo.** Si $d_{k+1} < 5$ entonces truncamos. Si no, sumamos $0, \underbrace{0 \cdots 05}_{k+1 \text{ dígitos}} \cdot 10^e = 5 \cdot 10^{-(k+1)} \cdot 10^e$ a x y truncamos.

  En este último caso lo que queda es

$$fl(x) = [(0, d_1 \cdots d_k) + 10^{-k}] \cdot 10^e$$

Una forma equivalente de enunciar estos criterios es la siguiente. Sea x^- es el máximo número de máquina menor o igual que x. Análogamente, sea x^+ el mínimo número de máquina mayor o igual que x. Entonces, el truncamiento aproxima por x^- mientras que el redondeo aproxima por aquel valor de x^- o x^+ más cercano a x.

En general, las computadoras utilizan la aproximación por redondeo.

1.3. Distribución de los números de máquina sobre la recta real

Observación 1.1. Un número de máquina con exponente e cae en el intervalo $[10^{e-1}, 10^e)$.

Lema 1.1. *Dado $x \in \mathcal{M}$ con exponente e, el número de máquina que lo sucede es $x' = x + 10^{-k} \cdot 10^e$.*

Demostración. Por la observación anterior, x cae en el intervalo $[10^{e-1}, 10^e)$. Si x es el máximo número de máquina en dicho intervalo, entonces x' es el número de máquina más pequeño en $[10^e, 10^{e+1})$, i. e., $x' = (0, 1) \cdot 10^{e+1}$. En caso contrario, el sucesor x' también cae en $[10^{e-1}, 10^e)$, con lo cual tiene exponente e. Lo mínimo que puede incrementarse la mantisa de x es $0, \underbrace{0 \cdots 01}_{k \text{ dígitos}}$, con lo cual $x' = x + (0, \underbrace{0 \cdots 01}_{k \text{ dígitos}}) \cdot 10^e = x + 10^{-k} \cdot 10^e$. □

La distribución de $\mathcal{M}$ no es uniforme sobre la recta real. Para ver por qué, contemos la cantidad de números de máquina que hay en el intervalo $[10^i, 10^{i+1})$ con $i \in \mathbb{Z}$ una constante entera.

Proposición 1.1. *La cantidad de números de máquina $x \in \mathcal{M}$ tal que $x \in [10^i, 10^{i+1})$ es $9 \cdot 10^{k-1}$.*

Demostración. El menor número de máquina en $[10^i, 10^{i+1})$ es $(0, 1) \cdot 10^{i+1} = 10^i$, y el mayor es $(0, 99 \cdots 99) \cdot 10^{i+1}$. Entre estos dos están

$$\begin{gathered}(0, 10 \cdots 00) \cdot 10^{i+1} \\ (0, 10 \cdots 01) \cdot 10^{i+1} \\ \vdots \\ (0, 99 \cdots 98) \cdot 10^{i+1} \\ (0, 99 \cdots 99) \cdot 10^{i+1}\end{gathered}$$

Observemos que estos son todos los números de máquina positivos con exponente $i + 1$. Afirmamos que no hay otros números de máquina en $[10^i, 10^{i+1})$. Si un número en $\mathcal{M}$ positivo tiene exponente menor o igual que i, como la mantisa es estrictamente menor que 1, entonces el número será estrictamente menor que 10^i. En caso contrario, si tiene exponente mayor o igual que $i + 2$, como el primer dígito decimal de la mantisa es no nulo, el número será al menos 10^{i+1}.

Para ver cuántos números hay en esta lista, notemos que hay tantos como números enteros entre 10^{k-1} y $10^k - 1$, y estos son $9 \cdot 10^{k-1}$. □

Corolario 1.1. *Los números de máquina no están uniformemente distribuidos.*

Demostración. Como $\#(\mathcal{M} \cap [10^i, 10^{i+1})) = 9 \cdot 10^{k-1}$ no depende de i, resulta que en los intervalos $[1, 10)$, $[10, 100)$, $[100, 1000)$, $\cdots$, $[10^i, 10^{i+1})$, $\cdots$, hay igual cantidad de números de máquina. □

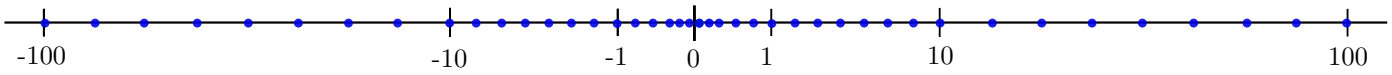
Intuitivamente, cuanto más nos alejemos de 0, más esparcidos estarán los números de máquina. Lo próximo que queremos es medir cuán esparcidos están, según el intervalo $[10^i, 10^{i+1})$ en el que caigan.

Lema 1.2. *Sea $x \in \mathcal{M} \cap [10^i, 10^{i+1})$ y sea x' su sucesor. Entonces $x' - x = 10^{i+1-k}$.*

Demostración. Como el exponente de x es $i + 1$ entonces $x' = x + 10^{-k} \cdot 10^{i+1}$, y concluimos $x' - x = 10^{i+1-k}$. □

Corolario 1.2. *En cada intervalo $[10^i, 10^{i+1})$ los números de máquina consecutivos estan equiespaciados.*

Más aún, a medida que i crece, la brecha entre elementos consecutivos en $[10^i, 10^{i+1})$ se hace mayor. De aquí se deduce que la distribución de los números de máquina tiene el siguiente aspecto



Esta distribución puede parecer extraña. Contrariamente a lo intuitivo, que haría pensar que una distribución uniforme sería más útil, esta distribución exponencial de los números de máquina resulta práctica pues se basa en la idea de que cuanto más chicos sean los números del rango en el que estamos trabajando, más pequeñas serán las variaciones que estaremos interesados en hacer. Contrariamente, cuanto más grandes sean los números con los que trabajemos, más grandes serán las variaciones con las que vayamos a trabajar. Esta noción se formaliza en la siguiente sección.

1.4. Error absoluto y error relativo

Definición 1.3. Sea $x \in \mathbb{R}$. Sea $x^* \in \mathbb{R}$ un valor que pretende aproximar a x.

- El error absoluto de aproximar x por x^* es $|x - x^*|$.
- El error relativo de aproximar x por x^* es $\frac{|x-x^*|}{|x|}$.

Notemos que la diferencia entre estas dos medidas de error, es que el absoluto no contempla el tamaño del valor que estamos aproximando mientras que el relativo sí lo hace. Esto determina que, en general, el error absoluto no sea una buena medida para el error, puesto que en ciertos contextos podemos tener un error absoluto aparentemente grande aunque la aproximación sea buena (por ejemplo, si aproximáramos la distancia entre la Tierra y la Luna con un error absoluto de 1000m) y en otros contextos podemos tener un error absoluto aparentemente chico aunque la aproximación sea mala (por ejemplo, si aproximáramos el perímetro de una célula vegetal con un error absoluto de 1mm).

Supongamos que queremos aproximar una magnitud que vale $x = 1000$, y nos piden que el error relativo sea a lo sumo $0,01$. Si nuestra medición es $x^* = 990$ entonces el error relativo será

$$\frac{|x - x^*|}{|x|} = \frac{10}{1000} = 0,01$$

con lo cual esta medición satisface lo pedido. Es fácil ver que cualquiera sea $x^* \in [990, 1010]$ cumple.

Supongamos que ahora estamos buscando medir una magnitud $x = 10$. La precisión buscada es la misma que antes, lo cual tiene sentido pues está expresada en términos relativos. Notemos que la medición $x^* = 9$ no será aceptada, porque el error relativo será de 0,1. Refinando la medición a $x^* = 9{,}9$ conseguimos un error relativo de 0,01 y es claro que cualquier valor $x^* \in [9{,}9; 10{,}1]$ no sobrepasa este error.

En ambas situaciones los errores relativos eran los mismos, sin embargo la diferencia en valor absoluto entre x y los valores de x^* aceptados era mucho mayor en el primer caso que en el segundo caso. La discusión del final de la sección anterior cobra ahora mucho más sentido.

A continuación estudiamos el error relativo al aproximar un número real por un número de máquina.

Proposición 1.2. *Si $fl(x)$ se obtiene truncando x entonces*

$$\frac{|x - fl(x)|}{|x|} \leq 10^{1-k}$$

Demostración. Supongamos sin pérdida de generalidad que $x > 0$. Sea $x = (0, d_1 \cdots d_k d_{k+1} \cdots) \cdot 10^e$. Entonces

$$\begin{aligned}\frac{|x - fl(x)|}{|x|} &= \frac{|(0, d_1 \cdots d_k d_{k+1} \cdots) \cdot 10^e - (0, d_1 \cdots d_k) \cdot 10^e|}{|(0, d_1 \cdots d_k d_{k+1} \cdots) \cdot 10^e|} \\ &= \frac{(0, 0 \cdots 0 d_{k+1} d_{k+2} \cdots)}{(0, d_1 \cdots d_k d_{k+1} \cdots)}\end{aligned}$$

Como $(0, d_1 \cdots d_k d_{k+1} \cdots) \geq 0,1 = 10^{-1}$ y $(0, 0 \cdots 0 d_{k+1} d_{k+2} \cdots) \leq 0, \underbrace{0 \cdots 01}_{k \text{ dígitos}} = 10^{-k}$, entonces

$$\frac{|x - fl(x)|}{|x|} \leq 10 \cdot 10^{-k} = 10^{1-k}$$

□

Proposición 1.3. *Si $fl(x)$ se obtiene redondeando x entonces*

$$\frac{|x - fl(x)|}{|x|} \leq \frac{1}{2} \cdot 10^{1-k}$$

Demostración. Separamos en casos, según el valor de d_{k+1}.

Si $d_{k+1} < 5$ entonces $fl(x) = (0, d_1 \cdots d_k) \cdot 10^e$. Entonces, haciendo la misma cuenta que antes,

$$\frac{|x - fl(x)|}{|x|} = \frac{(0, 0 \cdots 0 d_{k+1} d_{k+2} \cdots)}{(0, d_1 \cdots d_k d_{k+1} \cdots)}$$

Como $d_{k+1} < 5$ entonces $(0, 0 \cdots 0 d_{k+1} d_{k+2} \cdots) \leq 5 \cdot 10^{-(k+1)}$. Luego

$$\frac{|x - fl(x)|}{|x|} \leq 5 \cdot 10^{-(k+1)} \cdot 10 = \frac{10}{2} \cdot 10^{-k} = \frac{1}{2} \cdot 10^{1-k}$$

Veamos el caso $d_{k+1} \geq 5$. Ahora tenemos $fl(x) = [(0, d_1 \cdots d_k) + 10^{-k}] \cdot 10^e$. Entonces

$$\begin{aligned}\frac{|x - fl(x)|}{|x|} &= \frac{|(0, d_1 \cdots d_k d_{k+1} \cdots) - [(0, d_1 \cdots d_k) + 10^{-k}]|}{(0, d_1 \cdots d_k d_{k+1} \cdots)} \\ &= \frac{|(0, 0 \cdots 0 d_{k+1} d_{k+2} \cdots) - 10^{-k}|}{(0, d_1 \cdots d_k d_{k+1} \cdots)} \\ &= \frac{10^{-k} - (0, 0 \cdots 0 d_{k+1} d_{k+2} \cdots)}{(0, d_1 \cdots d_k d_{k+1} \cdots)} \\ &= \frac{(0, 0 \cdots 0 (9 - d_{k+1})(9 - d_{k+2}) \cdots)}{(0, d_1 \cdots d_k d_{k+1} \cdots)}\end{aligned}$$

Como $d_{k+1} \geq 5$ entonces $(0, 0 \cdots 0 (9 - d_{k+1})(9 - d_{k+2}) \cdots) \leq (0, \underbrace{0 \cdots 0}_{k \text{ dígitos}} 499 \cdots) = 5 \cdot 10^{-(k+1)}$. Luego

$$\frac{|x - fl(x)|}{|x|} \leq 5 \cdot 10^{-(k+1)} \cdot 10 = \frac{1}{2} \cdot 10^{1-k}$$

□

Corolario 1.3. *En una máquina con representación de punto flotante decimal con mantisa de k dígitos, el error de redondeo es* $\mathcal{O}(10^{-k})$.

El máximo error relativo que se puede cometer, según lo calculado en la Proposición 1.3, tiene un nombre particular.

1.5. Epsilon de máquina

Definición 1.4. Llamamos epsilon de máquina al máximo error relativo que puede cometerse por redondeo. Lo notamos ε.

En el caso de la máquina $\mathcal{M}$ con la que venimos trabajando, el epsilon de máquina es $\varepsilon = \frac{1}{2} \cdot 10^{1-k}$.

Proposición 1.4. *ε es el mínimo real positivo x tal que $fl(1 + x) \neq 1$.*

Demostración. La observación clave es que el número representable que sigue al 1 es $1 + 10^{-k} \cdot 10 = 1 + 10^{1-k}$ (Lema 1.1). Entonces, el punto medio entre 1 y su sucesor es $1 + \frac{1}{2} \cdot 10^{1-k}$. De aquí se deduce que si $0 \leq x < \frac{1}{2} \cdot 10^{1-k}$ entonces al redondear $1 + x$ obtenemos $fl(1 + x) = 1$, y si $x = \frac{1}{2} \cdot 10^{1-k}$ y redondeamos $1 + x$ obtenemos el sucesor. Entonces $x = \frac{1}{2} \cdot 10^{1-k} = \varepsilon$ es el mínimo que cumple lo buscado. □

Observación 1.2. La validez de este resultado proviene de la fuerte relación que hay entre la cota superior para el error relativo y la distancia entre 1 y el número de máquina que lo sucede.

Observación 1.3. El epsilon de máquina no tiene ninguna relación con el mínimo real positivo representable. De hecho, el mínimo representable depende del rango en el que se mueve el exponente, y no de la precisión de la mantisa.

La noción de ε permite dar cotas superiores sobre el error cometido al realizar distintas operaciones en una máquina, independientemente de las características del sistema de representación de números que utilice la misma. Las cuatro operaciones estándar que realiza una computadora son,

$$\begin{aligned} x \oplus y &= fl(fl(x) + fl(y)) \\ x \ominus y &= fl(fl(x) - fl(y)) \\ x \otimes y &= fl(fl(x) \times fl(y)) \\ x \oslash y &= fl(fl(x)/fl(y)) \end{aligned}$$

que representan, respectivamente, la suma, la resta, el producto y el cociente de dos números reales x e y.

Proposición 1.5. *Sean $x, y \in \mathbb{R}$ no nulos, con igual signo. En cualquier máquina con suma $\oplus$, que utilice redondeo, vale*

$$\frac{|(x + y) - (x \oplus y)|}{|x + y|} \leq 2\varepsilon + \varepsilon^2$$

Demostración. En efecto,

$$
\begin{aligned}
\frac{|(x+y)-(x\oplus y)|}{|x+y|} &= \frac{|(x+y)-(x\oplus y)|}{|x+y|} \\
&= \frac{|(x+y)-fl(fl(x)+fl(y))|}{|x+y|} \\
&= \frac{|x+y-(fl(x)+fl(y))+(fl(x)+fl(y))-fl(fl(x)+fl(y))|}{|x+y|} \\
&= \frac{\left|x\frac{x-fl(x)}{x}+y\frac{y-fl(y)}{y}+(fl(x)+fl(y))\frac{fl(x)+fl(y)-fl(fl(x)+fl(y))}{fl(x)+fl(y)}\right|}{|x+y|} \\
&\leq \frac{1}{|x+y|}\left(|x|\frac{|x-fl(x)|}{|x|}+|y|\frac{|y-fl(y)|}{|y|}+|fl(x)+fl(y)|\frac{|fl(x)+fl(y)-fl(fl(x)+fl(y))|}{|fl(x)+fl(y)|}\right) \\
&\leq \frac{1}{|x+y|}\left(|x|\varepsilon+|y|\varepsilon+|fl(x)+fl(y)|\varepsilon\right) \\
&= \frac{1}{|x+y|}\left(|x|+|y|+|fl(x)+fl(y)|\right)\varepsilon \\
&= \frac{1}{|x+y|}\left(|x|+|y|+\left|x\frac{fl(x)-x}{x}+y\frac{fl(y)-y}{y}+x+y\right|\right)\varepsilon \\
&\leq \frac{1}{|x+y|}\left(|x|+|y|+|x|\frac{|x-fl(x)|}{|x|}+|y|\frac{|y-fl(y)|}{|y|}+|x|+|y|\right)\varepsilon \\
&\leq \frac{1}{|x+y|}\left(2|x|+2|y|+|x|\varepsilon+|y|\varepsilon\right)\varepsilon \\
&= \frac{1}{|x+y|}(|x|+|y|)(2+\varepsilon)\varepsilon \\
&= \frac{|x|+|y|}{|x+y|}(2\varepsilon+\varepsilon^2)
\end{aligned}
$$

Como x e y tienen igual signo, $\frac{|x|+|y|}{|x+y|}=1$, lo cual termina la demostración. □

1.6. Errores de redondeo clásicos

1.6.1. Suma de números de órdenes muy distintos

Supongamos que tenemos un número real de valor absoluto pequeño, y otro de valor absoluto grande. Entonces la suma de máquina de los dos puede hacer que el más pequeño desaparezca.

Para ejemplificar, supongamos que nuestra aritmética tiene una precisión de $k=5$ dígitos de mantisa. Sean $x = 0,88888888\cdot 10^7$ e $y = 0,1\cdot 10^2$. Entonces

$$
\begin{aligned}
x\oplus y &= fl(fl(x)+fl(y)) \\
&= fl(0,88888\cdot 10^7 + 0,1\cdot 10^2) \\
&= fl(0,888881\cdot 10^7) \\
&= 0,88888\cdot 10^7
\end{aligned}
$$

El término x ha absorbido a y.

1.6.2. Resta de números cercanos

Al restar dos números cercanos, el resultado estará próximo a cero, lo que puede ocasionar que se pierdan dígitos significativos. Este fenómeno se conoce como *cancelación catastrófica*, y tiene un gran impacto en el error relativo.

A modo de ejemplo, supongamos nuevamente que la precisión es de $k=5$ dígitos, y sean que $x = 0,12346923$ e $y = 0,12345175$ entonces

$$\begin{aligned} x \ominus y &= fl(fl(x) - fl(y)) \\ &= fl(0,12347 - 0,12345) \\ &= fl(0,00002) \\ &= 0,00002 \end{aligned}$$

Sin embargo $x - y = 0,00001748$, lo que muestra que hemos perdido 3 dígitos significativos del resultado en la operación con redondeo.

1.6.3. Multiplicación por números grandes o dividisión por números pequeños

En este caso, se produce una amplificación del error absoluto acarreado. Supongamos que x^* es una aproximación de máquina de x. Dividiendo a x^* por un número muy pequeño, digamos 10^{-n} para cierto $n > 0$, obtenemos el número de máquina $x^*/10^{-n}$ que aproxima a $x/10^{-n}$ con un error absoluto de

$$|x^*/10^{-n} - x/10^{-n}| = |x^* - x| \cdot 10^n$$

El error absoluto $|x^* - x|$ del primer redondeo se ve amplificado en un factor de 10^n.

1.7. Generalización

Como hemos dicho al principio de esta sección, hemos estudiado la representación de los números de máquina utilizando un sistema de base 10 exclusivamente por comodidad. Todos los resultados que vimos son fácilmente generalizables a una base arbitraria $b > 1$. A modo de ejemplo, en un conjunto análogo a $\mathcal{M}$ pero de base 2, el epsilon de máquina resulta ser $\varepsilon = \frac{1}{2} \cdot 2^{1-k} = 2^{-k}$.

Otro ejemplo interesante es la representación IEEE de precisión doble, cuyo conjunto de números de máquina es similar a $\mathcal{M}$ pero es de base 2 y tiene $k = 52 + 1$ dígitos de precisión (52 dígitos de mantisa más 1 dígito de precisión que agrega la normalización de la misma). Entonces, el epsilon de máquina en una computadora que use representación IEEE de precisión doble será $\varepsilon = 2^{-53}$. Éste es el máximo error relativo que una computadora típica cometerá por redondeo.

2. Ceros de función

2.1. Problema

Dada $f : \mathbb{R} \to \mathbb{R}$, buscamos un $x^* \in \mathbb{R}$ tal que $f(x^*) = 0$. El número x^* se llama cero o raíz de f.

En ciertos casos, como por ejemplo para $f(x)$ un polinomio de grado menor o igual que 2, el problema tiene una solución que sabemos calcular en forma exacta. En otros casos, por ejemplo para $f(x) = e^x - \ln(x^2)$, no parece tan claro cómo calcular una raíz, si es que existe una.

2.2. Propuesta

Para calcular una raíz, construiremos una sucesión $\{x_n\}_{n\in\mathbb{N}_0}$ de modo tal que $x_n \xrightarrow[n\to\infty]{} x^*$. Más aún, definiremos dicha sucesión en forma recurrente, de modo tal que a partir de x_0 podamos computar x_1, luego x_2 y así sucesivamente. Esto nos da un método iterativo, en el cual a medida que el número de iteraciones n crece, la aproximación x_n de x^* es cada vez mejor, en el sentido de que el valor absoluto $|x_n - x^*|$ es cada vez más pequeño.

Sin saber aún cómo construir $\{x_n\}_n$, surgen algunas preguntas:

1. ¿Cuán rápido convergerá x_n a x^*? En otras palabras, ¿con qué velocidad la diferencia $|x_n - x^*|$ converge a 0?
2. ¿Cuánto es necesario iterar para obtener una buena aproximación? Recordemos que la raíz x^*, a la que queremos converger, no es conocida de antemano, con lo cual, en general, no sabemos cuán cerca está x_n de x^* en un instante dado de la iteración.
3. Dado que no podemos iterar infinitamente, necesitamos establecer criterios para decidir cuándo finalizar la iteración, que no dependan de la distancia entre x_n y x^*.

2.3. Velocidad u orden de convergencia

Como hemos planteado, nos interesa medir la velocidad con la que $\{x_n\}_n$ se aproxima a x^*.

Definición 2.1. Sea $\{x_n\}_n$ una sucesión que converge a x^*, pero $x_n \neq x^*$ para todo n. Decimos que $\{x_n\}_n$ tiene orden de convergencia $p > 0$ si

$$\lim_{n\to\infty} \frac{|x_{n+1} - x^*|}{|x_n - x^*|^p} = c \neq 0$$

para cierta constante $c \in \mathbb{R}$. Si $p = 1$ decimos que la convergencia es lineal. Si $1 < p < 2$ decimos que la convergencia es supralineal. Si $p = 2$ decimos que la convergencia es cuadrática.

Observemos que si el orden de convergencia es p entonces $|x_{n+1} - x^*|$ es asintóticamente equivalente a $|x_n - x^*|^p$. Esto implica que, asintóticamente, el error absoluto se reduce polinomialmente, con exponente p.

Dado que algunas sucesiones convergen con velocidad variable, no siempre es posible aplicar la anterior definición. Hay una medida más general de la velocidad de convergencia, dada por la siguiente

Definición 2.2. Sea $\{\alpha_n\}_n$ convergente a α. Sea $\{\beta_n\}_n$ convergente a 0. Decimos que $\{\alpha_n\}_n$ tiene orden de convergencia $\mathcal{O}(\beta_n)$ (o que α_n converge tan rápido como β_n) si existe una constante $c > 0$ tal que $|\alpha_n - \alpha| \leq c\beta_n$ para todo n suficientemente grande.

En este caso, si $\{\beta_n\}_n$ tiene orden de convergencia p, decimos que $\{\alpha_n\}_n$ tiene orden de convergencia al menos p.

2.3.1. Interpretación

¿Qué significa que $\{x_n\}_n$ converja a x^* con orden p? Llamemos $e_n = x_n - x^*$. Según la definición de antes, esto significa que $\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} = c$ para cierta constante $c \neq 0$. Que c no sea infinito, significa que $|e_n|^p$ no tiende a 0 más rápido de lo que lo hace $|e_{n+1}|$. Que c sea no nulo, significa que $|e_{n+1}|$ tampoco lo hace más rápido que $|e_n|^p$. Por lo tanto, $|e_{n+1}|$ y $|e_n|^p$ convergen a 0 con la misma velocidad o, dicho de otro modo, son asintóticamente equivalentes.

A su vez, es posible interpretar el significado de esta velocidad, en términos prácticos. Dada la equivalencia asintótica de $|e_{n+1}|$ y $|e_n|^p$, vamos a suponer que para n suficientemente grande $|e_{n+1}| \approx |e_n|^p$. Supongamos que hasta el término n-ésimo llevamos calculados k dígitos decimales del valor x^*, es decir que $|e_n| \approx 10^{-k}$. Entonces, $|e_{n+1}| \approx (10^{-k})^p = 10^{-kp}$, es decir que en el $(n+1)$-ésimo término, la cantidad de decimales calculados se multiplica por p.

Entonces, por ejemplo, que una sucesión converja cuadráticamente significa, a nivel práctico, que la cantidad de dígitos decimales calculados se duplica a cada paso.

2.4. Criterios de parada

Respondemos a la tercera pregunta que nos hicimos previamente.

- Fijar un número máximo N de iteraciones.
- Fijar $\varepsilon > 0$ y terminar cuando $|x_n - x_{n-1}| < \varepsilon$.

  Esto es, terminar cuando los saltos de la sucesión sean chicos.

  Este criterio puede fallar. Por ejemplo, tomemos $x_n = \sum_{k=1}^{n} \frac{1}{k}$ la sucesión de números armónicos. Como $x_n - x_{n-1} = \frac{1}{n} \xrightarrow[n\to\infty]{} 0$ entonces para cualquier $\varepsilon > 0$ existirá un valor de n suficientemente grande para el cual dos términos sucesivos disten menos de ε. Sin embargo, es bien sabido que los números armónicos divergen.
- Fijar $\varepsilon > 0$ y terminar cuando $\frac{|x_n - x_{n-1}|}{|x_{n-1}|} < \varepsilon$.

  Esto es, terminar cuando la sucesión no de saltos demasiado grandes en términos relativos.
- Fijar $\varepsilon > 0$ y terminar cuando $|f(x_n)| < \varepsilon$.

  Este criterio se basa en que si x_n converge a una raíz, entonces cuando la sucesión converja, $f(x_n)$ será pequeño.

  También puede fallar. Por ejemplo, en el caso en que $f(x)$ está muy cerca de cero en un punto pero no tiene una raíz en el entorno (e. g. $f(x) = \frac{1}{x}$).
- Fijar $\varepsilon > 0$ y terminar cuando $|f(x_n) - f(x_{n-1})| < \varepsilon$.
- Fijar $\varepsilon > 0$ y terminar cuando $\frac{|f(x_n) - f(x_{n-1})|}{|f(x_{n-1})|} < \varepsilon$.

Dado que todos estos criterios son heurísticos, pueden fallar. La elección del criterio adecuado dependerá del contexto.

A continuación exploraremos algunas formas de encontrar una tal sucesión $\{x_n\}_n$ que converja a una raíz.

2.5. Método de Bisección

Supongamos que tenemos una función $f : [a, b] \to \mathbb{R}$ contínua, a la que le queremos calcular una raíz. Supongamos, además, que $f(a)f(b) < 0$, es decir, $f(a)$ y $f(b)$ tienen signos opuestos. Consideremos el Algoritmo 1, que asume que estamos en estas condiciones.

Algorithm 1: Método de Bisección

1 Sean $a_0 = a$, $b_0 = b$
2 **for** $k = 1$ **to** N **do**
3 $\quad c_{k-1} = \frac{a_{k-1} + b_{k-1}}{2}$
4 $\quad$ Si $f(c_{k-1})$ cumple con el criterio de parada, terminar
5 $\quad$ Si $f(c_{k-1})f(a_{k-1}) < 0$, entonces $a_k = a_{k-1}$, $b_k = c_{k-1}$
6 $\quad$ Si $f(c_{k-1})f(b_{k-1}) < 0$, entonces $a_k = c_{k-1}$, $b_k = b_{k-1}$
7 **end**

El algoritmo determina tres sucesiones $\{a_n\}_n$, $\{b_n\}_n$ y $\{c_n\}_n$. En cada paso elige a c_n como el punto medio entre a_n y b_n.

Observación 2.1. Las lineas 5 y 6 garantizan que

- $a_n \leq c_n \leq b_n$ para todo n.

- $f(a_n)f(b_n) < 0$ para todo n.
- La sucesión $\{a_n\}_n$ es creciente, mientras que $\{b_n\}_n$ es decreciente.

Veamos que la sucesión $\{c_n\}_n$ converge a una raíz de f.

Lema 2.1.

$$b_n - a_n \leq \frac{b_0 - a_0}{2^n}$$

Demostración. Por inducción en $n \in \mathbb{N}_0$.

Si $n = 0$, no hay nada que ver.

Sea $n > 0$. Es fácil ver que las lineas 5 y 6 del algoritmo eligen a_n y b_n de modo tal que $b_n - a_n \leq \frac{b_{n-1} - a_{n-1}}{2}$. Por hipótesis inductiva, $b_{n-1} - a_{n-1} \leq \frac{b_0 - a_0}{2^{n-1}}$ y usando la desigualdad anterior llegamos al resultado buscado. □

En particular, esto muestra que $\lim_{n\to\infty}(b_n - a_n) = 0$.

Proposición 2.1. *La sucesion $\{c_n\}_n$ es convergente. Más aún $\lim_{n\to\infty} c_n$ es una raíz de f.*

Demostración. Como $a_n \leq b_n$ para todo n, y $\{b_n\}_n$ es decreciente, entonces $\{a_n\}_n$ está acotada superiormente, por ejemplo por b_0. Además $\{a_n\}_n$ es creciente. Luego, es una sucesión convergente. Sea $\ell = \lim_{n\to\infty} a_n$.

Análogamente, $\{b_n\}_n$ es una sucesión decreciente acotada inferiormente, por ejemplo por a_0, con lo cual es convergente. Luego $0 = \lim_{n\to\infty}(b_n - a_n) = \lim_{n\to\infty} b_n - \ell$, con lo cual $\lim_{n\to\infty} b_n = \ell$.

Como $a_n \leq c_n \leq b_n$, entonces $\{c_n\}_n$ está acotada entre dos sucesiones que convergen al mismo límite. Por Sandwich, debe ser $\lim_{n\to\infty} c_n = \ell$. En definitiva, $\{a_n\}_n$, $\{b_n\}_n$ y $\{c_n\}_n$ convergen las tres y tienen el mismo límite.

Falta ver que $f(\ell) = 0$. Si fuera $f(\ell) > 0$ entonces, como f es contínua y a_n tiende a ℓ, tendríamos $f(a_n) > 0$ para todo n suficientemente grande. Análogamente, $f(b_n) > 0$ para todo n suficientemente grande. Pero esto contradice el hecho de que $f(a_n)f(b_n) < 0$ para todo n. Análogamente, no puede ser $f(\ell) < 0$. Entonces no queda otra que $f(\ell) = 0$.

□

Esto demuestra que el método siempre converge a una raíz de la función. En lo que sigue llamaremos $x^* = \lim_{n\to\infty} c_n$.

Observación 2.2. De la monotonía de las sucesiones se deduce que $a_n \leq x^* \leq b_n$ para todo n.

Proposición 2.2.

$$|c_n - x^*| \leq \frac{b_0 - a_0}{2^{n+1}}$$

Demostración. Como $a_n \leq x^* \leq b_n$ y c_n es el punto medio, entonces $|c_n - x^*| \leq \frac{b_n - a_n}{2}$. Por el Lema 2.1, se deduce $|c_n - x^*| \leq \frac{b_0 - a_0}{2^{n+1}}$. □

Corolario 2.1. *El orden de convergencia de $\{c_n\}_n$ es al menos lineal.*

Demostración. Se desprende de que $\frac{b_0 - a_0}{2^{n+1}}$ converge linealmente. □

2.5.1. Ventajas y desventajas

Ventajas

- Para cada a_n y b_n, nos alcanza con conocer el signo de $f(a_n)$ y el de $f(b_n)$ con lo cual podría no ser necesario evaluar la función f en esos puntos. Esto es conveniente en contextos en los cuales la evaluación es una operación costosa y es posible conocer el signo por alguna vía sencilla.
- Tenemos una cota para el error absoluto.
- Es fácil encontrar puntos iniciales a_0 y b_0 factibles.

Desventajas

- Asegura una convergencia de orden al menos lineal, aunque puede resultar lenta.

2.6. Problemas de punto fijo

Definición 2.3. Sea $g : [a, b] \to \mathbb{R}$. Un punto $p \in [a, b]$ se llama punto fijo de g si $g(p) = p$.

Un problema de cálculo de raíces de una función $f(x)$ puede ser transformado en un problema de punto fijo. Por ejemplo si $g(x) = x + f(x)$ entonces p es raíz de f si y solo si p es punto fijo de g.

La conveniencia de transformar el problema de cálculo de una raíz en un problema de punto fijo, se debe a que se conocen varios métodos para resolver esto último, como veremos a continuación.

Proposición 2.3. *Sea $g : [a, b] \to [a, b]$ contínua. Entonces g tiene punto fijo en $[a, b]$. Si además es derivable y $|g'(x)| < 1$ para todo $x \in (a, b)$, entonces el punto fijo es único.*

Demostración. Primero veamos que g tiene un punto fijo en $[a, b]$. Notemos que como $a \leq g(x) \leq b$, entonces $g(a) \geq a$ y $g(b) \leq b$. Si $g(a) = a$, listo. Si no, $g(a) > a$. Análogamente, si $g(b) = b$, terminamos. Si no, $g(b) < b$. Tenemos, entonces, $g(a) - a > 0$ y $g(b) - b < 0$.

Consideremos $f(x) = g(x) - x$. Por lo anterior, $f(a) > 0$ y $f(b) < 0$. Además f es contínua en $[a, b]$ pues g lo es. Luego, por Teorema de Bolzano, existe $p \in (a, b)$ tal que $f(p) = 0$, es decir que $g(p) - p = 0$ o bien $g(p) = p$, con lo cual p es un punto fijo de g.

Veamos la unicidad. Sean $p_1, p_2 \in [a, b]$ puntos fijos de g, entonces $g(p_1) = p_1$ y $g(p_2) = p_2$. Como g es derivable, por el Teorema del Valor Medio tenemos que

$$|g(p_1) - g(p_2)| = |p_1 - p_2| \, |g'(\xi)|$$

para cierto ξ entre p_1 y p_2. Pero $|g(p_1) - g(p_2)| = |p_1 - p_2|$, entonces

$$|p_1 - p_2| = |p_1 - p_2| \, |g'(\xi)|$$

Por hipótesis, $|g'(\xi)| < 1$, con lo cual no queda otra que $|p_1 - p_2| = 0$, i. e., $p_1 = p_2$.

□

Dadas estas condiciones de existencia y unicidad de punto fijo, construimos una sucesión convergente a un punto fijo.

Proposición 2.4. *Sea $g : [a, b] \to [a, b]$ contínua y derivable, tal que existe una constante no negativa k tal que $|g'(x)| \leq k < 1$ para todo $x \in (a, b)$. Sea $\{x_n\}_n$ una sucesión tal que*

$$x_0 \in [a, b]$$
$$x_{n+1} = g(x_n) \text{ si } n \geq 0$$

Entonces $\{x_n\}_n$ converge al único punto fijo de g.

Demostración. Por la Proposición anterior, g posee un único punto fijo en $[a, b]$, que llamamos p. Observemos que como $g(x) \in [a, b]$ y $x_0 \in [a, b]$, entonces todo término de la sucesión cae en $[a, b]$.

Queremos ver que $\lim_{n\to\infty} x_n = p$. Notemos que si $n > 0$,

$$|x_n - p| = |g(x_{n-1}) - g(p)|$$

Como g es derivable, por el Teorema del Valor Medio existe ξ_{n-1} entre x_{n-1} y p, tal que

$$|g(x_{n-1}) - g(p)| = |x_{n-1} - p| \, |g'(\xi_{n-1})|$$

Como $|g'(\xi_{n-1})| \leq k$ entonces

$$|g(x_{n-1}) - g(p)| \leq |x_{n-1} - p|\, k$$

Por la primera de las igualdades, concluimos que

$$|x_n - p| \leq |x_{n-1} - p|\, k$$

Inductivamente, es fácil ver que

$$|x_n - p| \leq |x_0 - p|\, k^n$$

Como $0 \leq k < 1$, $|x_0 - p|\, k^n$ tiende a 0 a medida que $n \to \infty$. Por el Teorema del Sandwich, $\lim\limits_{n\to\infty} |x_n - p| = 0$, con lo cual $\lim\limits_{n\to\infty} x_n = p$.

□

Esta técnica para encontrar el punto fijo se llama *iteración de punto fijo*. Bajo ciertas condiciones podemos asegurar el orden de convergencia de la sucesión.

Proposición 2.5. *Sea $g \in C^r([a,b])$ tal que $p \in (a,b)$ es punto fijo y*

$$g'(p) = g''(p) = \cdots = g^{(r-1)}(p) = 0,\ g^{(r)}(p) \neq 0$$

Entonces, si $x_{n+1} = g(x_n)$ converge a p, su orden de convergencia es r.

Demostración. Como $g \in C^r([a,b])$, consideramos el desarrollo de Taylor de g de orden $r-1$, centrado en p,

$$g(x) = g(p) + g'(p)(x-p) + \cdots + \frac{g^{(r-1)}(p)}{(r-1)!}(x-p)^{r-1} + \frac{g^{(r)}(\xi_x)}{r!}(x-p)^r = g(p) + \frac{g^{(r)}(\xi_x)}{r!}(x-p)^r$$

con ξ_x entre x y p. Evaluando en x_n obtenemos

$$g(x_n) = g(p) + \frac{g^{(r)}(\xi_n)}{r!}(x_n - p)^r$$

con ξ_n entre x_n y p. Como $g(x_n) = x_{n+1}$ y $g(p) = p$,

$$x_{n+1} - p = \frac{g^{(r)}(p)}{r!}(x_n - p)^r$$

o sea que

$$\frac{x_{n+1} - p}{(x_n - p)^r} = \frac{g^{(r)}(p)}{r!}$$

Tomando límite en ambos miembros,

$$\lim_{n\to\infty} \frac{x_{n+1} - p}{(x_n - p)^r} = \frac{g^{(r)}(p)}{r!}$$

Como esta expresión converge, su módulo también lo hace, y

$$\lim_{n\to\infty} \frac{|x_{n+1} - p|}{|x_n - p|^r} = \frac{|g^{(r)}(p)|}{r!}$$

Como $g^{(r)}(p) \neq 0$ entonces el límite $\frac{|g^{(r)}(p)|}{r!}$ es no nulo. Luego, el orden de convergencia de $\{x_n\}_n$ es r.

□

2.7. Método de Newton

Supongamos que dada una $f : \mathbb{R} \to \mathbb{R}$, queremos hallar una raíz. Para esto queremos encontrar una función $g : \mathbb{R} \to \mathbb{R}$, cuya expresión involucre a f, de modo tal que los puntos fijos de g sean raíces de f. Más aún, vamos a pedir que la iteración de punto fijo asociada a g converja cuadráticamente.

Proponemos $g(x) = x - h(x)f(x)$ siendo h una función que todavía no conocemos. Notemos que si p es punto fijo de g entonces $p = g(p) = p - h(p)f(p)$, con lo cual si $h(p) \neq 0$ resulta que $f(p) = 0$. Por lo tanto, todo punto fijo de g es una raíz de f, sea cual sea la función h tal que $h(p) \neq 0$.

Como queremos un orden de convergencia cuadrático, pedimos que $g'(p) = 0$. Como $g'(x) = 1 - h'(x)f(x) - h(x)f'(x)$, entonces

$$\begin{aligned} g'(p) = 0 &\Leftrightarrow 1 - h'(p)f(p) - h(p)f'(p) = 0 \\ &\Leftrightarrow 1 - h(p)f'(p) = 0 \\ &\Leftrightarrow h(p) = \frac{1}{f'(p)} \end{aligned}$$

suponiendo que $f'(p) \neq 0$. Entonces podemos tomar $h(x) = \frac{1}{f'(x)}$, que cumple $h(p) \neq 0$. Así, queda

$$g(x) = x - \frac{f(x)}{f'(x)}$$

con lo cual la iteración de punto fijo asociada es

$$x_{n+1} = g(x_n) = x_n - \frac{f(x_n)}{f'(x_n)}$$

En lo que sigue vamos a ver que esta sucesión efectivamente converge a un punto fijo de g (bajo ciertas condiciones), y lo hace cuadráticamente.

Proposición 2.6. *Sea $f \in C^2([a, b])$. Sea $p \in (a, b)$ tal que $f(p) = 0$, $f'(p) \neq 0$. Entonces existe $\delta > 0$ tal que toda sucesión $\{x_n\}_n$ con*

$$\begin{aligned} x_0 &\in [p - \delta, p + \delta] \\ x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \text{ si } n \geq 0 \end{aligned}$$

está bien definida y converge.

Demostración. Sea $g(x) = x - \frac{f(x)}{f'(x)}$.

- g **está bien definida en un entorno de** p**.** Como f' es contínua en $[a, b]$ y $p \in (a, b)$ entonces f' es contínua en p. En particular, existe $\delta_1 > 0$ tal que $f'(x) \neq 0$ para todo $x \in [p - \delta_1, p + \delta_1]$.
- g **es estable en un entorno de** p**.** Notemos que $g'(x) = \frac{f(x)f''(x)}{f'(x)^2}$, que está bien definida en $[p - \delta_1, p + \delta_1]$.

  Como $g'(p) = 0$ y g' es contínua en p, entonces existe $\delta_2 > 0$, $\delta_2 \leq \delta_1$ tal que, para alguna constante k, $|g'(x)| \leq k < 1$ para todo $x \in (p - \delta_2, p + \delta_2)$.

  Luego, si $x \in [p - \delta_2, p + \delta_2]$, utilizando el Teorema del Valor Medio, deducimos que

  $$|g(x) - g(p)| = |g'(\xi_x)|\,|x - p|$$

  para ξ_x entre x y p. Entonces $|g'(\xi_x)| < 1$. Como además $g(p) = p$ resulta que

  $$|g(x) - p| < |x - p| \leq \delta_2$$

  Luego $|g(x) - p| \leq \delta_2$, es decir que $g(x) \in [p - \delta_2, p + \delta_2]$.

Por lo tanto, $g : [p-\delta_2, p+\delta_2] \to [p-\delta_2, p+\delta_2]$ está bien definida y además $|g'(x)| \leq k < 1$ para todo $x \in (p-\delta_2, p+\delta_2)$. Poniendo $\delta = \delta_2$, una sucesión $\{x_n\}_n$ tal que $x_0 \in [p-\delta, p+\delta]$ y $x_{n+1} = g(x_n) = x_n - \frac{f(x_n)}{f'(x_n)}$ está en las condiciones de la Proposición 2.4, con lo cual converge a un punto fijo de g.

□

Observemos que este resultado sólo asegura la convergencia cuando comenzamos la iteración dentro de un entorno suficientemente chico del punto fijo. En general, no conocemos la localización del punto fijo así como tampoco cuán chico debe ser el entorno. Por esta razón, se suele partir desde un punto arbitrario, e iterar, revisando si se logra converger. En caso negativo, se elige otro punto de partida y se repite la iteración.

Otra forma de sortear el problema es aplicar, inicialmente, algunas iteraciones del método de Bisección y cuando el entorno que contiene a la raíz es suficientemente chico, proceder con Newton.

Proposición 2.7. *Sea $f \in C^2$ y $p \in \mathbb{R}$ una raíz de f. Sea $\{x_n\}_n$ la iteración de punto fijo dada por $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$. Supongamos que está bien definida y que nunca vale p. Si $e_n = x_n - p$, entonces*

$$\frac{e_{n+1}}{e_n^2} = \frac{1}{2}\frac{f''(\xi_n)}{f'(x_n)}$$

con ξ_n un valor entre x_n y p. Más aún, si $f'(p), f''(p) \neq 0$ y la iteración converge a p, entonces lo hace con orden cuadrático.

Demostración. Para cada $n \in \mathbb{N}_0$, consideremos el desarrollo de Taylor de f de orden 1 centrado en x_n,

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{f''(\xi_n)}{2}(x - x_n)^2$$

con ξ_n entre x_n y x. Evaluando en p obtenemos

$$f(p) = f(x_n) + f'(x_n)(p - x_n) + \frac{f''(\xi_n)}{2}(p - x_n)^2$$

con ξ_n entre x_n y p. Luego,

$$0 = f(x_n) - f'(x_n)e_n + \frac{f''(\xi_n)}{2}e_n^2$$

Con un poco de aritmética llegamos a la ecuación buscada,

$$\frac{e_{n+1}}{e_n^2} = \frac{1}{2}\frac{f''(\xi_n)}{f'(x_n)}$$

Resta ver que si $f'(p) \neq 0$ y $\{x_n\}_n$ converge a p, entonces lo hace cuadráticamente. Para esto, debemos ver que $\frac{|e_{n+1}|}{|e_n|^2}$ converge a una constante no nula. Tomando módulo en la ecuación anterior,

$$\frac{|e_{n+1}|}{|e_n|^2} = \frac{1}{2}\frac{|f''(\xi_n)|}{|f'(x_n)|}$$

Como f' es contínua, y $\lim_{n\to\infty} x_n = p$, entonces

$$\lim_{n\to\infty} f'(x_n) = f'(p) \neq 0$$

Por otro lado, como ξ_n siempre cae entre x_n y p, y f'' es contínua, resulta que

$$\lim_{n\to\infty} f''(\xi_n) = f''(p)$$

Por lo tanto, estas dos últimas expresiones convergen en módulo y, más aún,

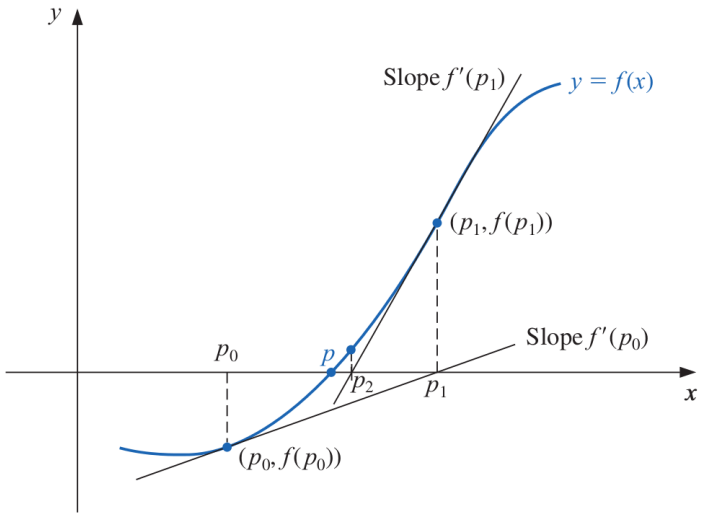
$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^2} = \lim_{n\to\infty} \frac{1}{2}\frac{|f''(\xi_n)|}{|f'(x_n)|} = \frac{1}{2}\frac{|f''(p)|}{|f'(p)|} \neq 0$$

que es lo que queríamos probar.

□

2.7.1. Interpretación geométrica

Este método tiene una muy linda interpretación geométrica. Dada una función f, tomamos un punto p_0 y consideramos la recta tangente que pasa por el punto $(p_0, f(p_0))$. Definimos p_1 como la intersección entre dicha recta y el eje x. Ahora tomamos la recta tangente que pasa por el punto $(p_1, f(p_1))$ y definimos p_2 como la intersección entre esta recta y el eje x. Continuando de esta forma nos iremos aproximando cada vez más a una raíz p de f. La siguiente imágen ilustra este proceso.



Sea $\{x_n\}_n$ una iteración de Newton. Dado x_n, definimos x_{n+1} como la abscisa de la intersección entre la recta tangente a f en x_n y el eje x. La recta tangente a f en x_n es $y = f'(x_n)x + b$ siendo b la ordenada al origen. Para $x = x_n$, esta recta toma el valor $y = f(x_n)$, con lo cual $f(x_n) = f'(x_n)x_n + b$, i. e., $b = f(x_n) - x_n f'(x_n)$. Luego, la recta tangente a f en x_n es

$$y = f'(x_n)x + f(x_n) - x_n f'(x_n)$$

La intersección entre esta recta y el eje x se da para un valor x tal que $f'(x_n)x + f(x_n) - x_n f'(x_n) = 0$, i. e., $x = x_n - \frac{f(x_n)}{f'(x_n)}$. Habíamos definido x_{n+1} como este valor, con lo cual

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

2.7.2. Ventajas y desventajas

Ventajas

- Bajo ciertas condiciones, el método converge con velocidad cuadrática, lo que significa que, a nivel práctico, el número de cifras decimales correctas calculadas se duplica a cada paso.

Desventajas

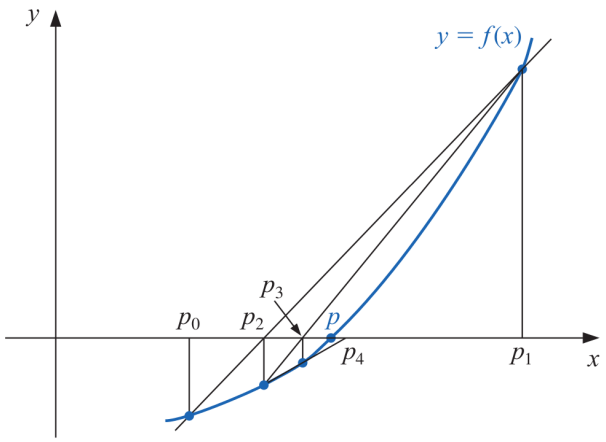
- La convergencia está asegurada sólo en un entorno del punto fijo buscado. En general, no conocemos dicho punto y mucho menos cuán cerca del mismo debemos comenzar a iterar.
- La velocidad de convergencia, a nivel práctico, puede resultar lenta lejos del punto fijo. Esto se debe a que, como vimos antes,

$$\frac{e_{n+1}}{e_n^2} = \frac{1}{2}\frac{f''(\xi_n)}{f'(x_n)}$$

  con lo cual, si $|f'(x_n)|$ es chico, el error decrece lento.
- Es necesario computar $f'(x_n)$ en cada iteracion. En ciertas situaciones esto puede ser caro de computar o directamente imposible. El método de la Secante viene a salvar este problema.

2.8. Método de la Secante

El método comienza con dos puntos x_0 y x_1. La iteración x_{n+1} es igual a la iteración de Newton, excepto que aproximamos $f'(x_n)$ por la pendiente de la recta secante que pasa por $f(x_n)$ y $f(x_{n-1})$.



Explícitamente, la aproximación que estamos usando es

$$f'(x_n) = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

Por lo tanto, la sucesión que usa este método está dada por

$$x_{n+1} = x_n - \frac{f(x_n)}{\frac{f(x_n)-f(x_{n-1})}{x_n - x_{n-1}}} = x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}$$

Se puede demostrar que la velocidad de este método es supralineal.

2.8.1. Ventajas y desventajas

Ventajas

- No es necesario computar ningún valor de f'.

Desventajas

- La velocidad de convergencia es más lenta que la del método de Newton.
- A medida que n crece y x_n va convergiendo, $x_n \approx x_{n-1}$ y $f(x_n) \approx f(x_{n-1})$, lo que produce una pérdida de dígitos significativos debido al redondeo en el término $\frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}$. El método Regula Falsi provee una solución a este problema.

2.9. Método Regula Falsi

Genera aproximaciones del mismo modo que el método de la Secante, pero en cada paso verifica que la raíz buscada quede entre dos iteraciones sucesivas, en forma análoga al método de Bisección.

Comienza con dos puntos x_0 y x_1 tal que $f(x_0)f(x_1) < 0$. Dados x_{n-1} y x_n, el punto x_{n+1} se calcula mediante la intersección de la recta secante que pasa por $f(x_{n-1})$ y $f(x_n)$, y el eje x. Luego, redefine x_n como aquel valor x_{n-1} o x_n que tiene distinto signo (evaluado en f) que x_{n+1}. La fórmula de la iteración sigue siendo la misma del método de la Secante.

En otras palabras, es idéntico al método de Bisección excepto que, en lugar de quedarse con el punto medio, toma la intersección entre la secante y el eje x.

Se puede probar que la velocidad de convergencia del método es al menos $\varphi = \frac{1+\sqrt{5}}{2}$.

2.9.1. Ventajas y desventajas

Ventajas

- Desde el punto de vista numérico, ahora $f(x_n)$ y $f(x_{n-1})$ siempre tendrán distinto signo, con lo cual $f(x_n) - f(x_{n-1})$ es una resta entre números de distinto signo (podemos pensar que es una suma), evitando una cancelación catastrófica.

Desventajas

- Requiere más operaciones que el método de la secante.

3. Sistemas de ecuaciones lineales

3.1. Problema

Dadas n ecuaciones, cada una con n incógnitas

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ \vdots \\ a_{1n}x_1 + \cdots + a_{nn}x_n = b_n \end{cases}$$

con coeficientes reales, queremos hallar valores de $x_1, \cdots, x_n$ que satisfagan todas las ecuaciones simultaneamente. En forma matricial escribimos el problema definiendo

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

de modo tal que buscamos soluciones de la ecuación $Ax = b$.

3.2. Existencia y unicidad de la solución

Notemos que

$$Ax = b \Leftrightarrow x_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{nn} \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

En otras palabras, $Ax = b$ tiene solución si y sólo si b es combinación lineal de las columnas de A. Usando esto podemos separar en casos:

- Si las columnas de A son linealmente independientes, entonces forman una base de $\mathbb{R}^n$ (porque son n), y en consecuencia la combinación lineal existe y es única.
- Si las columnas de A no son linealmente independientes, entonces una tal combinación lineal puede existir como no. Si existe alguna entonces hay infinitas combinaciones lineales que cumplen la condición.

En definitiva $Ax = b$ tiene solución única si y sólo si las columnas de A son l. i. Si no son l. i. entonces hay inifinitas soluciones o no hay ninguna.

3.3. Resolución de un sistema lineal

1. **Caso A diagonal**. El sistema tiene la forma

$$\begin{cases} a_{11}x_1 & & = b_1 \\ & \ddots & \\ & & a_{nn}x_n = b_n \end{cases}$$

   Pueden pasar dos cosas:

   - Si $a_{ii} \neq 0$ para todo i, entonces la única solución es $x_i = \frac{b_i}{a_{ii}}$, $i = 1, \cdots, n$.
   - Si no, existe $i_0 \in \{1, \cdots, n\}$ tal que $a_{i_0 i_0} = 0$. Entonces la ecuación $a_{i_0 i_0} x_{i_0} = b_{i_0}$ es equivalente a $b_{i_0} = 0$.
     - Si $b_{i_0} \neq 0$ entonces tenemos una contradicción y el sistema no tiene solución.

- Si $b_{i_0} = 0$ entonces la ecuación se cumple para cualquier x_{i_0}. Si existe solución (depende de lo que suceda con las restantes ecuaciones) entonces habrá infinitas.

El costo del cómputo en este caso es claramente $\mathcal{O}(n)$.

2. **Caso A triangular superior**. El sistema tiene la forma

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ \qquad\quad a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \qquad\qquad\quad \vdots \\ \qquad\qquad\qquad\quad a_{nn}x_n = b_n \end{cases}$$

Pueden pasar dos cosas:

- Si $a_{ii} \neq 0$ para todo i, entonces las columnas de A son l. i. y la solución es única. Más aún, es

$$\begin{aligned} x_n &= \frac{b_n}{a_{nn}} \\ x_{n-1} &= \frac{b_{n-1} - a_{(n-1)n}x_n}{a_{(n-1)(n-1)}} \\ &\vdots \\ x_1 &= \frac{b_1 - a_{12}x_2 - a_{13}x_3 - \cdots - a_{1n}x_n}{a_{11}} \end{aligned}$$

  Este algoritmo para computar las soluciones de un sistema de ecuaciones triangular superior determinado, recibe el nombre de *backward substitution*. Cuando el sistema es triangular inferior determinado, la sustitución se realiza comenzando desde la primer ecuación, y recibe el nombre de *forward substitution*.
- Si no, existe $i_0 \in \{1, \cdots, n\}$ tal que $a_{i_0 i_0} = 0$. Entonces la ecuación $a_{i_0 i_0}x_{i_0} = b_{i_0} - a_{i_0(i_0+1)}x_{i_0+1} - \cdots - a_{i_0 n}x_n$ es equivalente a $0 = b_{i_0} - a_{i_0(i_0+1)}x_{i_0+1} - \cdots - a_{i_0 n}x_n$.
  - Si el lado derecho es distinto de 0 entonces tenemos una contradicción y el sistema no tiene solución.
  - Si no, la ecuación se cumple para cualquier x_{i_0}. Si existe solución (depende de lo que suceda con las restantes ecuaciones) entonces habrá infinitas.

El costo del cómputo en este caso es $\mathcal{O}(n^2)$, pues para la incógnita x_i ($i = 1, \cdots, n$) se realizan $\mathcal{O}(n)$ operaciones, y son n incógnitas en total.

3. **Caso general**. Para resolver el problema para una matriz A cualquiera, la transformaremos mediante operaciones elementales de filas y luego usaremos backward substitution. El algoritmo para transformar el sistema en uno triangular superior se conoce como *eliminación gaussiana*.

3.4. Eliminación gaussiana sin pivoteo

El algoritmo opera sobre la matriz ampliada del sistema. Llamamos

$$M^{(k)} = \left(\begin{array}{ccc|c} a_{11}^{(k)} & \cdots & a_{1n}^{(k)} & b_1^{(k)} \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1}^{(k)} & \cdots & a_{nn}^{(k)} & b_n^{(k)} \end{array}\right)$$

a la matriz ampliada luego de k pasos del algoritmo. Inicialmente, asumamos que $a_{(k+1)(k+1)}^{(k)} \neq 0$ para todo k. La entrada es $M^{(0)} = \left(\begin{array}{c|c} A & b \end{array}\right)$. En el primer paso, el algoritmo de eliminación gaussiana fuerza, mediante operaciones de filas, la aparición de ceros en la primer columna, debajo de la diagonal,

`Paso 1:` $F_i \leftarrow F_i - \frac{a_{i1}^{(0)}}{a_{11}^{(0)}} F_1$ para $i = 2, \cdots, n$

`Paso 2:` $F_i \leftarrow F_i - \frac{a_{i2}^{(1)}}{a_{22}^{(1)}} F_2$ para $i = 3, \cdots, n$

En el paso $k < n$, colocamos ceros en la columna k, debajo de la diagonal,

`Paso k:` $F_i \leftarrow F_i - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} F_k$ para $i = k+1, \cdots, n$

El último será el

`Paso n - 1:` $F_i \leftarrow F_i - \frac{a_{i(n-1)}^{(n-2)}}{a_{(n-1)(n-1)}^{(n-2)}} F_{n-1}$ para $i = n$

El invariante fundamental de este proceso es que al principio del paso k, las primeras $k-1$ columnas tienen únicamente ceros debajo de la diagonal. Al concluir los $n-1$ pasos obtendremos la matriz $M^{(n-1)}$ que es triangular superior.

Algorithm 2: Eliminación gaussiana sin pivoteo

```
for k = 1 to n − 1 do
    for i = k + 1 to n do
        m_ik = a_ik / a_kk
        F_i ← F_i − m_ik F_k
    end
end
```

La complejidad del algoritmo es claramente $\mathcal{O}(n^3)$, pues para cada una de las n columnas se hacen $\mathcal{O}(n)$ operaciones de fila, y cada una de estas operaciones de fila son $\mathcal{O}(n)$ operaciones escalares. Un análisis más fino muestra que la cantidad de operaciones de punto flotante que realiza la eliminación gaussiana es aproximadamente $\frac{1}{3}n^3$.

3.5. Eliminación gaussiana con pivoteo

Se llama *pivote* al elemento $a_{kk}^{(k-1)}$ de cada iteración del algoritmo, utilizado para computar los coeficientes m_{ik} e introducir ceros debajo de la diagonal. Puede suceder que al principio de cierto paso de la eliminación gaussiana, digamos el k, el elemento $a_{kk}^{(k-1)}$ sea 0. En este caso no podemos usar $a_{kk}^{(k-1)}$ como pivote. Tenemos dos posibilidades:

- Si $a_{ik}^{(k-1)} = 0$ para todo $i = k+1, \cdots, n$, entonces la columna k ya tiene todos ceros debajo de la diagonal, y por lo tanto no es necesario realizar ninguna acción en este paso.
- Si no, debemos intercambiar la fila k por alguna otra fila $i_0 > k$ tal que $a_{i_0 k}^{(k-1)} \neq 0$, y continuar con la eliminación normalmente.

Repitiendo esto cada vez que ocurra el problema a lo largo de la ejecución del algoritmo, llegaremos nuevamente a una matriz triangular inferior.

Observación 3.1. Si ahora esas mismas permutaciones de filas fueran realizadas en el mismo orden sobre A, entonces se puede probar que la matriz que se obtiene admite eliminación gaussiana sin pivoteo. En términos matriciales, existe una matriz $P \in \mathbb{R}^{n \times n}$ que es producto de matrices elementales de permutación tal que PA admite eliminación gaussiana sin pivoteo.

Observación 3.2. Toda matriz $A \in \mathbb{R}^{n \times n}$ admite eliminación gaussiana con pivoteo, aunque no necesariamente sin pivoteo.

3.6. Pivoteo parcial

Los coeficientes m_{ik} de cada paso dependen de $a_{ik}^{(k-1)}$ y $a_{kk}^{(k-1)}$ que son resultados de operaciones realizadas a lo largo de las iteraciones $1, \cdots, k-1$ y que, en consecuencia, acarrean un error de redondeo. Si al computar el cociente $\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}$ resulta que el divisor $a_{kk}^{(k-1)}$ es un número con valor absoluto pequeño, el error que acumulaba a_{ik} se ve amplificado. Por esta razón, es deseable que a_{kk} tenga valor absoluto lo más grande posible. Con este objetivo se implementa la estrategia de *pivoteo parcial*: al comienzo del k-ésimo paso, se permuta la fila k con alguna fila $i_0 \geq k$ tal que $|a_{i_0 k}^{(k-1)}| = \max_{k \leq i \leq n} |a_{ik}^{(k-1)}|$.

Cuanto mayor sea el pivote, más pequeño será m_{ik} y en consecuencia la operación $F_i \leftarrow F_i - m_{ik}F_k$ sobre la matriz ampliada del sistema no hará variar sus entradas en gran medida. Intuitivamente, la variación del valor absoluto de los elementos de la matriz ampliada a lo largo de las iteraciones es una medida de la efectividad del pivoteo parcial. En este sentido definimos lo siguiente.

Definición 3.1. Sea $A \in \mathbb{R}^{n\times n}$. Como antes, sea $A^{(k)}$ la matriz ampliada de un sistema que tiene a A como matriz, luego de k pasos de eliminación gaussiana con pivoteo parcial. Sea $a_k = \max\limits_{1\leq i,j\leq n} |a_{ij}^{(k)}|$. Definimos el factor de crecimiento de A como

$$\rho = \frac{\max\limits_{1\leq k\leq n-1} a_k}{a_0}$$

Por lo dicho arriba, es conveniente que ρ sea lo más pequeño posible. El siguiente resultado establece una cota para este factor en ciertos tipos de matrices.

Proposición 3.1. *Sea $A \in \mathbb{R}^{n\times n}$.*

- *Si A es una matriz arbitraria entonces $\rho \leq 2^{n-1}$.*
- *Si A es una matriz de Hessemberg (sus coeficientes debajo de la primer subdiagonal son nulos) entonces $\rho \leq n$.*
- *Si A es una matriz tridiagonal entonces $\rho \leq 2$.*

Si bien en una matriz arbitraria el factor de crecimiento es a lo sumo 2^{n-1}, este máximo ocurre rara vez y, en general, la estrategia de pivoteo parcial es una estrategia numéricamente estable.

3.7. Factorización LU

Definición 3.2. Sea $A \in \mathbb{R}^{n\times n}$. Se llama factorización LU de A a una escritura de la forma $A = LU$ con $L \in \mathbb{R}^{n\times n}$ triangular inferior con unos en la diagonal y $U \in \mathbb{R}^{n\times n}$ triangular superior.

Veamos cómo calcular la factorización LU de una matriz A. Supongamos que A admite eliminación gaussiana sin pivoteo. Notemos que en el paso k de la eliminación se realizan las operaciones de fila $F_i = F_i - m_{ik}F_k$ para $i = k+1, \cdots, n$. En términos matriciales esto es multiplicar $A^{(k)}$ a izquierda por las matrices de elementales

$$M_k = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & 0 \\ \vdots & & 1 & & \vdots \\ 0 & \cdots & -m_{(k+1)k} & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & 0 \\ \vdots & & 1 & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & -m_{nk} & \cdots & 1 \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & 0 \\ \vdots & & 1 & & \vdots \\ 0 & \cdots & -m_{(k+1)k} & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & -m_{nk} & \cdots & 1 \end{pmatrix}$$

Al cabo de los $n-1$ pasos de eliminación gaussiana tendremos que $M_{n-1}\cdots M_1 A = U$ con U triangular superior.

Observemos que M_k es inversible para cada k, puesto que M_k es triangular inferior con unos en la diagonal. Entonces podemos escribir $A = M_1^{-1}\cdots M_{n-1}^{-1}U = LU$, con $L = M_1^{-1}\cdots M_{n-1}^{-1}$. Si bien ya hemos llegado a la factorización LU de A, podemos simplificar la expresión de L. Notemos que $M_k = I_n - m_k e_k^t$ con $m_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ m_{(k+1)k} \\ \vdots \\ m_{nk} \end{pmatrix}$ y e_k el k-ésimo vector canónico de $\mathbb{R}^n$. Usando esta escritura es fácil ver que $M_k^{-1} = I_n + m_k e_k^t$, ya que

$$\begin{aligned}(I_n - m_k e_k^t)(I_n + m_k e_k^t) &= I_n^2 + m_k e_k^t - m_k e_k^t - m_k e_k^t m_k e_k^t \\ &= I_n^2 - m_k e_k^t m_k e_k^t \\ &= I_n \qquad\qquad (\text{pues } e_k^t m_k = 0)\end{aligned}$$

Luego,

$$\begin{aligned}L &= M_1^{-1} \cdots M_{n-1}^{-1} \\ &= (I_n + m_1 e_1^t) \cdots (I_n + m_{n-1} e_{n-1}^t) \\ &= I_n + m_1 e_1^t + \cdots + m_{n-1} e_{n-1}^t \qquad\qquad (\text{pues } e_i^t m_j = 0 \text{ si } i \leq j)\end{aligned}$$

En definitiva $A = LU$ con $L = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 \\ m_{21} & \ddots & 0 & \cdots & 0 \\ \vdots & & 1 & & \vdots \\ m_{(k+1)1} & \cdots & m_{(k+1)k} & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ m_{n1} & \cdots & m_{nk} & \cdots & 1 \end{pmatrix}$.

Debido a lo inherente que es la factorización LU al proceso de eliminación gaussiana el costo de computar esta factorización es claramente $\mathcal{O}(n^3)$.

3.7.1. Existencia y unicidad de la factorización LU

Hemos probado lo siguiente,

Proposición 3.2. *Sea $A \in \mathbb{R}^{n\times n}$. Si A admite eliminación gaussiana sin pivoteo entonces A tiene factorización LU.*

Observación 3.3. La recíproca no es cierta. Por ejemplo, la matriz nula tiene factorización LU pero obviamente no admite eliminación gaussiana sin pivoteo.

Observación 3.4. No toda matriz tiene factorización LU. Si $A \in \mathbb{R}^{2\times 2}$ tiene factorización LU entonces existen $a, b, c, d \in \mathbb{R}$ tal que

$$A = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix} \begin{pmatrix} b & c \\ 0 & d \end{pmatrix} = \begin{pmatrix} b & c \\ ab & ac + d \end{pmatrix}$$

Teniendo en cuenta esta condición necesaria, es fácil ver que la matriz $\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ no tiene factorización LU.

Observación 3.5. Remitiéndonos a la Observación 3.1, deducimos que existe $P \in \mathbb{R}^{n\times n}$ producto de matrices de permutación tal que PA tiene factorización LU. Pero P es producto de matrices inversibles, con lo cual existe P^{-1}, y llegamos a la escritura $A = P^{-1}LU$, que se llama factorización PLU. Observemos que como una tal matriz P siempre existe entonces también existirá siempre una factorización PLU.

Con respecto a la unicidad, se tiene el siguiente resultado,

Proposición 3.3. *Sea $A \in \mathbb{R}^{n\times n}$ inversible. Si A tiene factorización LU entonces dicha escritura es única.*

Demostración. Sean $A = L_1U_1 = L_2U_2$ dos escrituras LU de A. Como A, L_1 y L_2 son inversibles, entonces U_1 y U_2 también lo son. Luego $L_1U_1 = L_2U_2 \Rightarrow U_1U_2^{-1} = L_1^{-1}L_2$. Como U_2 es triangular superior entonces U_2^{-1} también lo es. Como producto de matrices triangulares superiores es triangular superior, entonces $U_1U_2^{-1}$ es triangular superior. Análogamente $L_1^{-1}L_2$ resulta triangular inferior. Sin embargo este último producto tiene una característica distintiva. Como L_1 tiene unos en la diagonal, L_1^{-1} los preserva, y como L_2 también tiene unos en la diagonal, entonces el producto $L_1^{-1}L_2$ también tiene unos en la diagonal.

Entonces $U_1U_2^{-1} = L_1^{-1}L_2$, siendo $U_1U_2^{-1}$ triangular superior y $L_1^{-1}L_2$ triangular inferior. Luego, ambos productos deben ser una matriz diagonal. Más aún, como $L_1^{-1}L_2$ tiene unos en la diagonal, entonces esta matriz diagonal es la identidad. Finalmente, $U_1U_2^{-1} = I_n \Rightarrow U_1 = U_2$ y $L_1^{-1}L_2 = I_n \Rightarrow L_1 = L_2$, es decir, las escrituras son idénticas.

□

Observación 3.6. En general, la factorización LU de una matriz no es única. Por ejemplo, la matriz nula admite infinitas factorizaciones LU.

3.7.2. Aplicación

Supongamos que dada una matriz A con factorización LU y vectores $b_1, \cdots, b_k$ queremos calcular una solución de $Ax = b_i$ para cada $i = 1, \cdots, k$. Si usáramos el algoritmo de eliminación gaussiana junto con backward substitution para resolver cada uno de los sistemas, tendríamos un costo de $\mathcal{O}(n^3)$ por cada sistema.

Por otro lado, factoricemos $A = LU$. Esta escritura se puede calcular, como dijimos antes, en $\mathcal{O}(n^3)$. Para cada sistema $Ax = b_i$ hacemos:

- Calculamos una solución de $Ly = b_i$, mediante forward substitution.
- Calculamos una solución de $Ux = y$, mediante backward substitution.

Esto requiere tiempo $\mathcal{O}(n^2)$. Notemos que como L es inversible, $Ux = y \Leftrightarrow LUx = Ly \Leftrightarrow Ax = b_i$. En definitiva, podemos resolver $Ax = b_i$ en $\mathcal{O}(n^2)$, para cada i, pagando una única vez, inicialmente, un costo de $\mathcal{O}(n^3)$.

3.7.3. Familias de matrices que admiten factorización LU

Definición 3.3. Una matriz $A \in \mathbb{R}^{n\times n}$ se dice estrictamente diagonal dominante por filas si

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$$

para todo $i = 1, \cdots, n$.

Proposición 3.4. *Si $A \in \mathbb{R}^{n\times n}$ es e. d. d. f. entonces A admite factorización LU.*

Demostración. Lo probamos por inducción en $n \in \mathbb{N}$. Escribamos $A = (a_{ij})_{i,j}$.

Si $n = 1$ la factorización es trivial.

Sea $n > 1$. Escribimos la matriz A por bloques en la forma

$$A = \left(\begin{array}{c|c} a_{11} & v^t \\ \hline u & A_{n-1} \end{array}\right)$$

donde $u = \begin{pmatrix} a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}$, $v^t = (a_{12} \ \cdots \ a_{1n})$ y A_{n-1} es la submatriz que se obtiene sacandole la primera fila y columna a A. Como $a_{11} \neq 0$, podemos aplicar el primer paso de eliminación gaussiana sin pivoteo,

$$\left(\begin{array}{c|c} 1 & 0\cdots 0 \\ \hline -u/a_{11} & I_{n-1} \end{array}\right)\left(\begin{array}{c|c} a_{11} & v^t \\ \hline u & A_{n-1} \end{array}\right) = \left(\begin{array}{c|c} a_{11} & v^t \\ \hline 0 & \\ \vdots & -uv^t/a_{11} + A_{n-1} \\ 0 & \end{array}\right)$$

Veamos que el bloque $B = -uv^t/a_{11} + A_{n-1}$ es una matriz e. d. d. f. Por comodidad supondremos que los índices de las matrices de dimensión $(n-1)\times(n-1)$ que estamos considerando corren de 2 hasta n. Si $2 \leq i, j \leq n$ son cualesquiera entonces

$$\begin{aligned} B_{ij} &= (-uv^t/a_{11} + A_{n-1})_{ij} \\ &= (-uv^t/a_{11})_{ij} + (A_{n-1})_{ij} \\ &= -1/a_{11}(uv^t)_{ij} + a_{ij} \\ &= a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}} \end{aligned}$$

Luego,

$$\begin{aligned} \sum_{\substack{j=2 \\ j\neq i}}^{n} |B_{ij}| &= \sum_{\substack{j=2 \\ j\neq i}}^{n} \left| a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}} \right| \\ &\leq \sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{ij}| + \sum_{\substack{j=2 \\ j\neq i}}^{n} \left| \frac{a_{i1}a_{1j}}{a_{11}} \right| \\ &= \sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{ij}| + \left| \frac{a_{i1}}{a_{11}} \right| \sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{1j}| \end{aligned}$$

Pero $\sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{ij}| = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| - |a_{i1}| < |a_{ii}| - |a_{i1}|$, esto último por ser A estrictamente diagonal dominante por filas. Por la misma razón se tiene $\sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{1j}| = \sum_{j=2}^{n} |a_{1j}| - |a_{1i}| < |a_{11}| - |a_{1i}|$. Entonces,

$$\begin{aligned} \sum_{\substack{j=2 \\ j\neq i}}^{n} |B_{ij}| &< |a_{ii}| - |a_{i1}| + \left| \frac{a_{i1}}{a_{11}} \right| (|a_{11}| - |a_{1i}|) \\ &= |a_{ii}| - |a_{i1}| + |a_{i1}| - \left| \frac{a_{i1}a_{1i}}{a_{11}} \right| \\ &= |a_{ii}| - \left| \frac{a_{i1}a_{1i}}{a_{11}} \right| \end{aligned}$$

Finalmente, como $|a_{ii}| - \left| \frac{a_{i1}a_{1i}}{a_{11}} \right| \leq \left| a_{ii} - \frac{a_{i1}a_{1i}}{a_{11}} \right|$,

$$\sum_{\substack{j=2 \\ j\neq i}}^{n} |B_{ij}| < \left| a_{ii} - \frac{a_{i1}a_{1i}}{a_{11}} \right| = |B_{ii}|$$

Que es lo que queríamos probar. Entonces, por hipótesis inductiva, B admite factorización LU. Sea $B = L_{n-1}U_{n-1}$ una tal escritura. Sean

$$L = \left(\begin{array}{c|c} 1 & 0 \cdots 0 \\ \hline v/a_{11} & L_{n-1} \end{array} \right) \qquad U = \left(\begin{array}{c|c} a_{11} & u^t \\ \hline 0 & \\ \vdots & U_{n-1} \\ 0 & \end{array} \right)$$

Notemos que L es triangular inferior con unos en la diagonal y U es triangular superior. Se puede ver, haciendo el producto por bloques, que $A = LU$, con lo cual esta es la factorización LU de A. Esto concluye el paso inductivo.

□

Proposición 3.5. *Sea $A \in \mathbb{R}^{n\times n}$ inversible. Entonces A tiene factorización LU si y sólo si sus n menores principales son no nulos.*

Demostración. Denotamos A_i a la submatriz de A formada por las primeras i filas e i columnas.

($\Rightarrow$) Queremos ver que $\det(A_i) \neq 0$ para todo $i = 1, \cdots, n$. Procedemos por inducción en $n \in \mathbb{N}$.

Si $n = 1$ entonces $A = \begin{pmatrix} a \end{pmatrix}$ para cierto $a \in \mathbb{R}$, que es no nulo puesto que A es inversible. Entonces $\det(A) = \det(A_1) \neq 0$.

Sea $n > 1$. Como A es inversible entonces $\det(A) = \det(A_n) \neq 0$. Sea $A = LU$ una factorización LU de A. Descomponemos a L y U en bloques del siguiente modo,

$$L = \left(\begin{array}{c|c} L_{n-1} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ \hline v^t & 1 \end{array}\right) \qquad U = \left(\begin{array}{c|c} U_{n-1} & u \\ \hline 0 \cdots 0 & x \end{array}\right)$$

con $L_{n-1}, U_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$, $v, u \in \mathbb{R}^{n-1}$ y $x \in \mathbb{R}$. Entonces

$$A = LU = \left(\begin{array}{c|c} L_{n-1}U_{n-1} & L_{n-1}u \\ \hline v^t U_{n-1} & v^t u + x \end{array}\right)$$

Por lo tanto $A_{n-1} = L_{n-1}U_{n-1}$, es decir que esta submatriz tiene factorización LU. Veamos que, además, es inversible. Por un lado, como L_{n-1} es triangular inferior con unos en la diagonal, entonces es inversible. Por otro lado, como A y L son inversibles, entonces U es inversible. Luego $0 \neq \det U = \det(U_{n-1}) \det(x)$, con lo cual $\det(U_{n-1}) \neq 0$, es decir, U_{n-1} es inversible. Por ende L_{n-1} y U_{n-1} son inversibles, y en consecuencia A_{n-1} también lo es.

Todo esto hace que valga la hipótesis inductiva sobre A_{n-1}, y por lo tanto sus $n-1$ menores principales son no nulos. Pero estos menores principales son exactamente $\det(A_1), \cdots, \det(A_{n-1})$.

($\Leftarrow$) Inducción en $n \in \mathbb{N}$.

Si $n = 1$ la factorización LU de A es trivial.

Sea $n > 1$. Por hipótesis, la submatriz A_{n-1} tiene todos sus menores principales no nulos. Entonces, por hipótesis inductiva, admite una factorización LU que la escribimos $A_{n-1} = L_{n-1}U_{n-1}$. Descomponemos A en bloques del siguiente modo

$$A = \left(\begin{array}{c|c} L_{n-1}U_{n-1} & u \\ \hline v^t & a \end{array}\right)$$

Definimos los vectores $\alpha, \beta \in \mathbb{R}^{n-1}$ como

$$\alpha = (L_{n-1})^{-1} u$$
$$\beta = (U_{n-1}^t)^{-1} v$$

Esto es posible gracias a que las dimensiones son adecuadas y las matrices L_{n-1} y U_{n-1}^t son inversibles, por ser A_{n-1} inversible. Finalmente definimos el número real

$$x = a - \beta^t \alpha$$

A partir de estos elementos vamos a obtener la factorización LU de A. Sean

$$L = \left(\begin{array}{c|c} L_{n-1} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ \hline \beta^t & 1 \end{array}\right) \qquad U = \left(\begin{array}{c|c} U_{n-1} & \alpha \\ \hline 0 \cdots 0 & x \end{array}\right)$$

Haciendo la multiplicación es fácil ver que $A = LU$, que es lo que queríamos probar.

□

3.8. Matrices simétricas definidas positivas y factorización de Cholesky

3.8.1. Matrices simétricas y definidas positivas

Definición 3.4. Una matriz $A \in \mathbb{R}^{n\times n}$ se dice definida positiva si $x^tAx > 0$ para todo $x \neq 0$.

Proposición 3.6. *Si $A \in \mathbb{R}^{n\times n}$ es definida positiva entonces es inversible.*

Demostración. Veamos el contrarrecíproco. Si A no es inversible entonces existe $x \in \mathbb{R}^n$ no nulo tal que $Ax = 0$. Entonces $x^tAx = 0$, y como $x \neq 0$ esto implica que A no es definida positiva. □

Proposición 3.7. *Si $A \in \mathbb{R}^{n\times n}$ es definida positiva entonces sus submatrices principales son definidas positivas e inversibles.*

Demostración. Sea $1 \leq k \leq n$ cualquiera. Sea A_k la submatriz A formada por sus primeras k filas y k columnas, queremos ver que A_k es definida positiva e inversible. Si $k = n$ no hay nada que ver. Supongamos $k < n$. Sea $x_k \in \mathbb{R}^k$ no nulo. A partir de este vector construimos un nuevo vector $x \in \mathbb{R}^n$ agregando ceros, es decir,

$$x = \begin{pmatrix} x_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Multiplicando por bloques se puede ver que $x^tAx = x_k^tAx_k$. Como $x_k \neq 0$ entonces $x \neq 0$, y como A es definida positiva entonces $x^tAx > 0$. Luego $x_k^tA_kx_k > 0$. Dado que x_k es cualquiera, concluimos que A_k es definida positiva, como queríamos.

La inversibilidad de A_k se deduce inmediatamente de la proposición anterior.

□

Corolario 3.1. *Si $A \in \mathbb{R}^{n\times n}$ es definida positiva entonces admite factorización LU y es única.*

Demostración. Por la última proposición, todas las submatrices principales son inversibles. Pero vimos que toda matriz que cumple esto, admite factorización LU. Además por ser A inversible, la escritura es única. □

Definición 3.5. Una matriz $A \in \mathbb{R}^{n\times n}$ se dice simétrica si $A = A^t$.

Proposición 3.8. *Sea $A \in \mathbb{R}^{n\times n}$ inversible y simétrica, que admite factorización LU. Entonces existen matrices $L \in \mathbb{R}^{n\times n}$ triangular inferior con unos en la diagonal y $D \in \mathbb{R}^{n\times n}$ diagonal tal que $A = LDL^t$. Esta escritura se conoce como factorización LDL.*

Demostración. Sea $A = LU$ una factorización LU. Como A es simétrica entonces $A = A^t \Rightarrow LU = (LU)^t = U^tL^t$. Por definición L es inversible. Como A es inversible, U resulta inversible. Entonces $(U^t)^{-1}L = L^tU^{-1}$. Como U es triangular superior, entonces U^t es triangular inferior y $(U^t)^{-1}$ también. Análogamente, como L es triangular inferior, L^t es triangular superior. Además U^{-1} es triangular superior. Luego $(U^t)^{-1}L$ es triangular inferior y L^tU^{-1} es triangular superior, con lo cual ambas matrices resultan ser diagonales. Sea $L^tU^{-1} = D$ diagonal. Como L^t y U^{-1} son inversibles, entonces D también lo es. Luego $D^{-1}L^t = U$, y reemplazando en la factorización original queda $A = LU = LD^{-1}L^t$. Como D^{-1} también es diagonal, esta es la escritura buscada. □

Corolario 3.2. *Si $A \in \mathbb{R}^{n\times n}$ es simétrica y tiene todos sus menores principales no nulos, entonces admite factorización LDL.*

Demostración. Por la Proposición 3.5, A admite factorización LU. Además es inversible. Luego, por la Proposición 3.8, A admite factorización LDL. □

3.8.2. Factorización de Cholesky

Definición 3.6. Sea $A \in \mathbb{R}^{n\times n}$. Se llama factorización de Cholesky de A a una escritura de la forma $A = LL^t$ con L una matriz triangular inferior con elementos en la diagonal positivos.

Proposición 3.9. *Sea $A \in \mathbb{R}^{n\times n}$. Entonces A es simétrica definida positiva si y sólo si admite factorización de Cholesky.*

Demostración. ($\Rightarrow$) Como A es definida positiva entonces, por la Proposición 3.7, todos los menores principales de A son no nulos. Como además A es simétrica, por el Corolario 3.2, A tiene factorización LDL. Sea $A = LDL^t$ una tal escritura.

Veamos que todos los elementos de la diagonal de D son positivos. Sea $1 \leq i \leq n$ arbitrario. Como L^t es inversible, existe $x \in \mathbb{R}^n$ tal que $L^t x = e_i$, y que necesariamente es no nulo. Como $x \neq 0$ entonces $x^t Ax > 0$. Luego

$$0 < x^t Ax = x^t(LDL^t)x = (x^t L)D(L^t x) = e_i^t De_i = D_{ii}$$

En definitiva $D_{ii} > 0$ y como i es cualquiera, todos los elementos de la diagonal de D son positivos. Definimos

$$\sqrt{D} = \begin{pmatrix} \sqrt{D_{11}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{D_{nn}} \end{pmatrix}$$

Notemos que $D = \sqrt{D}\sqrt{D}$, con lo cual $A = LDL^t = L\sqrt{D}\sqrt{D}L^t = L\sqrt{D}\sqrt{D}^t L^t = (L\sqrt{D})(L\sqrt{D})^t$. Pero $L\sqrt{D}$ sigue siendo triangular inferior y, más aún, los elementos de su diagonal son positivos, pues tanto los de L como los de $\sqrt{D}$ lo son. Así llegamos a la factorización deseada.

($\Leftarrow$) Sea $A = LL^t$ una factorización de Cholesky. Sea $x \neq 0$, queremos ver que $x^t Ax > 0$. Se tiene que $x^t Ax = x^t LL^t x = (L^t x)^t(L^t x) = \|L^t x\|_2^2$ y esta norma es estrictamente mayor que cero, ya que como L^t es triangular con diagonal no nula y x es no nulo, entonces $L^t x \neq 0$.

□

Supongamos que A tiene factorización de Cholesky, de modo tal que podemos escribir

$$A = \begin{pmatrix} a_{11} & \cdots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} \ell_{11} & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \cdots & \ell_{nn} \end{pmatrix} \begin{pmatrix} \ell_{11} & \ell_{21} & \cdots & \ell_{n1} \\ 0 & \ell_{22} & \cdots & \ell_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \ell_{nn} \end{pmatrix} = LL^t$$

con $\ell_{ii} > 0$ para todo i. Para calcular el coeficiente de la fila i y columna j ($j \leq i$), computamos el producto de la fila i de L contra la columna j de L^t:

$$a_{ij} = \begin{pmatrix} \ell_{i1} & \cdots & \ell_{ij} & \cdots & \ell_{ii} & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \ell_{j1} \\ \vdots \\ \ell_{jj} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \ell_{i1}\ell_{j1} + \cdots + \ell_{ij}\ell_{jj} = \sum_{k=1}^{j} \ell_{ik}\ell_{jk}$$

Usando esta ecuación se deduce que:

- Si $j = i = 1$ entonces

$$a_{11} = \ell_{11}^2 \Leftrightarrow \ell_{11} = \sqrt{a_{11}}$$

- Si $j = 1 < i$ entonces

$$a_{i1} = \ell_{i1}\ell_{11} \Leftrightarrow \ell_{i1} = \frac{a_{i1}}{\ell_{11}}$$

- Si $1 < j = i$ entonces

$$a_{jj} = \ell_{j1}^2 + \cdots + \ell_{jj}^2 \Leftrightarrow \ell_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} \ell_{jk}^2}$$

- Si $1 < j < i$ entonces

$$a_{ij} = \ell_{i1}\ell_{j1} + \cdots + \ell_{ij}\ell_{jj} \Leftrightarrow \ell_{ij} = \frac{1}{\ell_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} \ell_{ik}\ell_{jk} \right)$$

Para calcular cada una de las entradas de la descomposición procedemos por columnas, primero calculando los coeficientes ℓ_{i1}, luego los ℓ_{i2} y así sucesivamente. Se puede ver que de esta forma respetamos las dependencias entre las ecuaciones anteriores.

Esto nos da un algoritmo para computar la matriz L de la factorización. Más aún, como cada ℓ_{ij} está unívocamente determinado, hemos probado que la factorización de Cholesky es única.

Algorithm 3: Factorización de Cholesky

```
ℓ_11 = √a_11;
for i = 2 to n do
    ℓ_i1 = a_i1 / ℓ_11;
end
for j = 2 to n do
    ℓ_jj = √(a_jj − Σ_{k=1}^{j−1} ℓ_jk^2);
    for i = j + 1 to n do
        ℓ_ij = (1/ℓ_jj) (a_ij − Σ_{k=1}^{j−1} ℓ_ik ℓ_jk);
    end
end
```

La complejidad del algoritmo es cúbica, que es la misma que la de LU. Sin embargo, en términos prácticos el cómputo de la factorización de Cholesky resulta ser el doble de rápido que el de LU. Además, al igual que LU, permite resolver eficientemente el problema de la resolución sucesiva de sistemas de misma matriz A.

3.9. Estabilidad numérica de Cholesky

A diferencia de LU, Cholesky no depende de pivoteos, manteniendo así el error acotado. El único problema que puede sufrir esta escritura es la presencia del cómputo de raíces cuadradas. Si bien los números a los que se les toma raíz son siempre positivos, si éstos son suficientemente pequeños se pueden tornar negativos a lo largo del cómputo, debido al error de redondeo, siendo imposible continuar con el algoritmo. Pese a que ésto sólo sucede en matrices muy mal condicionadas, para evitarlo puede optarse por sumarle a la matriz A otra matriz diagonal con entradas de valor absoluto chico, y así reforzar su condición de definida positiva. La desventaja en este caso es que se pierde precisión en el resultado.

4. Normas

4.1. Definiciones

Definición 4.1. Una función $f : \mathbb{R}^n \to \mathbb{R}$ es una norma vectorial si cumple:

1. $f(x) > 0$ para todo $x \neq 0$ y $f(0) = 0$.
2. $f(\lambda x) = |\lambda| f(x)$.
3. $f(x + y) \leq f(x) + f(y)$.

Algunos ejemplos de normas vectoriales son:

- $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
- $\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$
- $\|x\|_\infty = \max\limits_{1 \leq i \leq n} |x_i|$

El concepto se extiende naturalmente para matrices.

Definición 4.2. Una función $F : \mathbb{R}^{n \times n} \to \mathbb{R}$ es una norma matricial si cumple:

1. $F(A) > 0$ para toda $A \neq 0$ y $F(0) = 0$.
2. $F(\lambda A) = |\lambda| F(A)$.
3. $F(A + B) \leq F(A) + F(B)$.

Un ejemplo de norma matricial es la norma de Frobenius, una extensión aparentemente natural de la norma 2:

- $\|A\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^2\right)^{1/2}$

Una propiedad no necesaria, aunque deseable, es la de submultiplicatividad.

Definición 4.3. Si $F : \mathbb{R}^{n \times n} \to \mathbb{R}$ es una norma matricial, F se dice submultiplicativa si

$$F(AB) \leq F(A)F(B)$$

para todas $A, B \in \mathbb{R}^{n \times n}$.

Otra propiedad notable de las normas es la de consistencia.

Definición 4.4. Si $F : \mathbb{R}^{n \times n} \to \mathbb{R}$ es una norma matricial y $f : \mathbb{R}^n \to \mathbb{R}$ es una norma vectorial, F se dice consistente con f si

$$f(Ax) \leq F(A)f(x)$$

para todos $A \in \mathbb{R}^{n \times n}$, $x \in \mathbb{R}^n$.

4.2. Normas inducidas

Definición 4.5. Sea $f : \mathbb{R}^n \to \mathbb{R}$ una norma vectorial. Definimos $F : \mathbb{R}^{n \times n} \to \mathbb{R}$ como

$$F(A) = \max_{x : f(x) = 1} f(Ax)$$

F resulta una norma matricial y se la llama norma inducida por f.

A partir de ahora simbolizaremos tanto las normas vectoriales como sus respectivas normas matriciales inducidas mediante $\|\cdot\|$, y el significado quedará claro por el contexto.

Las normas inducidas son de especial interés por lo siguiente.

Proposición 4.1. *Si $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ es una norma matricial inducida, entonces $\|\cdot\|$ es submultiplicativa y consistente, es decir que*

1. $\|AB\| \leq \|A\| \, \|B\|$
2. $\|Ax\| \leq \|A\| \, \|x\|$

Demostración. 1. Sea $x_0 \in \mathbb{R}^n$ unitario, que realiza la norma de AB. Entonces

$$\|AB\| = \|ABx_0\| = \left\| A \frac{Bx_0}{\|Bx_0\|} \right\| \|Bx_0\|$$

Como $\frac{Bx_0}{\|Bx_0\|}$ es un vector unitario, entonces $\left\| A \frac{Bx_0}{\|Bx_0\|} \right\| \leq \|A\|$. Análogamente, $\|Bx_0\| \leq \|B\|$. Finalmente, como todos los factores son positivos, $\|AB\| \leq \|A\| \, \|B\|$.

2. Sea $x \in \mathbb{R}^n$ cualquiera. Entonces $\|A\| \geq \left\| A \frac{x}{\|x\|} \right\| = \frac{\|Ax\|}{\|x\|}$ pues $\frac{x}{\|x\|}$ es unitario. Luego $\|Ax\| \leq \|A\| \, \|x\|$.

□

4.3. Normas matriciales clásicas

Proposición 4.2. *Sea $A \in \mathbb{R}^{n\times n}$. Entonces*

$$\|A\|_1 = \max_{1\leq j\leq n} \|col_j(A)\|_1$$

Demostración. ($\geq$) Sea $1 \leq j \leq n$ arbitrario, entonces

$$\|A\|_1 = \max_{x:\|x\|_1=1} \|Ax\|_1 \geq \|Ae_j\|_1 = \|col_j(A)\|_1$$

Como j es cualquiera, resulta que $\|A\|_1 \geq \max_{1\leq j\leq n} \|col_j(A)\|_1$.

($\leq$)

$$
\begin{aligned}
\|A\|_1 &= \max_{x:\|x\|_1=1} \|Ax\|_1 \\
&= \max_{x:\|x\|_1=1} \|(fil_1(A)x, \cdots, fil_n(A)x)\|_1 \\
&= \max_{x:\|x\|_1=1} \sum_{i=1}^{n} |fil_i(A)x| \\
&= \max_{x:\|x\|_1=1} \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \\
&\leq \max_{x:\|x\|_1=1} \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}||x_j| \\
&= \max_{x:\|x\|_1=1} \sum_{j=1}^{n} |x_j| \left(\sum_{i=1}^{n} |a_{ij}| \right) \\
&\leq \max_{x:\|x\|_1=1} \sum_{j=1}^{n} |x_j| \left(\max_{1\leq k\leq n} \sum_{i=1}^{n} |a_{ik}| \right) \\
&= \max_{x:\|x\|_1=1} \left(\max_{1\leq k\leq n} \sum_{i=1}^{n} |a_{ik}| \right) \left(\sum_{j=1}^{n} |x_j| \right) \\
&= \max_{x:\|x\|_1=1} \left(\max_{1\leq k\leq n} \|col_k(A)\|_1 \right) \|x\|_1 \\
&= \max_{1\leq j\leq n} \|col_j(A)\|_1
\end{aligned}
$$

□

Proposición 4.3. *Sea $A \in \mathbb{R}^{n\times n}$. Entonces*

$$\|A\|_\infty = \max_{1\leq i\leq n} \|fil_i(A)\|_1$$

Demostración. Si $A = 0$ no hay nada que ver. Supongamos $A \neq 0$.

($\geq$) Sea i_0 la fila de A con norma 1 máxima. Definimos el vector $\alpha \in \mathbb{R}^n$ de modo tal que $\alpha_j = \text{sgn}(A_{i_0 j})$ para todo $j = 1, \cdots, n$. Entonces

$$fil_{i_0}(A)\alpha = \sum_{j=1}^{n} A_{i_0 j}\alpha_j = \sum_{j=1}^{n} \text{sgn}(A_{i_0 j})A_{i_0 j} = \sum_{j=1}^{n} |A_{i_0 j}| = \|fil_{i_0}(A)\|_1$$

y si i es una fila cualquiera de A entonces

$$|fil_i(A)\alpha| = \left| \sum_{j=1}^{n} A_{ij}\alpha_j \right| \leq \sum_{j=1}^{n} |A_{ij}||\alpha_j| \leq \sum_{j=1}^{n} |A_{ij}| = \|fil_i(A)\|_1$$

Luego, como $\|fil_i(A)\|_1 \leq \|fil_{i_0}(A)\|_1$ para todo i, resulta que

$$\max_{1\leq i\leq n} |fil_i(A)\alpha| = \|fil_{i_0}(A)\|_1 = \max_{1\leq i\leq n} \|fil_i(A)\|_1$$

Por otro lado, como $A \neq 0$ entonces $fil_{i_0}(A) \neq 0$ con lo cual α tiene alguna componente no nula. Como todas las coordenadas de α son 0, 1 ó -1, entonces $\|v\|_\infty = 1$. Luego

$$\|A\|_\infty = \max_{x:\|x\|_\infty=1} \|Ax\|_\infty \geq \|A\alpha\|_\infty = \max_{1\leq i\leq n} |fil_i(A)\alpha| = \max_{1\leq i\leq n} \|fil_i(A)\|_1$$

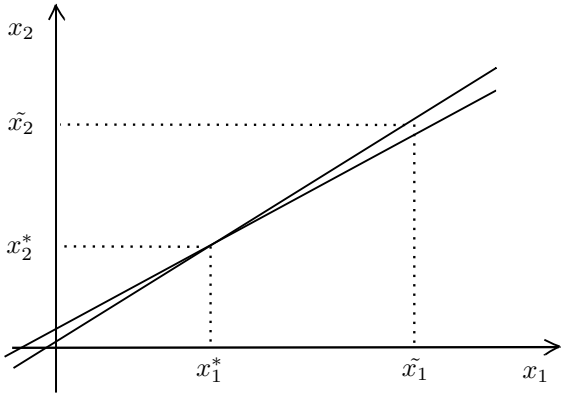
($\leq$)

$$\begin{aligned}
\|A\|_\infty &= \max_{x:\|x\|_\infty=1} \|Ax\|_\infty \\
&= \max_{x:\|x\|_\infty=1} \max_{1\leq i\leq n} |fil_i(A)x| \\
&= \max_{x:\|x\|_\infty=1} \max_{1\leq i\leq n} \left|\sum_{j=1}^{n} a_{ij}x_j\right| \\
&\leq \max_{x:\|x\|_\infty=1} \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}||x_j| \\
&\leq \max_{x:\|x\|_\infty=1} \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|\,\|x\|_\infty \\
&= \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \\
&= \max_{1\leq i\leq n} \|fil_i(A)\|_1
\end{aligned}$$

□

4.4. Estabilidad de un sistema y número de condición

Consideremos un sistema lineal $Ax = b$ de 2×2. Geométricamente se representa mediante dos rectas en el plano. Supongamos que estas rectas son *casi paralelas*.



Este sistema tiene solución única $x^* = \begin{pmatrix} x_1^* \\ x_2^* \end{pmatrix}$. Al resolverlo numéricamente encontramos una aproximación de la solución dada por $\tilde{x} = \begin{pmatrix} \tilde{x_1} \\ \tilde{x_2} \end{pmatrix}$ que, debido a la proximidad entre rectas, es tal que $A\tilde{x} = \tilde{b}$ con $\tilde{b}$ parecido a b. Sin embargo, pese a este parecido, $\tilde{x}$ está muy lejos de la solución exacta x^*.

En este caso, en que una pequeña variación respecto del resultado exacto significa una gran imprecisión en la aproximación obtenida, se dice que el sistema está mal condicionado.

La siguiente proposición formaliza esta idea intuitiva.

Proposición 4.4. *Sea $A \in \mathbb{R}^{n\times n}$ inversible. Sea x^* solución de $Ax = b$. Sea $\tilde{x}$ solución de $Ax = \tilde{b}$. Si $\|\cdot\|$ es una norma inducida cualquiera, entonces*

1. $\|x^* - \tilde{x}\| \leq \left\|b - \tilde{b}\right\| \left\|A^{-1}\right\|$
2. $\frac{\|x^*-\tilde{x}\|}{\|x^*\|} \leq \|A\| \left\|A^{-1}\right\| \frac{\left\|b-\tilde{b}\right\|}{\|b\|}$

Demostración. 1. Como A es inversible y $\|\cdot\|$ consistente,

$$\|x^* - \tilde{x}\| = \left\|A^{-1}b - A^{-1}\tilde{b}\right\| = \left\|A^{-1}(b - \tilde{b})\right\| \leq \left\|A^{-1}\right\| \left\|b - \tilde{b}\right\|$$

2. Usando la desigualdad anterior,

$$\|x^* - \tilde{x}\| \leq \left\|A^{-1}\right\| \left\|b - \tilde{b}\right\| = \frac{\|Ax^*\|}{\|Ax^*\|} \left\|A^{-1}\right\| \left\|b - \tilde{b}\right\| \leq \|A\| \|x^*\| \left\|A^{-1}\right\| \frac{\left\|b - \tilde{b}\right\|}{\|b\|}$$

Para terminar basta dividir la desigualdad por $\|x^*\|$.

□

Definición 4.6. Si $A \in \mathbb{R}^{n\times n}$ es inversible y $\|\cdot\|$ es una norma inducida, definimos el número de condición de A como

$$\kappa(A) = \|A\| \left\|A^{-1}\right\|$$

Con esta definición, podemos reescribir el resultado anterior como

Corolario 4.1.

$$\frac{\|x^* - \tilde{x}\|}{\|x^*\|} \leq \kappa(A) \frac{\left\|b - \tilde{b}\right\|}{\|b\|}$$

Este resultado nos dice que cuanto más chico sea $\kappa(A)$ entonces bajo un pequeño error relativo $\frac{\left\|b-\tilde{b}\right\|}{\|b\|}$ (es decir, que el resultado obtenido es similar al buscado) podremos asegurar con mayor certeza un pequeño error relativo $\frac{\|x^*-\tilde{x}\|}{\|x^*\|}$ (es decir, que la aproximación es precisa).

Observación 4.1. $\kappa(A) \geq \|I\| = 1$

5. Factorización QR

5.1. Matrices ortogonales

Definición 5.1. Una matriz $Q \in \mathbb{R}^{n\times n}$ se dice ortogonal si $Q^{-1} = Q^t$.

Observación 5.1. Si Q es ortogonal entonces $Q^t = Q^{-1}$ también, pues $((Q^t)^{-1})^t = ((Q^t)^t)^{-1} = Q^{-1} = Q^t$.

En lo que sigue, Q es una matriz ortogonal de $n \times n$.

Lema 5.1. *Q preserva norma 2, i. e.,* $\|Qx\|_2 = \|x\|_2$.

Lema 5.2. $\|Q\|_2 = 1$.

Corolario 5.1. $\kappa_2(Q) = 1$.

Esto nos dice que las matrices ortogonales son estables, lo cual se condice con el hecho de que preservan norma 2, es decir, que no deforman vectores, haciendo que el error de las componentes escalares no sea amplificado. Este motivo hace a las matrices ortogonales atractivas desde el punto de vista numérico. Además, el cómputo de su inversa no requiere más que una transposición de elementos, y por lo tanto no introduce error.

Definición 5.2. Sea $A \in \mathbb{R}^{n\times n}$. Se llama factorización QR de A a una escritura de la forma $A = QR$ con $Q \in \mathbb{R}^{n\times n}$ una matriz ortogonal y $R \in \mathbb{R}^{n\times n}$ triangular superior.

Notemos que de tener la factorización QR, tenemos la cadena de equivalencias $Ax = b \Leftrightarrow QRx = b \Leftrightarrow Rx = Q^t b$ y este último sistema se puede resolver mediante backward substitution, con error mínimo.

5.2. Método de rotaciones (Givens)

Dado $\theta \in [0, \frac{\pi}{2}]$ queremos encontrar una transformación lineal $W : \mathbb{R}^2 \to \mathbb{R}^2$ que rote en un ángulo θ todo vector en el plano. Es fácil ver en forma geométrica que W actúa del siguiente modo sobre la base canónica,

$$W\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}\cos\theta\\-\sin\theta\end{pmatrix}$$
$$W\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}\sin\theta\\\cos\theta\end{pmatrix}$$

Entonces

$$W = \begin{pmatrix}\cos\theta & \sin\theta\\-\sin\theta & \cos\theta\end{pmatrix}$$

Se puede probar que esto vale para cualquier $\theta \in \mathbb{R}$ (separando en casos, según el cuadrante en que se encuentre θ). De este modo quedan caracterizadas todas las rotaciones en $\mathbb{R}^2$. Notemos que W es ortogonal.

Supongamos que dado $v = \begin{pmatrix}v_1\\v_2\end{pmatrix} \in \mathbb{R}^2$ no nulo, queremos encontrar una rotación $W : \mathbb{R}^2 \to \mathbb{R}^2$ que rote v sobre el eje positivo de las abscisas, es decir, tal que $Wv = \begin{pmatrix}\|v\|_2\\0\end{pmatrix}$. Usando la forma de una rotación en $\mathbb{R}^2$, tenemos que

$$Wv = \begin{pmatrix}\|v\|_2\\0\end{pmatrix} \Leftrightarrow \begin{pmatrix}\cos\theta & \sin\theta\\-\sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}v_1\\v_2\end{pmatrix} = \begin{pmatrix}\|v\|_2\\0\end{pmatrix} \Leftrightarrow \begin{cases}v_1\cos\theta + v_2\sin\theta = \|v\|_2\\-v_1\sin\theta + v_2\cos\theta = 0\end{cases}$$

Tomemos θ tal que $\cos\theta = \frac{v_1}{\|v\|_2}$ y $\sin\theta = \frac{v_2}{\|v\|_2}$. Este θ existe y es único ($\cos\theta$ y $\sin\theta$ se pueden pensar como las coordenadas x e y respectivamente de un punto sobre la circunferencia unitaria), y se puede verificar que así tomado satisface las ecuaciones previas. De este modo, hemos probado lo siguiente,

Proposición 5.1. *Sea* $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \in \mathbb{R}^2$ *un vector no nulo. La única rotación* $W : \mathbb{R}^2 \to \mathbb{R}^2$ *que cumple* $Wv = \begin{pmatrix} \|v\|_2 \\ 0 \end{pmatrix}$ *es*

$$W = \begin{pmatrix} \frac{v_1}{\|v\|_2} & \frac{v_2}{\|v\|_2} \\ -\frac{v_2}{\|v\|_2} & \frac{v_1}{\|v\|_2} \end{pmatrix}$$

Sea ahora $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ cualquiera y pongamos $v = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}$. Si $v = 0$ entonces A es triangular superior y por lo tanto $I_2 A$ es una factorización QR de A. Si $v \neq 0$ construimos, usando la proposición anterior, una rotación W tal que $Wv = \begin{pmatrix} \|v\|_2 \\ 0 \end{pmatrix}$, con lo cual

$$WA = \begin{pmatrix} \|v\|_2 & * \\ 0 & * \end{pmatrix} = R$$

Luego $WA = R$ con R triangular superior y en definitiva $A = W^t R$ es una factorización QR de A.

Proposición 5.2. *Si* $A \in \mathbb{R}^{2\times 2}$ *entonces* A *tiene factorización QR.*

5.2.1. Extensión al caso general

Definimos para cada $1 \leq i < j \leq n$ y $v \in \mathbb{R}^2$ no nulo, las matrices

$$W_{ij}(v) = \begin{pmatrix} I_{(i-1)} & & & & & & \\ & \frac{v_1}{\|v\|_2} & 0 & \cdots & 0 & \frac{v_2}{\|v\|_2} & \\ & 0 & 1 & \cdots & 0 & 0 & \\ & \vdots & \vdots & \ddots & \vdots & \vdots & \\ & 0 & 0 & \cdots & 1 & 0 & \\ & -\frac{v_2}{\|v\|_2} & 0 & \cdots & 0 & \frac{v_1}{\|v\|_2} & \\ & & & & & & I_{(n-j)} \end{pmatrix}$$

Los espacios en blanco representan ceros. Es fácil ver que estas matrices siempre son ortogonales.

Sea $A = (a_{ij})_{i,j} \in \mathbb{R}^{n\times n}$. Vamos a ir multiplicando sucesivamente a izquierda la matriz A por matrices ortogonales W_{ij}, de modo tal de colocar ceros en la primer columna, luego en la segunda, y así continuando. Escribamos $A^{(k)} = (a_{ij}^{(k)})_{i,j}$ a la matriz que se obtiene luego de k multiplicaciones sobre A. A cada una de estas matrices $A^{(k)}$ la llamamos *matriz transitoria.*

`Paso 1:` Definimos $v = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}$ y sea $W_{12} = W_{12}(v)$. Entonces

$$W_{12}A = \begin{pmatrix} a_{11}^{(1)} & * & \cdots & * \\ 0 & * & \cdots & * \\ a_{31}^{(1)} & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ a_{n1}^{(1)} & * & \cdots & * \end{pmatrix} = A^{(1)}$$

`Paso 2:` Definimos $v = \begin{pmatrix} a_{11}^{(1)} \\ a_{31}^{(1)} \end{pmatrix}$ y sea $W_{13} = W_{13}(v)$. Entonces

$$W_{13}W_{12}A = \begin{pmatrix} a_{11}^{(2)} & * & \cdots & * \\ 0 & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ a_{n1}^{(2)} & * & \cdots & * \end{pmatrix} = A^{(2)}$$

Continuando así, llegaremos al

`Paso n - 1:` Definimos $v = \begin{pmatrix} a_{11}^{(n-2)} \\ a_{n1}^{(n-2)} \end{pmatrix}$ y sea $W_{1n} = W_{1n}(v)$. Entonces

$$W_{1n} \cdots W_{12} A = \begin{pmatrix} a_{11}^{(n-1)} & * & \cdots & * \\ 0 & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{pmatrix} = A^{(n-1)}$$

Si en alguno de los pasos resulta que $v = 0$ entonces no es necesario multiplicar por ninguna matriz para colocar un cero en una determinada posición, y podemos saltar al paso siguiente.

Análogamente, para poner ceros en la segunda columna definimos W_{2j}, $j = 3, \cdots, n$. En general, al poner ceros en la columna i definimos W_{ij}, $j = i+1, \cdots, n$. Notar que al multiplicar por las matrices W_{ij}, las columnas $1, \cdots, i-1$ no son modificadas.

El último paso será el

`Paso n(n - 1)/2:` Definimos $W_{(n-1)n}$ como se indica arriba, de modo tal que

$$W_{(n-1)n} \cdots W_{23} W_{1n} \cdots W_{12} A = R$$

con R triangular superior. Luego, como cada W_{ij} es ortogonal,

$$A = W_{12}^t \cdots W_{(n-1)n}^t R$$

y como producto de matrices ortogonales es una matriz ortogonal, definiendo $Q = W_{12}^t \cdots W_{(n-1)n}^t$, llegamos a que $A = QR$, que es la factorización QR de A.

Proposición 5.3. *Si $A \in \mathbb{R}^{n \times n}$ entonces A tiene factorización QR.*

5.2.2. Costo del algoritmo

El costo está dado por

$$\sum_{i=1}^{n-1} (\text{costo de colocar ceros en la columna } i)$$

Para colocar ceros en la columna i de la matriz transitoria la multiplicamos por $n - i$ matrices W_{ij}, $j = i+1, \cdots, n$. El costo de armar W_{ij} es $\mathcal{O}(1)$ suponiendo una representación óptima en el sentido espacial, pues sólo es necesario almacenar 4 elementos. Para computar el producto de W_{ij} por la matriz transitoria, sólo es necesario multiplicar las filas i y j de W_{ij}, pues el resto de las filas son vectores canónicos que dejan intactas las correspondientes filas de la matriz transitoria. Las filas i y j de W_{ij} sólo contienen dos entradas no nulas, que son las únicas que se multiplican y suman contra las $n - i + 1$ últimas columnas de la matriz transitoria (las únicas no nulas debajo de la diagonal).

Sumando estos costos, resultan en total

$$\mathcal{O}\left(\sum_{i=1}^{n-1} (n-i)(1 + 2 \cdot 2 \cdot (n-i+1)) \right) = \mathcal{O}(n^3)$$

Un análisis más fino permite ver que el método de rotaciones realiza aproximadamente $\frac{4}{3}n^3$ operaciones de punto flotante.

5.3. Método de reflexiones (Householder)

Dado un vector $u \in \mathbb{R}^n$, $\|u\|_2 = 1$, busco una transformación lineal $W : \mathbb{R}^n \to \mathbb{R}^n$ que refleje todo vector sobre el hiperplano H perpendicular a u. Notemos que una tal W cumple:

- $Wu = -u$
- $Wv = v$ para todo $v \in H$

Estas condiciones caracterizan unívocamente a W, porque indican cómo actúa sobre u y sobre una base de H, y esta información define a W sobre una base de $\mathbb{R}^n$. Calculemos entonces esta transformación lineal.

Consideremos $P = uu^t \in \mathbb{R}^{n\times n}$. Se tiene

$$Pu = uu^t u = u(u^t u) = u$$
$$Pv = uu^t v = u(u^t v) = 0$$

A partir de P, definimos $W = I - 2P$, que cumple

$$Wu = u - 2Pu = u - 2u = -u$$
$$Wv = v - 2Pv = v - 0 = v$$

Entonces $W = I - 2P = I - 2uu^t$ es la única transformación lineal que cumple lo pedido. Esta transformación lineal se llama reflexión sobre el hiperplano ortogonal a u. De este modo, hemos caracterizado todas las reflexiones en $\mathbb{R}^n$. Notemos que esta matriz W es ortogonal.

Supongamos ahora que tenemos dos vectores $v, w \in \mathbb{R}^n$ con misma norma 2. Queremos encontrar una reflexión W, que refleje v en w. En otras palabras, buscamos un vector $u \in \mathbb{R}^n$ tal que la reflexión sobre el hiperplano perpendicular a u mande v en w.

Proposición 5.4. *Sean $v, w \in \mathbb{R}^n$ tal que $\|v\|_2 = \|w\|_2$. Entonces existe una reflexión W tal que $Wv = w$.*

Demostración. Si $v = w = 0$ no hay nada que ver. Supongamos que son no nulos. Como $\|v\|_2 = \|w\|_2$, podemos tomar $u = \frac{v-w}{\|v-w\|_2}$ que es tal que $\|u\|_2 = 1$, y una cuenta indica que $W = I - 2uu^t$ es una reflexión que manda v en w. □

A partir de esto vamos a construir una factorización QR de una matriz $A = (a_{ij})_{i,j} \in \mathbb{R}^{n\times n}$. Vamos a ir multiplicando a la matriz A sucesivamente por matrices de reflexión, para poner ceros en la primera columna, luego en la segunda, y así sucesivamente. Como de costumbre, llamemos $A^{(k)}$ a la matriz A luego de k multiplicaciones.

`Paso 1:` Definimos v como la primera columna de A, es decir, $v = \begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix} \in \mathbb{R}^n$, y sea $w = \begin{pmatrix} \|v\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^n$. Entonces $\|v\|_2 = \|w\|_2$ con lo cual existe una reflexión $W_1 = I - 2u_1u_1^t \in \mathbb{R}^{n\times n}$ tal que $W_1 v = w$, que cumple

$$W_1 A = \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{pmatrix} = A^{(1)}$$

`Paso 2:` Definimos v como la primer columna de la submatriz de $A^{(1)}$ que se obtiene sacando su primera fila y columna, es decir, $v = \begin{pmatrix} a_{22}^{(1)} \\ \vdots \\ a_{n2}^{(1)} \end{pmatrix} \in \mathbb{R}^{n-1}$, y sea $w = \begin{pmatrix} \|v\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{n-1}$. Entonces existe $\overline{W}_2 = I - 2\overline{u}_2\overline{u}_2^t \in \mathbb{R}^{(n-1)\times(n-1)}$ reflexión tal que $\overline{W}_2 v = w$. Si llamamos $u_2 = \begin{pmatrix} 0 \\ \overline{u}_2 \end{pmatrix} \in \mathbb{R}^n$ y $W_2 = I - 2u_2u_2^t = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \overline{W}_2 & \\ 0 & & & \end{pmatrix}$, entonces

$$W_2W_1A = W_2A^{(1)} = \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ 0 & 0 & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & * \end{pmatrix} = A^{(2)}$$

El paso $k < n$ es análogo,

`Paso k:` Definimos v como la primer columna de la submatriz de $A^{(k-1)}$ que se obtiene sacando sus primeras $k-1$ filas y columnas, es decir, $v = \begin{pmatrix} a_{kk}^{(k-1)} \\ \vdots \\ a_{nk}^{(k-1)} \end{pmatrix} \in \mathbb{R}^{n-k+1}$, y sea $w = \begin{pmatrix} \|v\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{n-k+1}$. Con estos, definimos la correspondiente reflexión $\overline{W_k}$ de $\mathbb{R}^{(n-k+1)\times(n-k+1)}$ con vector normal $\overline{u}_k$ y luego la extiendo a una reflexión W_k de $\mathbb{R}^{n\times n}$ definiendo $u_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \overline{u}_k \end{pmatrix}$. Entonces

$$W_k \cdots W_1 A = A^{(k)}$$

El último es el

`Paso n - 1:` Definimos W_{n-1} como se indica arriba, de modo tal que

$$W_{n-1} \cdots W_2W_1A = R$$

con R triangular superior. Como toda reflexión es ortogonal, entonces transponiendo $W_1, \cdots, W_{n-1}$ llegamos a la factorización QR.

5.3.1. Costo del algoritmo

Si computáramos el producto matricial en forma naïve $\mathcal{O}(n^3)$ el costo de $W_{n-1} \cdots W_1A$ sería $\mathcal{O}(n^4)$. Podemos usar la escritura $W_i = I - 2u_iu_i^t$ de las matrices ortogonales para computar el producto en forma más eficiente. Notemos que $W_1A = (I - 2u_1u_1^t)A = A - 2u_iu_i^tA$. Si asociamos en la forma $A - 2u_i(u_i^tA)$, se puede ver que el costo de cada una de las operaciones es $\mathcal{O}(n^2)$. En definitiva, podemos computar $W_1A = A^{(1)}$ en $\mathcal{O}(n^2)$. Lo mismo sucede para $W_2A^{(1)}$ y así siguiendo. En total son $n-1$ multiplicaciones, totalizando un costo $\mathcal{O}(n^3)$.

El método de reflexiones realiza aproximadamente $\frac{2}{3}n^3$ operaciones de punto flotante.

5.4. Observaciones finales

El método de las rotaciones es particularmente útil cuando la matriz A es rala. Como dicho método coloca ceros de a una posición a la vez, podemos evitar hacerlo para cada una de las entradas de A que contengan ceros. En contraste, el método de reflexiones coloca en cada paso ceros en toda una columna.

Por otra parte, a lo largo del desarrollo hemos pedido que la matriz A fuese cuadrada. Sin embargo esto no es necesario, y es posible adaptar cualquiera de los métodos anteriores para obtener una factorización QR de una matriz no necesariamente cuadrada. Se tiene la siguiente generalización,

Proposición 5.5. *Si $A \in \mathbb{R}^{m\times n}$ entonces existen $Q \in \mathbb{R}^{m\times m}$ ortogonal y $R \in \mathbb{R}^{m\times n}$ triangular superior, tal que $A = QR$.*

6. Métodos iterativos para resolución de sistemas lineales

6.1. Definiciones

Definición 6.1. Sea $A \in \mathbb{R}^{n\times n}$. El polinomio característico de A es

$$\chi_A(x) = \det(xI_n - A)$$

Definición 6.2. $\lambda \in \mathbb{C}$ se dice autovalor de $A \in \mathbb{R}^{n\times n}$ si existe un vector $v \in \mathbb{C}^n$ no nulo tal que $Av = \lambda v$. El vector v se llama autovector asociado al autovalor λ.

Proposición 6.1. *$\lambda \in \mathbb{C}$ es autovalor de A sí y sólo si λ es raíz de χ_A.*

Demostración.

$$\begin{aligned}\lambda \text{ es autovalor de } A &\Leftrightarrow \exists v \neq 0 \text{ tal que } Av = \lambda v\\ &\Leftrightarrow \exists v \neq 0 \text{ tal que } (\lambda I_n - A)v = 0\\ &\Leftrightarrow \lambda I_n - A \text{ no es inversible}\\ &\Leftrightarrow \det(\lambda I_n - A) = 0\\ &\Leftrightarrow \chi_A(\lambda) = 0\end{aligned}$$

□

Definición 6.3. Sea $A \in \mathbb{R}^{n\times n}$. El radio espectral de A es

$$\rho(A) = \max_{\lambda \text{ autovalor de } A} |\lambda|$$

Proposición 6.2. *$\rho(A) \leq \|A\|$ para toda norma matricial inducida compleja.*

Demostración. Sea $\lambda_0 \in \mathbb{C}$ un autovalor de A de módulo máximo. Sea $v_0 \in \mathbb{C}^n$ un autovector asociado a λ_0. Entonces

$$\|A\| \geq \left\| A\frac{v_0}{\|v_0\|} \right\| = \frac{\|Av_0\|}{\|v_0\|} = \frac{\|\lambda_0 v_0\|}{\|v_0\|} = \frac{|\lambda_0|\,\|v_0\|}{\|v_0\|} = |\lambda_0| = \rho(A)$$

□

Definición 6.4. Decimos que una matriz $A \in \mathbb{R}^{n\times n}$ es convergente si

$$\lim_{k\to\infty} (A^k)_{ij} = 0$$

para todo i, j.

Proposición 6.3. *A es convergente si y sólo si $\rho(A) < 1$.*

Proposición 6.4. *Si $\rho(A) < 1$ entonces $I - A$ es inversible y además*

$$\sum_{k=0}^{\infty} A^k = (I - A)^{-1}$$

6.2. Problema

Dado un sistema lineal, queremos construir una sucesión de vectores $\{x^{(k)}\}_{k\in\mathbb{N}_0}$ tal que $x^{(k)} \xrightarrow[k\to\infty]{} x^*$, donde x^* sea una solución del sistema.

6.3. Métodos exactos vs. métodos iterativos

Ya hemos visto cómo resolver sistemas lineales en forma directa y exacta. La pregunta lógica es ¿por qué buscar una solución iterativa al problema si ya tenemos una directa? Para sistemas de ecuaciones pequeños, los métodos iterativos resultan más lentos que los directos, pues demandan más tiempo para realizar las suficientes iteraciones de modo de aproximar con exactitud la solución. Sin embargo, en dos situaciones los métodos iterativos resultan una mejor opción:

- **Sistemas de ecuaciones ralos (la matriz del sistema presenta muchos ceros).** Aquí los métodos iterativos son más eficientes tanto en términos temporales como espaciales. Si la matriz del sistema es rala, los métodos iterativos son compatibles con la utilización de representaciones adecuadas que reduzcan el espacio y tiempo de las operaciones, mientras que los métodos directos no. Por ejemplo, la eliminación gaussiana opera con filas completas, haciendo desaparecer los ceros presentes inicialmente en una fila. Otro ejemplo es la factorización LU que, aunque A sea rala, no asegura que L y U sean ralas.
- **Sistemas de ecuaciones muy grandes.** Dado que la solución se aproxima mediante iteraciones sucesivas, la cantidad de iteraciones necesarias para obtener una aproximación tan buena como se desee depende de nuestro criterio. Para sistemas grandes, acotar la cantidad de iteraciones es un factor determinante en el costo temporal.

6.4. Método de Jacobi

Sean $A = (a_{ij})_{i,j} \in \mathbb{R}^{n\times n}$ y $b = (b_i)_i \in \mathbb{R}^n$ los componentes del sistema $Ax = b$ que queremos resolver. Vamos a suponer que $a_{ii} \neq 0$ para todo i. Fijemos $x^{(0)} = (x_1^{(0)}, \cdots, x_n^{(0)})$ cualquiera. Definimos $x^{(1)}$ del siguiente modo. Para cada $i = 1, \cdots, n$ tomamos la ecuación i-ésima del sistema

$$a_{i1}x_1 + \cdots + a_{in}x_n = b_i$$

Reemplazamos x_j por $x_j^{(0)}$ para cada $j \neq i$ y despejamos x_i para definir

$$x_i^{(1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j\neq i}^{n} a_{ij}x_j^{(0)}\right)$$

En general, para definir $x^{(k)}$ usamos $x^{(k-1)}$:

$$x_i^{(k)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j\neq i}^{n} a_{ij}x_j^{(k-1)}\right)$$

Algorithm 4: Método de Jacobi

```
Definir x^0;
for k = 1 to N do
    for i = 1 to n do
        x_i^(k) = (1/a_ii) (b_i − Σ_{j≠i}^n a_ij x_j^(k−1));
    end
end
```

Es claro que el costo de cada iteración del método es $\mathcal{O}(n^2)$.

Con este algoritmo podemos aprovechar los ceros que presente la matriz A. Supongamos que tenemos una representación adecuada para dicha matriz en la cual, en cada fila, sólo almacenamos las entradas no nulas. Con esta representación, en cada iteración de Jacobi, sólo necesitamos computar los productos $a_{ij}x_j^{(k-1)}$ cuando a_{ij} sea no nulo. Observar que como la matriz A no es modificada a lo largo del proceso, nunca destruimos la representación.

6.4.1. Forma matricial

Nos proponemos dar una forma completamente matricial del método. Escribamos

$$A = D - L - U$$

donde

$$D = \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{nn} \end{pmatrix}$$

es la matriz que consta sólo de los elementos de la diagonal de A,

$$L = \begin{pmatrix} 0 & & 0 \\ & \ddots & \\ -a_{ij} & & 0 \end{pmatrix}$$

es la matriz que consta de los elementos debajo de la diagonal de A negados, y

$$U = \begin{pmatrix} 0 & & -a_{ij} \\ & \ddots & \\ 0 & & 0 \end{pmatrix}$$

es la matriz que consta de los elementos encima de la diagonal de A negados. Por hipótesis, D es inversible, pues todos los elementos de la diagonal son no nulos. Se tiene $Ax = b \Leftrightarrow (D - L - U)x = b \Leftrightarrow Dx - (L + U)x = b \Leftrightarrow Dx = b + (L + U)x \Leftrightarrow x = D^{-1}b + D^{-1}(L + U)x$. Esta igualdad motiva la definición de la iteración

$$x^{(k)} = D^{-1}b + D^{-1}(L + U)x^{(k-1)}$$

que en caso de converger, lo hace a una solución de $Ax = b$.

Observación 6.1. Podemos pensar esto como un problema de punto fijo. Si definimos la función $g(x) = D^{-1}b + D^{-1}(L + U)x$, entonces estamos buscando un punto fijo de g, utilizando la iteración $x^{(k)} = g(x^{(k-1)})$.

Calculemos $x^{(k)}$ en términos de los elementos de A y B y veamos que coincide con el algoritmo de Jacobi. Se tiene

$$D^{-1}b = \begin{pmatrix} \frac{1}{a_{11}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{a_{nn}} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} \frac{b_1}{a_{11}} \\ \vdots \\ \frac{b_n}{a_{nn}} \end{pmatrix}$$

$$D^{-1}(L+U) = \underbrace{\begin{pmatrix} \frac{1}{a_{11}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{a_{nn}} \end{pmatrix}}_{\text{multiplica la fila } i \text{ por } 1/a_{ii}} \begin{pmatrix} 0 & & -a_{ij} \\ & \ddots & \\ -a_{ij} & & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \cdots & -\frac{a_{2n}}{a_{22}} \\ \vdots & & \ddots & \vdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \cdots & 0 \end{pmatrix}$$

Entonces

$$x^{(k)} = D^{-1}b + D^{-1}(L + U)x^{(k-1)} = \begin{pmatrix} \frac{1}{a_{11}} \left(b_1 - \sum_{j \neq 1}^{n} a_{1j} x_j^{(k-1)} \right) \\ \vdots \\ \frac{1}{a_{nn}} \left(b_n - \sum_{j \neq n}^{n} a_{nj} x_j^{(k-1)} \right) \end{pmatrix}$$

como queríamos ver.

6.5. Método de Gauss - Seidel

Construimos la iteración en forma análoga a Jacobi, con la única diferencia de que para calcular $x_i^{(k)}$ usamos las coordenadas $x_1^{(k)}, \cdots, x_{i-1}^{(k)}$ ya calculadas:

$$x_i^{(k)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \right)$$

Algorithm 5: Método de Gauss - Seidel

```
Definir x^(0);
for k = 1 to N do
    for i = 1 to n do
        x_i^(k) = (1/a_ii) (b_i − Σ_{j=1}^{i−1} a_ij x_j^(k) − Σ_{j=i+1}^{n} a_ij x_j^(k−1));
    end
end
```

6.5.1. Forma matricial

Se tiene $Ax = b \Leftrightarrow (D - L - U)x = b \Leftrightarrow (D - L)x - Ux = b \Leftrightarrow (D - L)x = b + Ux \Leftrightarrow x = (D - L)^{-1}b + (D - L)^{-1}Ux$. Esta última equivalencia vale debido a que $D - L$ es inversible, pues D es inversible. Definimos entonces la iteración

$$x^{(k)} = (D - L)^{-1}b + (D - L)^{-1}Ux^{(k-1)}$$

Veamos que esta definición coincide con el algoritmo de Gauss - Seidel. Vamos a despejar las componentes de $x^{(k)}$ a partir de la igualdad $(D - L)x^{(k)} = b + Ux^{(k-1)}$. Tenemos

$$b + Ux^{(k-1)} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} + \begin{pmatrix} 0 & & -a_{ij} \\ & \ddots & \\ 0 & & 0 \end{pmatrix} \begin{pmatrix} x_1^{(k-1)} \\ \vdots \\ x_n^{(k-1)} \end{pmatrix} = \begin{pmatrix} b_1 - \sum_{j=2}^{n} a_{1j} x_j^{(k-1)} \\ b_2 - \sum_{j=3}^{n} a_{2j} x_j^{(k-1)} \\ \vdots \\ b_n \end{pmatrix}$$

Entonces

$$(D - L)x^{(k)} = b + Ux^{(k-1)} \Leftrightarrow \begin{pmatrix} a_{11} & & 0 \\ \vdots & \ddots & \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} x^{(k)} = \begin{pmatrix} b_1 - \sum_{j=2}^{n} a_{1j} x_j^{(k-1)} \\ b_2 - \sum_{j=3}^{n} a_{2j} x_j^{(k-1)} \\ \vdots \\ b_n \end{pmatrix}$$

$$\Leftrightarrow x^{(k)} = \begin{pmatrix} \frac{1}{a_{11}} \left(b_1 - \sum_{j=2}^{n} a_{1j} x_j^{(k-1)} \right) \\ \frac{1}{a_{22}} \left(b_2 - a_{21} x_1^{(k)} - \sum_{j=3}^{n} a_{2j} x_j^{(k-1)} \right) \\ \vdots \\ \frac{1}{a_{nn}} \left(b_n - \sum_{j=1}^{n-1} a_{nj} x_j^{(k)} \right) \end{pmatrix}$$

que es lo que queríamos ver.

6.6. Análisis de convergencia

Hasta ahora propusimos dos iteraciones distintas, aunque nunca probamos que efectivamente convergieran a una solución. Observemos que tanto Jacobi como Gauss - Seidel, son iteraciones del tipo

$$x^{(k)} = Tx^{(k-1)} + c$$

donde $T \in \mathbb{R}^{n\times n}$ y $c \in \mathbb{R}^n$ están fijos. A continuación proveemos una condición necesaria y suficiente sobre estas matrices-coeficientes para asegurar la convergencia.

Proposición 6.5. *Consideremos el sistema lineal $x = Tx + c$ y sea x^* una solución del mismo. Sea $\{x^{(k)}\}_{k\in\mathbb{N}_0}$ una sucesión de vectores tal que*

$$x^{(k)} = Tx^{(k-1)} + c$$

para todo $k > 0$. Entonces $\{x^{(k)}\}_k$ converge a x^* para todo $x^{(0)}$ si y sólo si T es convergente.

Demostración. Sea $r_k = x^* - x^{(k)}$. Queremos probar que $r_k \xrightarrow[k\to\infty]{} 0$ para todo $x^{(0)}$ si y sólo si T converge.

($\Leftarrow$) Notemos que $r_k = x^* - x^{(k)} = (Tx^* + c) - (Tx^{(k-1)} + c) = Tx^* - Tx^{(k-1)} = T(x^* - x^{(k-1)}) = Tr_{k-1}$, con lo cual $r_k = T^k r_0 = T^k(x^* - x^{(0)})$. Pero entonces $r_k \xrightarrow[k\to\infty]{} 0 \Leftrightarrow T^k(x^* - x^{(0)}) \xrightarrow[k\to\infty]{} 0$. Es claro que si T converge vale lo anterior.

($\Rightarrow$) Recíprocamente, si $T^k(x^* - x^{(0)}) \xrightarrow[k\to\infty]{} 0$ para todo $x^{(0)}$, podemos elegir $x^{(0)}$ de modo tal que $x^* - x^{(0)} = e_i$ y por lo tanto $T^k(x^* - x^{(0)}) = T^k e_i = col_i(T^k) \xrightarrow[k\to\infty]{} 0$, y como i es cualquiera resulta que T converge. □

Corolario 6.1. *$\{x^{(k)}\}_k$ converge a x^* para todo $x^{(0)}$ si y sólo si $\rho(T) < 1$.*

Este corolario nos brinda un criterio útil para determinar la convergencia de una iteración. Recordemos que Jacobi usaba $T = D^{-1}(L + U)$ mientras que Gauss - Seidel tomaba $T = (D - L)^{-1}U$, de modo tal que, fijada A, basta determinar si el radio espectral de la respectiva T es o no menor que 1.

6.7. Familias de matrices que aseguran la convergencia

Proposición 6.6. *Si $A \in \mathbb{R}^{n\times n}$ es e. d. d. f. entonces Jacobi converge.*

Demostración. Sea $T = D^{-1}(L + U)$ la matriz de la iteración de Jacobi. Queremos ver que $\rho(T) < 1$. Como $\|\cdot\|_\infty$ es consistente, basta probar que $\|T\|_\infty < 1$. Usando la Proposición 4.3, tenemos que

$$\|T\|_\infty = \max_{1\le i\le n} \|fil_i(T)\|_1 = \max_{1\le i\le n} \sum_{\substack{j=1\\ j\neq i}}^{n} \left|-\frac{a_{ij}}{a_{ii}}\right| = \max_{1\le i\le n} \frac{\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|}{|a_{ii}|}$$

Como A es e. d. d. f. entonces $\frac{\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|}{|a_{ii}|} < 1$ para todo i. Luego $\|T\|_\infty < 1$.

□

Proposición 6.7. *Si $A \in \mathbb{R}^{n\times n}$ es e. d. d. f. entonces Gauss - Seidel converge.*

Demostración. Sea $T = (D-L)^{-1}U$ la matriz de la iteración de Gauss - Seidel. Veamos que todo autovalor de T tiene módulo menor que 1. Sea $\lambda \in \mathbb{C}$ un autovector de T y sea $v \in \mathbb{C}^n$ un autovalor asociado, tal que $\|v\|_\infty = 1$. Entonces

$$Tv = \lambda v \Rightarrow (D - L)^{-1}Uv = \lambda v \Rightarrow Uv = \lambda(D - L)v$$

En la última igualdad, miramos una fila $1 \le i \le n$ arbitraria,

$$-\sum_{j=i+1}^{n} a_{ij}v_j = \lambda \sum_{j=1}^{i} a_{ij}v_j \Rightarrow \lambda a_{ii}v_i = -\lambda \sum_{j=1}^{i-1} a_{ij}v_j - \sum_{j=i+1}^{n} a_{ij}v_j$$

Sea i_0 una coordenada de v tal que $|v_{i_0}| = 1$. Tomando módulo en la expresión anterior,

$$|\lambda||a_{i_0 i_0}||v_{i_0}| = \left| -\lambda \sum_{j=1}^{i_0-1} a_{ij} v_j - \sum_{j=i_0+1}^{n} a_{ij} v_j \right|$$

$$\begin{aligned} \Rightarrow |\lambda||a_{i_0 i_0}| &\leq |\lambda| \sum_{j=1}^{i_0-1} |a_{i_0 j}||v_j| + \sum_{j=i_0+1}^{n} |a_{i_0 j}||v_j| \\ &\leq |\lambda| \sum_{j=1}^{i_0-1} |a_{i_0 j}| + \sum_{j=i_0+1}^{n} |a_{i_0 j}| \end{aligned}$$

$$\Rightarrow |\lambda| \left(|a_{i_0 i_0}| - \sum_{j=1}^{i_0-1} |a_{i_0 j}| \right) \leq \sum_{j=i_0+1}^{n} |a_{i_0 j}|$$

Como A es e. d. d. f. entonces $|a_{i_0 i_0}| - \sum_{j=1}^{i_0-1} |a_{i_0 j}| > 0$, y por ende

$$\Rightarrow |\lambda| \leq \frac{\sum_{j=i_0+1}^{n} |a_{i_0 j}|}{|a_{i_0 i_0}| - \sum_{j=1}^{i_0-1} |a_{i_0 j}|}$$

Usando nuevamente que A es e. d. d. f., el lado derecho de la desigualdad debe ser menor que 1, y así concluimos que $|\lambda| < 1$.

□

Proposición 6.8. *Si $A \in \mathbb{R}^{n \times n}$ es simétrica definida positiva entonces Gauss - Seidel converge.*

Observación 6.2. Si $A \in \mathbb{R}^{n \times n}$ es simétrica definida positiva entonces Jacobi **no necesariamente** converge.

6.8. Comparación entre los métodos

El método de Gauss - Seidel tiene dos ventajas sobre el método de Jacobi:

- Tiene un menor requerimiento espacial. Notar que es posible usar un sólo arreglo para almacenar las sucesivas iteraciones (uno que contiene parte de la iteración actual y parte de la anterior), a diferencia de Jacobi que necesita dos (uno para el actual y otro para el anterior).
- Se conocen más familias de matrices convergentes.

La ventaja de Jacobi es su mayor simplicidad.

7. Cálculo de autovalores y autovectores

7.1. Problema

Dada $A \in \mathbb{R}^{n\times n}$ queremos calcular su *autovalor principal* (aquel de módulo máximo) y un autovector asociado. Para esto, vamos a construir una sucesión $\{a_k\}_k$ de números reales convergente al autovalor principal, y una sucesión $\{z^{(k)}\}_k$ de vectores convergente a un autovector asociado.

7.2. Método de las potencias

Supongamos que existe una base $\{v_1, \cdots, v_n\}$ de $\mathbb{R}^n$ formada por autovectores de A (i. e. A es diagonalizable) y sean $\lambda_1, \cdots, \lambda_n \in \mathbb{C}$ los autovalores asociados. Notar que estos son todos los autovalores de A (con posibles repeticiones). Supongamos que A tiene un único autovalor de módulo máximo. Sin pérdida de generalidad, podemos suponer

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$$

Nuestro objetivo es calcular $|\lambda_1|$. Sea $x \in \mathbb{R}^n$ cualquiera, con componente en la dirección de v_1, el autovector de la base asociado al autovalor principal (de ahora en más simplemente *autovector principal*), no nula. Entonces existen $\alpha_1, \cdots, \alpha_n \in \mathbb{C}$, $\alpha_1 \neq 0$, tales que

$$x = \sum_{i=1}^{n} \alpha_i v_i$$

Multiplicando por A^k en ambos miembros:

$$\begin{aligned} A^k x &= \sum_{i=1}^{n} \alpha_i A^k v_i \\ &= \sum_{i=1}^{n} \alpha_i \lambda_i^k v_i \\ &= \lambda_1^k \left(\alpha_1 v_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k v_2 + \cdots + \alpha_n \left(\frac{\lambda_n}{\lambda_1}\right)^k v_n \right) \end{aligned}$$

Consideremos $\phi : \mathbb{R}^n \to \mathbb{R}$ una función contínua, que a lo sumo se anula en 0, y que saca escalares afuera (i. e. $\phi(\lambda v) = |\lambda|\phi(v)$). Ejemplos de funciones que cumplen esto son todas las normas en $\mathbb{R}^n$. Entonces

$$\begin{aligned} \frac{\phi(A^k x)}{\phi(A^{k-1}x)} &= \frac{|\lambda_1^k|\phi\left(\alpha_1 v_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k v_2 + \cdots + \alpha_n \left(\frac{\lambda_n}{\lambda_1}\right)^k v_n\right)}{|\lambda_1^{k-1}|\phi\left(\alpha_1 v_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1}\right)^{k-1} v_2 + \cdots + \alpha_n \left(\frac{\lambda_n}{\lambda_1}\right)^{k-1} v_n\right)} \\ &= |\lambda_1| \frac{\phi\left(\alpha_1 v_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k v_2 + \cdots + \alpha_n \left(\frac{\lambda_n}{\lambda_1}\right)^k v_n\right)}{\phi\left(\alpha_1 v_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1}\right)^{k-1} v_2 + \cdots + \alpha_n \left(\frac{\lambda_n}{\lambda_1}\right)^{k-1} v_n\right)} \end{aligned}$$

Como $|\frac{\lambda_i}{\lambda_1}| < 1$ para todo $i > 1$ entonces $\left(\frac{\lambda_i}{\lambda_1}\right)^k \xrightarrow[k\to\infty]{} 0$, y en definitiva

$$\frac{\phi(A^k x)}{\phi(A^{k-1}x)} \xrightarrow[k\to\infty]{} |\lambda_1| \frac{\phi(\alpha_1 v_1)}{\phi(\alpha_1 v_1)} = |\lambda_1|$$

Esto último vale por ser ϕ contínua, y además como a lo sumo se anula en 0 entonces $\phi(\alpha_1 v_1) \neq 0$ pues $\alpha_1 \neq 0$ y $v_1 \neq 0$. Hemos probado que,

Proposición 7.1. *Sea $A \in \mathbb{R}^{n\times n}$ diagonalizable y con un autovalor principal λ_1. Sea $\phi : \mathbb{R}^n \to \mathbb{R}^n$ como antes. Sea $x \in \mathbb{R}^n$ cualquiera, con componente no nula en la dirección de un autovector principal. Entonces $\left\{\frac{\phi(A^k x)}{\phi(A^{k-1}x)}\right\}_{k\in\mathbb{N}}$ converge a $|\lambda_1|$.*

Definimos las sucesiones $\{a_k\}_k$ y $\{x^{(k)}\}_k$ como sigue. El término inicial $x^{(0)}$ lo elegimos arbitrariamente pero con componente en la dirección de un autovector principal no nula, y además

$$x^{(k)} = Ax^{(k-1)}$$

Se tiene entonces $x^{(k)} = A^k x^{(0)}$. Por otra parte, definimos

$$a_k = \frac{\phi(x^{(k)})}{\phi(x^{(k-1)})}$$

Entonces

$$a_k = \frac{\phi(A^k x^{(0)})}{\phi(A^{k-1} x^{(0)})}$$

Por la proposición anterior, esta sucesión converge a $|\lambda_1|$. Utilizando estas dos sucesiones obtenemos el siguiente algoritmo:

Algorithm 6: Método de las potencias

```
Definir x^(0);
for k = 1 to N do
    x^(k) = A x^(k-1);
    a_k = φ(x^(k)) / φ(x^(k-1));
end
```

Observación 7.1. En la práctica es difícil escoger un $x^{(0)}$ cuya componente en la dirección de un autovector principal sea no nula. En general no conocemos una base de autovectores sin antes calcular los autovalores asociados, con lo cual se hace imposible la escritura de un vector en la base de autovectores. Por este motivo $x^{(0)}$ se suele elegir en forma completamente arbitraria y se ejecuta el método. En caso de que no converja, se elige nuevamente el término inicial. Repetimos este proceso sucesivamente hasta lograr la convergencia. Se puede probar que si el método converge lo hace necesariamente a $|\lambda_1|$.

7.3. Cálculo de un autovector asociado

Para calcular un autovector principal, observemos que como

$$A^k x = \lambda_1^k \left(\alpha_1 v_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k v_2 + \cdots + \alpha_n \left(\frac{\lambda_n}{\lambda_1}\right)^k v_n\right)$$

entonces si k es suficientemente grande

$$A^k x \approx \lambda_1^k \alpha_1 v_1$$

En otras palabras, si k es grande, $A^k x$ es un múltiplo del autovector principal v_1. Entonces $\frac{A^k x}{\|A^k x\|}$ es aproximadamente un múltiplo normalizado del autovector principal v_1. Esta normalización es conveniente por cuestiones numéricas, al evitar que las componentes del vector crezcan desmedidamente.

Definida la sucesión $\{x^{(k)}\}_k$ como antes, definimos una nueva sucesión $\{z^{(k)}\}_k$ con la siguiente regla,

$$z^{(k)} = \frac{x^{(k)}}{\left\|x^{(k)}\right\|}$$

Pero entonces

$$z^{(k)} = \frac{A^k x^{(0)}}{\left\| A^k x^{(0)} \right\|}$$

Luego, por lo anterior, para k grande $z^{(k)}$ se aproxima a un múltiplo de v_1 de norma igual a 1. Esta idea da lugar a una segunda versión del Método de las potencias, más completa.

Algorithm 7: Método de las potencias

```
Definir x^(0);
for k = 1 to N do
    x^(k) = A x^(k-1);
    z^(k) = x^(k) / ||x^(k)||;
    a_k = φ(x^(k)) / φ(x^(k-1));
end
```

7.4. Método de las potencias inversas

Supongamos que ahora los autovalores son tales que

$$|\lambda_1| \geq \cdots \geq |\lambda_{n-1}| > |\lambda_n|$$

y deseamos calcular el autovalor de módulo mínimo $|\lambda_n|$. Notemos que

$$\left| \frac{1}{\lambda_n} \right| > \left| \frac{1}{\lambda_{n-1}} \right| \geq \cdots \geq \left| \frac{1}{\lambda_1} \right|$$

Recordemos que si A es una matriz inversible y $\lambda \in \mathbb{C}$ no nulo es autovalor de A con autovector asociado v, entonces $\frac{1}{\lambda}$ es autovalor de A^{-1} con el mismo autovector v. Por lo tanto, aplicando el Método de las potencias para A^{-1} podemos calcular el autovalor de módulo máximo $\left| \frac{1}{\lambda_n} \right|$.

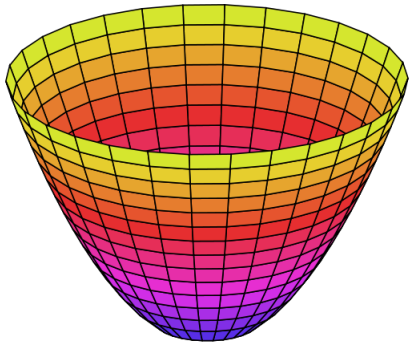
8. Método del gradiente conjungado

8.1. Problema

Dada $A \in \mathbb{R}^{n\times n}$ simétrica definida positiva, queremos calcular la única solución del sistema lineal $Ax = b$.

La idea que usaremos es la siguiente. Consideraremos una función $Q : \mathbb{R}^n \to \mathbb{R}$ que alcanza el mínimo absoluto en la solución $A^{-1}b$ del sistema dado. Comenzando con un punto cualquiera sobre el gráfico de Q, vamos a movernos en una dirección dada, una cierta distancia, acercándonos al mínimo. Repetimos este proceso sucesivas veces hasta llegar al mínimo.

La función Q que consideraremos no es cualquiera entre las que tienen un mínimo absoluto en A. Para poder asegurar que siempre que nos movamos estemos acercándonos al mínimo queremos que el gráfico de Q sea un paraboloide cóncavo.



8.2. El método

Definimos

$$Q(x) = x^t Ax - 2x^t b$$

Como A es simétrica definida positiva, Q resulta tener la forma deseada y además alcanza su mínimo absoluto en $x = A^{-1}b$. Para ver esto último notemos que podemos reescribir $Q(x) = (A^{-1}b - x)^t A(A^{-1}b - x) - b^t A^{-1}b$. Como A es definida positiva entonces $(A^{-1}b - x)^t A(A^{-1}b - x) \geq 0$, con lo cual $Q(x) \geq -b^t A^{-1}b$. Luego Q alcanzará el mínimo $-b^t A^{-1}b$ si y sólo si $(A^{-1}b - x)^t A(A^{-1}b - x) = 0 \Leftrightarrow A^{-1}b - x = 0$, nuevamente debido a que A es definida positiva.

Fijemos un punto inicial $x^{(0)} \in \mathbb{R}^n$ y recorramos en dirección $d^{(0)} \in \mathbb{R}^n$ una cierta distancia dada por $\alpha_0 \in \mathbb{R}$, definiendo así un nuevo punto $x^{(1)} = x^{(0)} + \alpha_0 d^{(0)}$. Con este mismo razonamiento definimos una sucesión $\{x^{(k)}\}_k$ que es tal que

$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$$

Supongamos que la dirección $d^{(k)}$ está fija y es no nula, y miremos $Q(x^{(k+1)}) = Q(x^{(k)} + \alpha d^{(k)})$ como función de α. Si pensamos en la forma que tiene $Q(x)$ al mirarla sobre la recta $t(\alpha) = x^{(k)} + \alpha d^{(k)}$ obtendremos una parábola cóncava, y el punto de dicha parábola más cercano al mínimo será el punto crítico. Calculemos $D(Q(t(\alpha)))$,

$$\begin{aligned}
D(Q(t(\alpha))) &= (\nabla Q(x^{(k)} + \alpha d^{(k)}))^t D(x^{(k)} + \alpha d^{(k)}) \\
&= (2A(x^{(k)} + \alpha d^{(k)}) - 2b)^t d^{(k)} \qquad \text{(pues } \nabla Q(x) = 2Ax - 2b\text{)} \\
&= 2(Ax^{(k)} - b + \alpha A d^{(k)})^t d^{(k)} \\
&= 2(-(b - Ax^{(k)}) + \alpha A d^{(k)})^t d^{(k)} \\
&= 2(-(b - Ax^{(k)})^t d^{(k)} + \alpha (d^{(k)})^t A^t d^{(k)}) \\
&= 2(-(b - Ax^{(k)})^t d^{(k)} + \alpha (d^{(k)})^t A d^{(k)})
\end{aligned}$$

Llamemos $r^{(k)} = b - Ax^{(k)}$ al residuo (lo que falta para llegar a la solución). Entonces

$$D(Q(t(\alpha))) = 2(-(r^{(k)})^t d^{(k)} + \alpha(d^{(k)})^t A d^{(k)})$$

Como $d^{(k)} \neq 0$ y A es definida positiva, entonces $(d^{(k)})^t A d^{(k)} \neq 0$. Luego, el mínimo se alcanza cuando

$$D(Q(t(\alpha))) = 0 \Leftrightarrow \alpha = \frac{(r^{(k)})^t d^{(k)}}{(d^{(k)})^t A d^{(k)}}$$

Entonces elegimos

$$\alpha_k = \frac{(r^{(k)})^t d^{(k)}}{(d^{(k)})^t A d^{(k)}}$$

De este modo, conociendo las direcciones, podemos calcular los α_k óptimos para cada paso.

Observación 8.1. Como A es simétrica definida positiva, la forma $\Phi(x, y) = x^t Ay$ es un producto interno. Notamos $\Phi(x, y) = \langle x, y \rangle_A$.

Observación 8.2. Recordemos que dado un producto interno $\langle , \rangle$ y vectores $u, v \in \mathbb{R}^n$, la proyección ortogonal de u sobre v es

$$\text{proy}_v(u) = \frac{\langle u, v \rangle}{\langle v, v \rangle} v$$

Si escribimos a u en una base de $\langle v \rangle \oplus \langle v \rangle^\perp$, entonces la proyección $\text{proy}_v(u)$ es la componente en la dirección de v. Geométricamente, esto es trazar un hiperplano perpendicular a $\langle v \rangle$ que pase por u y tomar su intersección con $\langle v \rangle$.

Al producto interno canónico en $\mathbb{R}^n$, $\Phi(x, y) = x^t y$, lo notamos $\Phi(x, y) = \langle x, y \rangle_2$. Entonces podemos escribir

$$\alpha_k = \frac{\langle r^{(k)}, d^{(k)} \rangle_2}{\langle d^{(k)}, d^{(k)} \rangle_A} = \frac{\langle e^{(k)}, d^{(k)} \rangle_A}{\langle d^{(k)}, d^{(k)} \rangle_A}$$

donde $e^{(k)} = A^{-1}b - x^{(k)}$ es el error. Entonces $\alpha_k d^{(k)}$ es la proyección ortogonal de $e^{(k)}$ sobre $d^{(k)}$ para el producto interno $\langle , \rangle_A$. Esto muestra que en cada paso, el método suma la componente del error en la dirección dada.

8.3. Elección de las direcciones

Notemos que dependiendo de cómo elijamos las direcciones, convergeremos más o menos rápido, o inclusive podemos no converger. Por lo tanto, queremos estudiar cómo elegir las direcciones para converger al mínimo lo más rápido posible. Pensemos el problema para algunos casos particulares de $Q(x)$ en $\mathbb{R}^2$:

- A diagonal y $b = 0$:

$$Q(x) = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = a_{11}x_1^2 + a_{22}x_2^2$$

  Las curvas de nivel de Q son elipses centradas en el origen. En este caso, conviene elegir una dirección paralela al eje x y otra paralela al eje y.

- A diagonal y $b \neq 0$:

$$Q(x) = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = a_{11}x_1^2 + a_{11}x_2^2 + b_1x_1 + b_2x_2$$

  Las curvas de nivel de Q son elipses con centro $(x_0, y_0) \neq 0$. Conviene elegir una dirección paralela al eje $x = x_0$ y otra paralela al eje $y = y_0$.

- En general:

  Las curvas de nivel de Q son elipses rotadas en un ángulo θ, con centro (x_0, y_0). Conviene elegir dos direcciones, cada una paralela a uno de los ejes rotados de las elipses.

Notemos que en todos los casos alcanzan con dos pasos para converger a la solución. Como es de esperarse, en $\mathbb{R}^n$ alcanzan n pasos para asegurar la convergencia. Esto es lo que probaremos a continuación.

Definición 8.1. Sea $A \in \mathbb{R}^{n\times n}$ definida positiva. Dos vectores $x, y \in \mathbb{R}^n$ se dicen direcciones A-conjugadas si $x^t Ay = 0$. En otras palabras, x e y son A-conjugadas si son vectores ortogonales para el producto interno $\langle , \rangle_A$.

Observación 8.3. Si $x, y \in \mathbb{R}^n$ entonces $\langle x, y\rangle_A = x^t Ay = (A^t x)^t y = \langle A^t x, y\rangle_2$. En particular, si A es simétrica resulta que $\langle x, y\rangle_A = \langle Ax, y\rangle_2$, es decir que x e y son A-conjugadas si y sólo si Ax e y son ortogonales para el producto interno canónico.

Lema 8.1. *Sea $\{v_1, \cdots, v_n\} \subset \mathbb{R}^n$ un conjunto ortogonal para un producto interno $\langle , \rangle$ y de elementos no nulos. Entonces el conjunto es linealmente independiente.*

Demostración. Sea $\sum_{i=1}^n \gamma_i v_i = 0$ una combinación lineal nula. Si $1 \le k \le n$ entonces

$$0 = \langle 0, v_k\rangle = \left\langle \sum_{i=1}^n \gamma_i v_i, v_k \right\rangle = \sum_{i=1}^n \gamma_i \langle v_i, v_k\rangle$$

Dado que el conjunto es ortogonal, todos los términos de la suma son cero salvo para $i = k$,

$$0 = \gamma_k \langle v_k, v_k\rangle = \gamma_k \|v_k\|^2$$

Como $v_k \neq 0$, deducimos que $\gamma_k = 0$. Como k es cualquiera, concluimos que el conjunto es l.i. □

Lema 8.2. *Sea $w \in \mathbb{R}^n$ y $\{v_1, \cdots, v_n\} \subset \mathbb{R}^n$ un conjunto ortogonal para un producto interno $\langle , \rangle$ y de elementos no nulos. Si $\langle w, v_i\rangle = 0$ para todo $i = 1, \cdots, n$, entonces $w = 0$.*

Demostración. Como $\{v_1, \cdots, v_n\}$ es ortogonal y no contiene al cero, entonces es l.i. Como tiene n elementos, entonces es una base. Sea $w = \sum_{i=1}^n \gamma_i v_i$ la escritura de w en la base dada. Entonces

$$0 = \langle w, v_k\rangle = \left\langle \sum_{i=1}^n \gamma_i v_i, v_k \right\rangle = \sum_{i=1}^n \gamma_i \langle v_i, v_k\rangle = \gamma_k \|v_k\|^2$$

En la última igualdad usamos la ortogonalidad de los elementos de la base. Como $v_k \neq 0$ entonces $\gamma_k = 0$, y como k es cualquiera, resulta que $w = 0$. □

El resultado fundamental es el siguiente,

Proposición 8.1. *Sea $A \in \mathbb{R}^{n\times n}$ s. d. p. Sean $d^{(0)}, \cdots, d^{(n-1)}$ direcciones A-conjugadas de a pares y no nulas. Sea $\{x^{(k)}\}_k$ definida como antes. Entonces $Ax^{(n)} = b$, es decir que el método del gradiente conjugado converge a lo sumo en n pasos.*

Demostración. Queremos ver que $x^{(n)} - A^{-1}b = 0$. Como las direcciones son ortogonales de a pares para el p. i. $\langle , \rangle_A$, entonces $\{d^{(0)}, \cdots, d^{(n-1)}\}$ forman un conjunto ortogonal para ese p. i. Como ninguna es nula, por el Lema 8.2 basta ver que $\langle x^{(n)} - A^{-1}b, d^{(k)}\rangle_A = 0$ para todo $k = 0, \cdots, n-1$. Equivalentemente, $\langle Ax^{(n)} - b, d^{(k)}\rangle_2 = 0$ para todo k.

Es fácil ver que si $k > 0$, $x^{(k)} = x^{(0)} + \sum_{i=0}^{k-1} \alpha_i d^{(i)}$. Luego

$$\begin{aligned}
\langle Ax^{(n)} - b, d^{(k)}\rangle_2 &= \left\langle A\left(x^{(0)} + \sum_{i=0}^{n-1} \alpha_i d^{(i)}\right) - b, d^{(k)} \right\rangle_2 \\
&= \left\langle Ax^{(0)} + \sum_{i=0}^{n-1} \alpha_i Ad^{(i)} - b, d^{(k)} \right\rangle_2 \\
&= \langle Ax^{(0)} - b, d^{(k)}\rangle + \sum_{i=0}^{n-1} \alpha_i \langle Ad^{(i)}, d^{(k)}\rangle_2
\end{aligned}$$

Como las direcciones son ortogonales resulta que $\langle Ad^{(i)}, d^{(k)}\rangle_2 = \langle d^{(i)}, d^{(k)}\rangle_A = 0$ si $i \neq k$. Obtenemos así,

$$\langle Ax^{(n)} - b, d^{(k)}\rangle_2 = \langle Ax^{(0)} - b, d^{(k)}\rangle_2 + \alpha_k \langle Ad^{(k)}, d^{(k)}\rangle_2$$

Calculemos $\alpha_k \langle Ad^{(k)}, d^{(k)}\rangle_2$,

$$\begin{aligned}
\alpha_k \langle Ad^{(k)}, d^{(k)}\rangle_2 &= \frac{\langle r^{(k)}, d^{(k)}\rangle_2}{\langle Ad^{(k)}, d^{(k)}\rangle_2} \langle Ad^{(k)}, d^{(k)}\rangle_2 \\
&= \langle r^{(k)}, d^{(k)}\rangle_2 \\
&= \langle b - Ax^{(k)}, d^{(k)}\rangle_2 \\
&= \langle b, d^{(k)}\rangle_2 - \langle Ax^{(k)}, d^{(k)}\rangle_2
\end{aligned}$$

Calculemos $\langle Ax^{(k)}, d^{(k)}\rangle_2$,

$$\begin{aligned}
\langle Ax^{(k)}, d^{(k)}\rangle_2 &= \left\langle Ax^{(0)} + \sum_{i=0}^{k-1} \alpha_i Ad^{(i)}, d^{(k)} \right\rangle_2 \\
&= \langle Ax^{(0)}, d^{(k)}\rangle_2 + \sum_{i=0}^{k-1} \alpha_i \langle Ad^{(i)}, d^{(k)}\rangle_2 \\
&= \langle Ax^{(0)}, d^{(k)}\rangle_2
\end{aligned}$$

Entonces

$$\alpha_k \langle Ad^{(k)}, d^{(k)}\rangle_2 = \langle b, d^{(k)}\rangle_2 - \langle Ax^{(0)}, d^{(k)}\rangle_2 = \langle b - Ax^{(0)}, d^{(k)}\rangle_2$$

Finalmente

$$\langle Ax^{(n)} - b, d^{(k)}\rangle_2 = \langle Ax^{(0)} - b, d^{(k)}\rangle_2 + \langle b - Ax^{(0)}, d^{(k)}\rangle_2 = 0$$

que es lo que queríamos probar.

□

8.4. Generación de direcciones A-conjugadas

La forma de generar direcciones A-conjugadas se basa en el proceso de ortogonalización de Gram-Schmidt. Repasemos este último. Dada una base $\{v_1, \cdots, v_n\}$ de $\mathbb{R}^n$, el proceso genera una base ortogonal $\{w_1, \cdots, w_n\}$ para un producto interno $\langle , \rangle$. Más aún, estos vectores generados son de la forma

$$w_k = v_k - \sum_{i=1}^{k-1} \frac{\langle v_k, w_i\rangle}{\langle w_i, w_i\rangle} w_i$$

Generamos una secuencia de direcciones A-conjugadas del siguiente modo. Fijado $x^{(0)}$, definimos $d^{(0)} = -r^{(0)}$. Para $k > 0$ definimos

$$d^{(k)} = -r^{(k)} - \frac{\langle -r^{(k)}, d^{(k-1)}\rangle_A}{\langle d^{(k-1)}, d^{(k-1)}\rangle_A} d^{(k-1)}$$

Los vectores $-r^{(k)}$ juegan el papel de los v_k, y los $d^{(k)}$ el de los w_k. La razón de que $d^{(k)}$ sólo dependa de $d^{(k-1)}$ y no de $d^{(j)}$ con $j < k-1$ (como en el esquema anterior de Gram-Schmidt) es que $\langle -r^{(k)}, d^{(j)}\rangle_A = 0$ para $j < k-1$. Entonces los vectores $d^{(0)}, \cdots, d^{(n-1)}$ así generados son ortogonales respecto del producto interno $\langle , \rangle_A$, es decir, son A-conjugados.

8.5. Comparación con Cholesky

Hemos visto que en el caso de A simétrica definida positiva, la resolución de $Ax = b$ se puede hacer vía la factorización de Cholesky. El método del gradiente conjugado es iterativo, con lo cual las ventajas frente al método directo son las anteriormente mencionadas.

9. Descomposición en valores singulares

9.1. Problema

Dada $A \in \mathbb{R}^{m\times n}$ queremos descomponer $A = U\Sigma V^t$ con $U \in \mathbb{R}^{m\times m}$ ortogonal, $\Sigma \in \mathbb{R}^{m\times n}$ diagonal y $V \in \mathbb{R}^{n\times n}$ ortogonal.

9.2. Lemas auxiliares

Para llegar al resultado principal de esta sección necesitaremos algunos resultados auxiliares.

Observación 9.1. El producto interno canónico en $\mathbb{C}^n$ es la forma $\Phi : \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$ tal que

$$\Phi(x, y) = x^t\overline{y}$$

donde $\overline{y}$ es la conjugación de y coordenada a coordenada. Notar que restringido a $\mathbb{R}^n$, coincide con el producto interno canónico en $\mathbb{R}^n$.

En lo que sigue, todos los productos internos considerados serán el canónico en $\mathbb{C}^n$, y lo notaremos con $\langle , \rangle$.

Observación 9.2. Si $\alpha \in \mathbb{C}$ y $x, y \in \mathbb{C}^n$, entonces

$$\langle \alpha x, y\rangle = (\alpha x)^t\overline{y} = \alpha x^t\overline{y} = \alpha\langle x, y\rangle$$

$$\langle x, \alpha y\rangle = x^t\overline{\alpha y} = x^t(\overline{\alpha}\ \overline{y}) = \overline{\alpha}x^t\overline{y} = \overline{\alpha}\langle x, y\rangle$$

Observación 9.3. Si $A \in \mathbb{R}^{n\times n}$ es simétrica y $v, w \in \mathbb{C}^n$, entonces

$$\langle Av, w\rangle = (Av)^t\overline{w} = v^tA^t\overline{w} = v^tA\overline{w} = v^t\overline{\overline{A}w} = v^t\overline{Aw} = \langle v, Aw\rangle$$

Lema 9.1. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica. Entonces todos sus autovalores son reales.*

Demostración. Sea $\lambda \in \mathbb{C}$ un autovalor de A. Sea $v \in \mathbb{C}^n$ un autovector asociado. Entonces

$$\langle Av, v\rangle = \langle \lambda v, v\rangle = \lambda\langle v, v\rangle = \lambda \|v\|^2$$

Por otro lado,

$$\langle Av, v\rangle = \langle v, Av\rangle = \langle v, \lambda v\rangle = \overline{\lambda}\langle v, v\rangle = \overline{\lambda} \|v\|^2$$

Luego $\lambda \|v\|^2 = \langle Av, v\rangle = \overline{\lambda} \|v\|^2$, es decir que $\lambda \|v\|^2 = \overline{\lambda} \|v\|^2$. Como $v \neq 0$ por ser autovector, entonces $\lambda = \overline{\lambda}$, por lo que concluimos que $\lambda \in \mathbb{R}$. □

Lema 9.2. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica. Si $\lambda \in \mathbb{R}$ es autovalor de A entonces λ tiene un autovector asociado con coordenadas reales.*

Demostración. Sea $v \in \mathbb{C}^n$ un autovector asociado a λ. Si todas las coordenadas de v son imaginarios puros, entonces definimos $w = iv$, que tiene todas las coordenadas reales. Además $w \neq 0$ porque $v \neq 0$. Se tiene

$$Aw = A(iv) = iAv = i(\lambda v) = \lambda(iv) = \lambda w$$

Entonces $w \in \mathbb{R}^n$ es un autovector asociado a λ.

Veamos el caso en que v tiene alguna coordenada que no es imaginario puro. Consideramos $w = v + \overline{v}$. Es claro que w es un vector de números reales. La coordenada que en v no era un imaginario pura tiene parte real no nula, con lo cual esa misma coordenada es no nula en w. Luego $w \neq 0$. Además, usando que A y λ son reales,

$$Aw = A(v + \overline{v}) = Av + A\overline{v} = Av + \overline{\overline{A}\overline{v}} = Av + \overline{Av} = \lambda v + \overline{\lambda v} = \lambda v + \overline{\lambda}\overline{v} = \lambda v + \lambda\overline{v} = \lambda(v + \overline{v}) = \lambda w$$

Luego $w \in \mathbb{R}^n$ es un autovector asociado a λ. □

Proposición 9.1. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica. Entonces existen $Q \in \mathbb{R}^{n\times n}$ ortogonal y $D \in \mathbb{R}^{n\times n}$ diagonal tal que*

$$A = QDQ^t$$

Demostración. Inducción en $n \in \mathbb{N}$.

Si $n = 1$, $A = \begin{pmatrix} a \end{pmatrix}$ para cierto $a \in \mathbb{R}$, y tomamos $Q = \begin{pmatrix} 1 \end{pmatrix}$ y $D = \begin{pmatrix} a \end{pmatrix}$.

Sea $n > 1$. Sea $\lambda \in \mathbb{R}$ un autovalor de A, y $v \in \mathbb{R}^n$ un autovector asociado con coordenadas reales, normalizado. Completamos v a una base de $\mathbb{R}^n$, con $v_2, \cdots, v_n$. A la base $\{v, v_2, \cdots, v_n\} \subset \mathbb{R}^n$ le aplicamos el proceso de ortogonalización de Gram-Schmidt, obteniendose $\{v, w_2, \cdots, w_n\} \subset \mathbb{R}^n$. Sea

$$W = \left(\begin{array}{c|c|c} & & \\ w_2 & \cdots & w_n \\ & & \end{array}\right) \in \mathbb{R}^{n\times(n-1)}$$

A partir de esta, definimos

$$U = \left(\begin{array}{c|c} & \\ v & W \\ & \end{array}\right) \in \mathbb{R}^{n\times n}$$

Entonces

$$\begin{aligned} U^tAU &= \left(\begin{array}{c} v^t \\ \hline \\ W^t \\ \\ \end{array}\right) A \left(\begin{array}{c|c} & \\ v & W \\ & \end{array}\right) \\ &= \left(\begin{array}{c} v^tA \\ \hline \\ W^tA \\ \\ \end{array}\right) \left(\begin{array}{c|c} & \\ v & W \\ & \end{array}\right) \\ &= \left(\begin{array}{c|c} v^tAv & v^tAW \\ \hline W^tAu & W^tAW \end{array}\right) \end{aligned}$$

Tenemos que

$$v^tAv = v^t(\lambda v) = \lambda v^tv = \lambda \|v\|^2 = \lambda$$

Además $W^tAv = W^t(\lambda v) = \lambda W^tv$, y como las filas de W^t son ortogonales a v entonces

$$W^tAv = 0$$

Análogamente

$$v^tAW = (A^tv)^tW = (Av)^tW = (\lambda v)^tW = \lambda v^tW = 0$$

Luego

$$U^tAU = \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & W^tAW & \\ 0 & & & \end{array}\right)$$

Resta calcular el bloque W^tAW. Como $W^tAW \in \mathbb{R}^{(n-1)\times(n-1)}$ es simétrica, entonces por hipótesis inductiva existen $\tilde{P} \in \mathbb{R}^{(n-1)\times(n-1)}$ ortogonal y $\tilde{D} \in \mathbb{R}^{(n-1)\times(n-1)}$ diagonal, tal que $W^tAW = \tilde{P}\tilde{D}\tilde{P}^t$. Extendemos $\tilde{P}$ definiendo

$$P = \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{P} & \\ 0 & & & \end{array}\right) \in \mathbb{R}^{n\times n}$$

A $\tilde{D}$ la extendemos del siguiente modo

$$D = \left(\begin{array}{c|ccc} \lambda & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{D} & \\ 0 & & & \end{array}\right) \in \mathbb{R}^{n\times n}$$

Notemos que, como $\tilde{P}$ es ortogonal, entonces P también lo es. Análogamente, como $\tilde{D}$ es triangular superior, entonces D también lo es. Luego

$$\begin{aligned} PDP^t &= \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{P} & \\ 0 & & & \end{array}\right)\left(\begin{array}{c|ccc} \lambda & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{D} & \\ 0 & & & \end{array}\right)\left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{P}^t & \\ 0 & & & \end{array}\right) \\ &= \left(\begin{array}{c|ccc} \lambda & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{P}\tilde{D} & \\ 0 & & & \end{array}\right)\left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{P}^t & \\ 0 & & & \end{array}\right) \\ &= \left(\begin{array}{c|ccc} \lambda & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{P}\tilde{D}\tilde{P}^t & \\ 0 & & & \end{array}\right) \\ &= \left(\begin{array}{c|ccc} \lambda & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & W^tAW & \\ 0 & & & \end{array}\right) \\ &= U^tAU \end{aligned}$$

Entonces $PDP^t = U^tAU$, con lo cual $A = (UP)D(UP)^t$. Poniendo $Q = UP \in \mathbb{R}^{n\times n}$ resulta que Q es ortogonal por ser producto de matrices ortogonales, y $A = QDQ^t$, que es lo que queríamos demostrar. □

Lema 9.3. *En las condiciones de la proposición anterior, las columnas de Q son autovectores de A y los elementos de la diagonal de D son autovalores de A. Más precisamente $col_i(Q)$ es autovector de A de autovalor D_{ii}.*

Demostración. Se tiene

$$Q^tAQ = D \Rightarrow QQ^tAQ = QD \Rightarrow AQ = QD \Rightarrow col_i(AQ) = col_i(QD) \Rightarrow A\ col_i(Q) = D_{ii}\ col_i(Q)$$

Como Q es ortogonal, en particular es inversible y por lo tanto no tiene columnas nulas. Luego, $col_i(Q) \neq 0$ es autovector de A de autovalor D_{ii}. □

Teorema 9.1. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica. Entonces existe una base ortonormal de $\mathbb{R}^n$ formada por autovectores reales de A.*

Demostración. Escribamos $A = QDQ^t$ como antes. Por el lema previo, las columnas de Q son autovectores de A. Como Q es es inversible, sus columnas son linealmente independientes y, por lo tanto, como son n, forman una base de $\mathbb{R}^n$. Más aún, son vectores ortogonales, pues Q es ortogonal. Para que la base sea ortonormal, basta dividir cada vector por su norma. □

9.3. Teorema de descomposición en valores singulares

Teorema 9.2. *Sea $A \in \mathbb{R}^{m\times n}$ arbitraria. Entonces existen $U \in \mathbb{R}^{m\times m}$ ortogonal, $V \in \mathbb{R}^{n\times n}$ ortogonal y $\Sigma \in \mathbb{R}^{m\times n}$ diagonal tal que*

$$A = U\Sigma V^t$$

Demostración. Vamos a descomponer la prueba en varios pasos.

1. **La matriz $AA^t \in \mathbb{R}^{m\times m}$ es simétrica.** Por la Proposición 9.1 existe una base ortonormal $\{u_1, \cdots, u_m\}$ de $\mathbb{R}^m$ formada por autovectores de AA^t. Sean $\lambda_1, \cdots, \lambda_m \in \mathbb{R}$ los autovalores asociados, todos reales por el Lema 9.1.

2. **Los autovalores $\lambda_1, \cdots, \lambda_m$ son no negativos.**

$$\begin{aligned} AA^t u_i = \lambda_i u_i &\Rightarrow u_i^t AA^t u_i = \lambda_i u_i^t u_i \\ &\Rightarrow (A^t u_i)^t (A^t u_i) = \lambda_i \|u_i\|_2^2 \\ &\Rightarrow \left\|A^t u_i\right\|_2^2 = \lambda_i \|u_i\|_2^2 \\ &\Rightarrow \lambda_i \geq 0 \end{aligned}$$

3. Supongamos sin pérdida de generalidad que $\lambda_1, \cdots, \lambda_r$ son los autovalores de AA^t no nulos y definamos para cada $1 \leq i \leq r$,

$$\begin{aligned} \sigma_i &= \sqrt{\lambda_i} \\ v_i &= \frac{1}{\sigma_i} A^t u_i \end{aligned}$$

   Notemos que $v_i \neq 0$ porque $Av_i = \frac{1}{\sigma_i} AA^t u_i = \frac{\lambda_i}{\sigma_i} u_i \neq 0$.

4. **Para cada $1 \leq i \leq r$, v_i es autovector de A^tA de autovalor λ_i.** Ya vimos que $v_i \neq 0$. Además

$$A^t A v_i = \frac{1}{\sigma_i} A^t AA^t u_i = \frac{\lambda_i}{\sigma_i} A^t u_i = \lambda_i v_i$$

5. **El conjunto $\{v_1, \cdots, v_r\}$ es ortonormal.**

$$
\begin{aligned}
v_i^t v_j = \left(\frac{A^t u_i}{\sigma_i}\right)^t \left(\frac{A^t u_j}{\sigma_j}\right) &= \frac{1}{\sigma_i \sigma_j} u_i^t A A^t u_j \\
&= \frac{1}{\sigma_i \sigma_j} u_i^t \lambda_j u_j \\
&= \frac{\lambda_j}{\sigma_i \sigma_j} u_i^t u_j \\
&= \frac{\lambda_j}{\sigma_i \sigma_j} \delta_{ij} \\
&= \frac{\lambda_j}{\sqrt{\lambda_i}\sqrt{\lambda_j}} \delta_{ij} \\
&= \delta_{ij}
\end{aligned}
$$

6. Extendemos $\{v_1, \cdots, v_r\}$ a una base ortonormal de $\mathbb{R}^n$ con $v_{r+1}, \cdots, v_n$. Definimos

$$
U = \begin{pmatrix} u_1 & \cdots & u_m \end{pmatrix} \qquad V = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & 0 & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & 0 & & & \ddots & \\ & & & & & 0 \end{pmatrix}
$$

7. Veamos que $A = U\Sigma V^t$. Como U y V son ortogonales, es lo mismo ver que $U^t A V = \Sigma$. Se tiene

$$
\begin{aligned}
(U^t A V)_{ij} &= fil_i(U^t A)\ col_j(V) \\
&= fil_i(U^t) A\ col_j(V) \\
&= col_i(U)^t A\ col_j(V) \\
&= u_i^t A v_j
\end{aligned}
$$

Separemos en casos:

- Si $j \leq r$:
  En este caso $v_j = \frac{1}{\sigma_j} A^t u_j$, con lo cual

$$
\begin{aligned}
u_i^t A v_j &= \frac{1}{\sigma_j} u_i^t A A^t u_j \\
&= \frac{1}{\sigma_j} u_i^t (\lambda_j u_j) \\
&= \frac{\lambda_j}{\sigma_j} u_i^t u_j \\
&= \frac{\sigma_j^2}{\sigma_j} \delta_{ij} \qquad \text{(pues } \{u_1, \cdots, u_m\} \text{ es ortonormal)} \\
&= \sigma_j \delta_{ij}
\end{aligned}
$$

- Si $j > r$:
  - Si $i \leq r$:
    En este caso $v_i = \frac{1}{\sigma_i} A^t u_i \Rightarrow \sigma_i v_i = A^t u_i$, con lo cual

$$\begin{aligned} u_i^t A v_j &= (u_i^t A v_j)^t \\ &= v_j^t A^t u_i \\ &= v_j^t (\sigma_i v_i) \\ &= \sigma_i v_j^t v_i \\ &= \sigma_i \delta_{ij} && \text{(pues } \{v_1, \cdots, v_n\} \text{ es ortonormal)} \\ &= 0 && \text{(pues } i \leq r < j\text{)} \end{aligned}$$

- Si $i > r$:
  Tenemos entonces que $\lambda_i = 0$. Veamos que necesariamente $A^t u_i = 0$. Por el absurdo supongamos que $A^t u_i \neq 0$, entonces $0 \neq \|A^t u_i\|_2^2 = (A^t u_i)^t (A^t u_i) = u_i^t A A^t u_i = u_i^t (\lambda_i u_i) = \lambda_i u_i^t u_i = \lambda_i \|u_i\|_2^2$. Luego $0 \neq \lambda_i \|u_i\|_2^2 \Rightarrow \lambda_i \neq 0$, absurdo.
  Como $A^t u_i = 0$ entonces

$$u_i^t A v_j = (u_i^t A v_j)^t = v_j^t A^t u_i = 0$$

□

Los elementos no nulos de la diagonal de Σ, $\sigma_1, \cdots, \sigma_r$, se llaman valores singulares de A. Si bien la descomposición en valores singulares no es única, las matrices U, Σ y V siempre cumplen ciertas propiedades

Proposición 9.2. *Sea $A = U\Sigma V^t$ una descomposición en valores singulares. Sean $\sigma_1, \cdots, \sigma_r$ los valores singulares. Sean $u_1, \cdots, u_m$ las columnas de U, y $v_1, \cdots, v_n$ las columnas de V. Entonces*

- *u_i es autovector de AA^t.*
- *v_i es autovector de $A^t A$.*

En ambos casos, si $i \leq r$ entonces el autovalor asociado es σ_i^2 y es 0 en caso contrario.

Demostración. Tenemos que

$$AA^t = (U\Sigma V^t)(U\Sigma V^t)^t = U\Sigma V^t V \Sigma^t U^t = U\Sigma\Sigma^t U^t$$

Notemos que $\Sigma\Sigma^t = \overline{\Sigma} = \text{diag}(\sigma_1^2, \cdots, \sigma_r^2, 0, \cdots, 0) \in \mathbb{R}^{m \times m}$. Entonces

$$AA^t u_i = U\overline{\Sigma}U^t u_i = U\overline{\Sigma}e_i = U \; col_i(\overline{\Sigma})$$

Si $i \leq r$ entonces $col_i(\overline{\Sigma}) = \sigma_i^2 e_i$ con lo cual $AA^t u_i = \sigma_i^2 U e_i = \sigma_i^2 u_i$, es decir que u_i es autovector de AA^t de autovalor σ_i^2. Si $i > r$ entonces $col_i(\overline{\Sigma}) = 0$ con lo cual u_i es autovector de autovalor 0.

Para los v_i la demostración es igual. □

10. Cuadrados mínimos lineales

10.1. Problema

Supongamos que disponemos de una muestra de valores provenientes de un experimento

x	y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_m	y_m

y queremos encontrar una relación entre los valores de x e y. Tenemos dos opciones:

1. Encontrar una función f tal que $f(x_i) = y_i$ para todo $i = 1, \cdots, m$.
2. Encontrar una función f tal que $f(x_i) \approx y_i$ para todo $i = 1, \cdots, m$.

Cuando tratamos con muestras observadas, en general estamos lidiando con valores que ya poseen un error, proveniente de la medición. Esto implica que buscar relaciones exactas (del tipo 1) pierda sentido pues en realidad sólo conocemos aproximaciones de los valores reales. Por este motivo, nos concentraremos en la opción 2.

Suponiendo que contamos con una familia de funciones $\mathcal{F}$ de la cual queremos extraer una de ellas, necesitamos un criterio para comparar funciones que contiene. Tres criterios posibles son:

- $\min\limits_{f \in \mathcal{F}} \max\limits_{1 \leq i \leq m} |f(x_i) - y_i|$

  Minimiza el máximo error en un punto. La desventaja de este criterio es que es muy susceptible a la presencia de outliers (valores atípicos).

- $\min\limits_{f \in \mathcal{F}} \sum_{i=1}^{m} |f(x_i) - y_i|$

  Minimiza el error. Este criterio sobrepasa el problema de los outliers. La desventaja que tiene es que queremos encontrar el mínimo de una función que involucra un valor absoluto, que sabemos que no es derivable en el origen, lo cual dificulta la tarea.

- $\min\limits_{f \in \mathcal{F}} \sum_{i=1}^{m} (f(x_i) - y_i)^2$

  Minimiza el error cuadrático. Este criterio no padece de ninguno de los problemas anteriores.

Utilizaremos este último criterio. La familia de funciones que vamos a considerar es

$$\mathcal{F} = \{a_1\phi_1 + \cdots + a_n\phi_n : a_1, \cdots, a_n \in \mathbb{R}\}$$

donde las funciones reales $\phi_1, \cdots, \phi_n$ están fijas. En otras palabras, estamos considerando el subespacio de funciones

$$\mathcal{F} = \langle \phi_1, \cdots, \phi_n \rangle_{\mathbb{R}}$$

Entonces el problema es encontrar los coeficientes $a_1, \cdots, a_n$ que realizan el mínimo

$$\min_{a_1, \cdots, a_n \in \mathbb{R}} \sum_{i=1}^{m} (a_1\phi_1(x_i) + \cdots + a_n\phi_n(x_i) - y_i)^2$$

Consideremos

$$A = \begin{pmatrix} \phi_1(x_1) & \cdots & \phi_n(x_1) \\ \phi_1(x_2) & \cdots & \phi_n(x_2) \\ \vdots & & \vdots \\ \phi_1(x_n) & \cdots & \phi_n(x_n) \end{pmatrix} \qquad x = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \qquad b = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$$

Entonces el mínimo a calcular lo podemos escribir como

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

Observación 10.1.

1. No necesitamos conocer el valor de las ϕ_i en todo su dominio, sino sólo en los puntos $x_1, \cdots, x_m$.
2. Para que el problema tenga sentido necesitamos más datos que incógnitas, i.e., $m \geq n$.

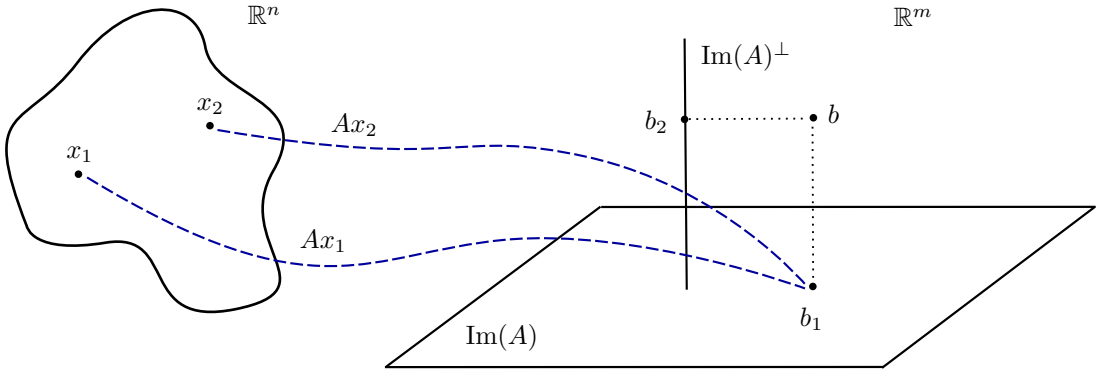
Abstrayéndonos de las funciones $\phi_1, \cdots, \phi_n$ y de las muestras $(x_1, y_1), \cdots, (x_m, y_m)$, definimos el problema de cuadrados mínimos lineales del siguiente modo. Dadas $A \in \mathbb{R}^{m\times n}$ y $b \in \mathbb{R}^m$, hallar $x \in \mathbb{R}^n$ que minimice $\|Ax-b\|_2^2$.

Lo estudiaremos y atacaremos desde un punto de vista puramente algebraico.

10.2. Intuición geométrica

En primer lugar, observemos que si el sistema $Ax = b$ tiene solución, entonces cualquiera de ellas realiza el mínimo, que es 0.

Si $Ax = b$ no tiene solución, entonces es evidente que el mínimo es mayor que 0. Para entender cómo elegir un vector x que lo realice, pensemos en Ax y b como vectores en $\mathbb{R}^m$. El mínimo se alcanza cuando la distancia euclídea entre estos dos vectores es mínima. Pero el único de estos dos vectores que se mueve es Ax, con lo cual hay que elegirlo de modo tal que esté lo más cerca posible de b. Recordemos que el conjunto de valores que puede tomar Ax es el subespacio $\mathrm{Im}(A) = \{Ax : x \in \mathbb{R}^n\}$. Luego, queremos encontrar la distancia del punto b al subespacio $\mathrm{Im}(A)$, y es sabido que el punto sobre el subespacio que realiza la distancia es la proyección ortogonal de b sobre $\mathrm{Im}(A)$.



En la figura, $b_1 = \mathrm{proy}_{\mathrm{Im}(A)}(b)$ es el punto que realiza la distancia, $b_2 = \mathrm{proy}_{\mathrm{Im}(A)^\perp}(b)$, y x_1 y x_2 son soluciones.

10.3. Solución

Como $\mathbb{R}^m = \mathrm{Im}(A) \oplus \mathrm{Im}(A)^\perp$ entonces b se escribe en forma única como $b = b_1 + b_2$ con $b_1 \in \mathrm{Im}(A)$ y $b_2 \in \mathrm{Im}(A)^\perp$. Recordemos que b_1 es la proyección ortogonal de b sobre $\mathrm{Im}(A)$ y b_2 es la proyección ortogonal de b sobre $\mathrm{Im}(A)^\perp$.

Proposición 10.1. *Sea $b = b_1 + b_2$ la escritura en forma única como $b_1 \in Im(A)$ y $b_2 \in Im(A)^\perp$. Entonces el mínimo $\min_{x\in\mathbb{R}^n} \|Ax-b\|_2^2$ se realiza únicamente cuando $Ax = b_1$.*

Demostración.

$$\begin{aligned}
\|Ax-b\|_2^2 &= \|Ax-(b_1+b_2)\|_2^2 \\
&= \|(Ax-b_1)+b_2\|_2^2 \\
&= \|Ax-b_1\|_2^2 + \|b_2\|_2^2 && \text{(por Pitágoras)} \\
&\geq \|b_2\|_2^2
\end{aligned}$$

Entonces el mínimo se realiza sí y sólo si $\|Ax - b_1\|_2^2 = 0 \Leftrightarrow Ax = b_1$.

□

Observemos que como $b_1 \in \text{Im}(A)$ entonces siempre existe un vector x que realiza el mínimo. Es decir, cuadrados mínimos lineales siempre tiene solución. Estudiemos ahora la unicidad de la misma.

Es importante notar que la unicidad no depende del conjunto $\text{Im}(A)$ si no de cómo la matriz A transforma los vectores. Más precisamente, existe único $x \in \mathbb{R}^n$ tal que $Ax = b_1$ si y sólo si la transformación lineal $f_A(x) = Ax$ es un monomorfismo. En términos de los elementos de la matriz, tenemos el siguiente resultado,

Proposición 10.2. *La solución de cuadrados mínimos lineales es única si y sólo si A es inversible.*

Demostración. Existirá un único $x \in \mathbb{R}^n$ tal que $Ax = b_1$ si y sólo si la matriz A es inversible. □

10.4. Ecuaciones normales

Nuestro próximo objetivo es caracterizar la solución en términos de A y de b.

Lema 10.1. *$Im(A)^\perp = Nu(A^t)$.*

Demostración. ($\subseteq$) Sea $v \in \text{Im}(A)^\perp$, entonces $\langle w, v\rangle_2 = 0$ para todo $w \in \text{Im}(A)$. Queremos ver que $v \in \text{Nu}(A^t)$.
Se tiene

$$A^t v = \begin{pmatrix} fil_1(A^t)v \\ \vdots \\ fil_n(A^t)v \end{pmatrix} = \begin{pmatrix} col_1(A)^t v \\ \vdots \\ col_n(A)^t v \end{pmatrix} = \begin{pmatrix} \langle col_1(A), v\rangle_2 \\ \vdots \\ \langle col_n(A), v\rangle_2 \end{pmatrix} = 0$$

pues $col_i(A) \in \text{Im}(A)$ para todo i.

($\supseteq$) Sea $v \in \text{Nu}(A^t)$, entonces $A^t v = 0$. Sea $w \in \text{Im}(A)$, queremos ver que $\langle w, v\rangle_2 = 0$.
Como $w \in \text{Im}(A)$, entonces $Az = w$ para algún $z \in \mathbb{R}^n$. Luego $w^t = z^t A^t$ con lo cual $\langle w, v\rangle_2 = w^t v = z^t A^t v = 0$. □

Proposición 10.3. *$x \in \mathbb{R}^n$ es solución de cuadrados mínimos lineales si y sólo si $A^t Ax = A^t b$.*

Demostración. ($\Rightarrow$) Sea x una solución. Escribamos $b = b_1 + b_2$ con $b_1 \in \text{Im}(A)$ y $b_2 \in \text{Im}(A)^\perp$. Entonces $Ax = b_1 \Rightarrow A^t Ax = A^t b_1$. Como $\text{Im}(A)^\perp = \text{Nu}(A^t)$ entonces $A^t b_2 = 0$, con lo cual $A^t Ax = A^t b_1 = A^t b_1 + A^t b_2 = A^t(b_1 + b_2) = A^t b$.

($\Leftarrow$) Sea x tal que $A^t Ax = A^t b = A^t(b_1 + b_2) = A^t b_1 \Rightarrow A^t Ax - A^t b_1 = 0 \Rightarrow A^t(Ax - b_1) = 0 \Rightarrow Ax - b_1 \in \text{Nu}(A^t) = \text{Im}(A)^\perp$. Entonces $Ax - b_1 \in \text{Nu}(A^t) = \text{Im}(A)^\perp \cap \text{Im}(A) = \{0\} \Rightarrow Ax = b_1$ entonces x es solución de cuadrados mínimos lineales. □

Entonces resolver cuadrados mínimos equivale a resolver el sistema $A^t Ax = A^t b$. Este sistema de ecuaciones recibe el nombre de *ecuaciones normales*. Sin embargo, la resolución vía este sistema puede no ser númericamente estable, pues la matriz $A^t A$ suele estar mal condicionada aún estando A bien condicionada. Por ejemplo,

$$A = \begin{pmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix} \qquad A^t A = \begin{pmatrix} 1+\varepsilon^2 & 1 \\ 1 & 1+\varepsilon^2 \end{pmatrix}$$

En este caso $A^t A$ está muy mal condicionada. Si ε es chico, entonces ε^2 es despreciable, y en un contexto de aritmética finita puede ser absorbido en la suma, obteniéndose $fl(fl(1) + fl(\varepsilon^2)) = 1$. Con este redondeo resulta que $A^t A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ de modo que las ecuaciones normales tendrán infinitas soluciones. Sin embargo A tiene columnas l. i. con lo cual la solución de cuadrados mínimos es única.

10.5. Resolución por QR

Sean $A \in \mathbb{R}^{m\times n}$ y $b \in \mathbb{R}^m$ y supongamos sin pérdida de generalidad que $m \geq n$ (si no fuera este el caso, agregamos filas de 0 en A y b). Sea $A = QR$ su factorización QR. Como Q^t preserva norma 2 entonces

$$\|Ax - b\|_2^2 = \left\|Q^t(Ax - b)\right\|_2^2 = \left\|Q^t(QRx - b)\right\|_2^2 = \left\|Rx - Q^tb\right\|_2^2$$

Tenemos dos casos:

- Las columnas de A son l. i.:

  Como Q es inversible entonces $\text{rg}(A) = \text{rg}(R)$. Entonces, en este caso, las columnas de R también son l. i. y por lo tanto tiene la forma

  $$R = \begin{pmatrix} & R_1 & \\ & & \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

  con $R_1 \in \mathbb{R}^{n\times n}$ triangular superior. Escribamos además

  $$Q^tb = \begin{pmatrix} c \\ d \end{pmatrix}$$

  con $c \in \mathbb{R}^n$ y $d \in \mathbb{R}^{m-n}$. Entonces

  $$\|Ax - b\|_2^2 = \left\|\begin{pmatrix} R_1x \\ 0 \end{pmatrix} - \begin{pmatrix} c \\ d \end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix} R_1x - c \\ -d \end{pmatrix}\right\|_2^2 = \|R_1x - c\|_2^2 + \|d\|_2^2$$

  Entonces el mínimo se realiza si y sólo si $R_1x = c$ y este sistema tiene solución única pues R_1 es una matriz cuadrada de rango máximo.

- Las columnas de A no son l. i.:

  Entonces las columnas de R no son l. i., y si $\text{rg}(A) = r < n$ entonces

  $$R = \begin{pmatrix} & R_1 & R_2 \\ & & \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

  con $R_1 \in \mathbb{R}^{r\times r}$ triangular superior y $R_2 \in \mathbb{R}^{r\times(n-r)}$. Además escribimos

  $$Q^tb = \begin{pmatrix} c \\ d \end{pmatrix} \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

  ahora con $c \in \mathbb{R}^r$, $d \in \mathbb{R}^{m-r}$, $x_1 \in \mathbb{R}^r$ y $x_2 \in \mathbb{R}^{n-r}$. Luego

  $$\|Ax - b\|_2^2 = \left\|\begin{pmatrix} R_1x_1 + R_2x_2 - c \\ -d \end{pmatrix}\right\|_2^2 = \|R_1x_1 + R_2x_2 - c\|_2^2 + \|d\|_2^2$$

  Entonces el mínimo se realiza si y sólo si $R_1x_1 + R_2x_2 = c$. Este sistema (de r ecuaciones y $n > r$ incógnitas) tiene infinitas soluciones, que se obtienen fijando $x_2 = \overline{x}_2$ y resolviendo $R_1x_1 = c - R_2\overline{x}_2$ que tiene solución única por ser R_1 cuadrada y de rango máximo.

10.6. Resolución por SVD

Consideremos ahora la descomposición en valores singulares $A = U\Sigma V^t$. Sean $\sigma_1, \cdots, \sigma_r$ los valores singulares. Tenemos que

$$\|Ax - b\|_2^2 = \left\|U^t(Ax - b)\right\|_2^2 = \left\|U^t(U\Sigma V^t x - b)\right\|_2^2 = \left\|\Sigma V^t x - U^t b\right\|_2^2$$

Como V^t es inversible entonces sustituyendo $y = V^t x$:

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2 = \min_{x\in\mathbb{R}^n} \left\|\Sigma V^t x - U^t b\right\|_2^2 = \min_{y\in\mathbb{R}^n} \left\|\Sigma y - U^t b\right\|_2^2$$

Separemos en los mismos dos casos de antes:

- Las columnas de A son l. i.:

  Notemos que como U y V son inversibles entonces $\operatorname{rg}(A) = \operatorname{rg}(\Sigma) = r$. Entonces, en este caso, $r = n$ y Σ tiene la forma

  $$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

  Escribiendo

  $$U^t b = \begin{pmatrix} c \\ d \end{pmatrix}$$

  con $c \in \mathbb{R}^n$ y $d \in \mathbb{R}^{m-n}$, tenemos

  $$\left\|\Sigma y - U^t b\right\|_2^2 = \left\| \begin{pmatrix} \sigma_1 y_1 \\ \vdots \\ \sigma_n y_n \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} c \\ d \end{pmatrix} \right\|_2^2 = \left\| \begin{pmatrix} \sigma_1 y_1 - c_1 \\ \vdots \\ \sigma_n y_n - c_n \end{pmatrix} \right\|_2^2 + \|d\|_2^2$$

  Entonces el mínimo se alcanza si y sólo si $y = \begin{pmatrix} c_1/\sigma_1 \\ \vdots \\ c_n/\sigma_n \end{pmatrix}$. Lo que resta es encontrar el único x tal que $V^t x = y$.

- Las columnas de A no son l. i.:

  En este caso hay $r < n$ valores singulares en la diagonal de Σ. Escribiendo

  $$U^t b = \begin{pmatrix} c \\ d \end{pmatrix}$$

  con $c \in \mathbb{R}^r$ y $d \in \mathbb{R}^{m-r}$, tenemos

$$\|\Sigma y - U^t b\|_2^2 = \left\| \begin{pmatrix} \sigma_1 y_1 \\ \vdots \\ \sigma_r y_r \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} c \\ d \end{pmatrix} \right\|_2^2 = \left\| \begin{pmatrix} \sigma_1 y_1 - c_1 \\ \vdots \\ \sigma_r y_n - c_r \end{pmatrix} \right\|_2^2 + \|d\|_2^2$$

Entonces el mínimo se alcanza si y sólo si $y = \begin{pmatrix} c_1/\sigma_1 \\ \vdots \\ c_r/\sigma_r \\ y_{r+1} \\ \vdots \\ y_n \end{pmatrix}$ siendo $y_{r+1}, \cdots, y_n \in \mathbb{R}$ arbitrarios. Las infinitas soluciones provienen de la libertad de elección para y_i con $i > r$.

11. Interpolación polinómica

11.1. Problema

Al igual que antes, supongamos que tenemos una muestra de valores $(x_0, y_0), \cdots, (x_n, y_n)$. Queremos encontrar un polinomio P que interpole dichos puntos, es decir, que cumpla $P(x_i) = y_i$ para cada $i = 0, \cdots, n$.

11.2. Polinomio interpolador de Lagrange

Teorema 11.1. *Consideremos $n+1$ puntos $(x_0, y_0), \cdots, (x_n, y_n) \in \mathbb{R}^2$ con $x_i \neq x_j$ si $i \neq j$. Entonces existe un único polinomio $P \in \mathbb{R}[x]$ de grado menor o igual que n tal que $P(x_i) = y_i$ para todo $i = 0, \cdots, n$.*

Demostración. Probemos primero la existencia. Definimos

$$L_{n,k}(x) = \frac{(x-x_0)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)} = \prod_{i\neq k}^{n} \frac{x-x_i}{x_k-x_i}$$

Este polinomio es de grado n y cumple

$$L_{n,k}(x_i) \begin{cases} 1 & \text{si } i = k \\ 0 & \text{si } i \neq k \end{cases}$$

Definimos

$$P(x) = \sum_{k=0}^{n} y_k L_{n,k}(x)$$

Tenemos que $\deg(P) = \deg(\sum_{k=0}^{n} y_k L_{n,k}) \leq \max\limits_{0\leq k\leq n} \deg(y_k L_{n,k}) \leq n$. Además

$$P(x_i) = \sum_{k=0}^{n} y_k L_{n,k}(x_i) = y_i L_{n,i}(x_i) = y_i$$

con lo cual $P \in \mathbb{R}[X]$ es un polinomio que cumple lo deseado.

Veamos la unicidad. Por el absurdo, supongamos que existe un polinomio $Q \in \mathbb{R}_n[X] - \{P\}$ que interpola los $n+1$ puntos. Sea $R = P - Q \neq 0$, que es tal que $\deg(R) \leq \max\{\deg(P), \deg(Q)\} \leq n$, con lo cual tiene a lo sumo n raíces distintas en $\mathbb{R}$. Sin embargo, para cada $i = 0, \cdots, n$, $R(x_i) = P(x_i) - Q(x_i) = 0$, que son $n+1$ valores distintos, lo cual es una contradicción. □

Este polinomio se conoce con el nombre de *polinomio interpolador de Lagrange*. En particular, este polinomio se puede utilizar para aproximar una función $f(x)$ interpolando puntos $(x_0, f(x_0)), \cdots, (x_n, f(x_n))$. Por este motivo, interesa conocer una expresión de la forma

$$f(x) = P(x) + R(x)$$

donde $R(x)$ es el error de la aproximación.

Proposición 11.1. *Sean $x_0, \cdots, x_n \in [a, b]$ distintos. Sea $f \in C^{n+1}([a, b])$. Sea P el polinomio interpolador de Lagrange en $(x_0, f(x_0)), \cdots, (x_n, f(x_n))$. Entonces para todo $x \in [a, b]$ existe $\xi(x) \in (a, b)$ tal que*

$$f(x) = P(x) + \frac{f^{n+1}(\xi(x))}{(n+1)!}(x-x_0)\cdots(x-x_n)$$

Por claridad, cuando tratemos con puntos $x_0, \cdots, x_n$, llamaremos $P_{m_1,\cdots,m_k}$ ($m_i \neq m_j$ si $i \neq j$ y $m_i \in \{0, \cdots, n\}$) al polinomio interpolador de Lagrange en los puntos $x_{m_1}, \cdots, x_{m_k}$.

A continuación damos una fórmula recursiva para calcular un polinomio interpolador de $n+1$ puntos, dados ciertos otros dos polinomios de n puntos.

Lema 11.1. *Sean* $(x_0, y_0), \cdots, (x_n, y_n) \in \mathbb{R}^2$ *con* $x_i \neq x_j$ *si* $i \neq j$. *Entonces, si* $i \neq j$,

$$P_{0,\cdots,n}(x) = \frac{(x - x_j)P_{0,\cdots,j-1,j+1,\cdots,n}(x) - (x - x_i)P_{0,\cdots,i-1,i+1,\cdots,n}(x)}{x_i - x_j}$$

Demostración. Es claro que $P_{0,\cdots,n}$ es un polinomio de grado menor o igual que n. Veamos que interpola los $n + 1$ puntos.

Si $k \neq i, j$, entonces

$$\begin{aligned} P_{0,\cdots,n}(x_k) &= \frac{(x_k - x_j)P_{0,\cdots,j-1,j+1,\cdots,n}(x_k) - (x_k - x_i)P_{0,\cdots,i-1,i+1,\cdots,n}(x_k)}{x_i - x_j} \\ &= \frac{(x_k - x_j)y_k - (x_k - x_i)y_k}{x_i - x_j} \\ &= \frac{(x_i - x_j)y_k}{x_i - x_j} \\ &= y_k \end{aligned}$$

En x_i el polinomio vale

$$\begin{aligned} P_{0,\cdots,n}(x_i) &= \frac{(x_i - x_j)P_{0,\cdots,j-1,j+1,\cdots,n}(x_i) - (x_i - x_i)P_{0,\cdots,i-1,i+1,\cdots,n}(x_i)}{x_i - x_j} \\ &= \frac{(x_i - x_j)y_i}{x_i - x_j} \\ &= y_i \end{aligned}$$

Análogamente, $P_{0,\cdots,n}(x_j) = y_j$.

□

11.3. Diferencias divididas

Planteamos un nuevo problema. Supongamos que se tiene una interpolación de n puntos y se la desea extender con un punto nuevo. La forma del polinomio interpolador vista antes no nos provee una forma de aprovechar el polinomio ya calculado, y nos obliga a computar un polinomio interpolador para $n + 1$ puntos desde cero. Queremos encontrar una forma para el polinomio interpolador que permita agregar secuencialmente nuevos puntos con un menor costo.

Definición 11.1. Sean $x_0, \cdots, x_n$ distintos y f una función real. Se define la diferencia dividida de orden 0 de f respecto de x_i como

$$f[x_i] = f(x_i)$$

Para $k > 0$ se define la diferencia dividida de orden k de f respecto de $x_i, \cdots, x_{i+k}$ como

$$f[x_i, \cdots, x_{i+k}] = \frac{f[x_{i+1}, \cdots, x_{i+k}] - f[x_i, \cdots, x_{i+k-1}]}{x_{i+k} - x_i}$$

La importancia de las diferencias divididas radica en el siguiente resultado:

Proposición 11.2.

$$P_{0,\cdots,n}(x) = a_0 + a_1(x - x_0) + \cdots + a_n(x - x_0)\cdots(x - x_{n-1})$$

con $a_k = f[x_0, \cdots, a_k]$.

Demostración. Hacemos inducción en la cantidad N de puntos interpolados. Si $N = 1$ no hay nada que ver, pues $P_0 = f(x_0) = f[x_0]$. Si $N = 2$ entonces

$$f[x_0] + f[x_0, x_1](x - x_0) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0)$$

y es fácil verificar que este polinomio interpola x_0 y x_1, y como tiene grado menor o igual a 1, entonces, por la unicidad del polinomio interpolador, es exactamente $P_{0,1}$.

Sea $N > 2$ y $n = N - 1$. Supongamos que P es un polinomio que interpola los N puntos $x_0, \cdots, x_n$, y que es de la forma

$$P(x) = P_{0,\cdots,n-1}(x) + \alpha(x - x_0)\cdots(x - x_{n-1})$$

con $\alpha \in \mathbb{R}$ cierta constante fija. Entonces $P(x_i) = P_{0,\cdots,n-1}(x_i) = f(x_i)$ para todo $i = 0, \cdots, n-1$, y $P(x_n) = P_{0,\cdots,n-1}(x_n) + \alpha \prod_{i=0}^{n-1}(x_n - x_i)$. Como $P(x_n) = f(x_n)$ por definición, entonces

$$\alpha = \frac{f(x_n) - P_{0,\cdots,n-1}(x_n)}{\prod_{i=0}^{n-1}(x_n - x_i)}$$

Se puede ver que bajo esta elección de α, el polinomio $P_{0,\cdots,n-1}(x) + \alpha \prod_{i=0}^{n-1}(x - x_i)$ interpola todos los puntos $x_0, \cdots, x_n$. Nuevamente por la unicidad del polinomio interpolador resulta que

$$P_{0,\cdots,n}(x) = P_{0,\cdots,n-1}(x) + \alpha \prod_{i=0}^{n-1}(x - x_i)$$

Por hipótesis inductiva

$$P_{0,\cdots,n-1}(x) = a_0 + a_1(x - x_0) + \cdots + a_{n-1}(x - x_0)\cdots(x - x_{n-2})$$

Luego, basta probar que $\alpha = a_n$. Notemos que α es el coeficiente del monomio x^n en $P_{0,\cdots,n}(x)$.

Por el Lema 11.1,

$$P_{0,\cdots,n}(x) = \frac{(x - x_0)P_{1,\cdots,n}(x) - (x - x_n)P_{0,\cdots,n-1}(x)}{x_n - x_0}$$

De aquí es fácil ver que el coeficiente de x^n está dado por los coeficientes de los monomios x^{n-1} en $P_{1,\cdots,n}(x)$ y $P_{0,\cdots,n-1}(x)$. Por hipótesis inductiva, el coeficiente en $P_{1,\cdots,n}(x)$ es $f[x_1, \cdots, x_n]$ y en $P_{0,\cdots,n-1}(x)$ es $f[x_0, \cdots, x_n]$. Entonces

$$\alpha = \frac{f[x_1, \cdots, x_n] - f[x_0, \cdots, x_n]}{x_n - x_0} = f[x_0, \cdots, x_n] = a_n$$

□

11.4. Interpolación segmentada

Los polinomios tienen una gran desventaja como interpoladores y es que cuanto mayor es su grado, más oscilan. Un procedimiento alternativo consiste en construir, dados $x_0 < \cdots < x_n$, un polinomio interpolador entre cada par consecutivo de puntos x_i y x_{i+1}, y a partir de estos construir una función que interpole todos los puntos. Es decir, construimos $S_0, \cdots, S_{n-1}$ polinomios, tal que S_i interpola x_i y x_{i+1}, y definimos

$$S(x) = \begin{cases} S_0(x) & \text{si } x \in [x_0, x_1] \\ \vdots & \\ S_{n-1}(x) & \text{si } x \in [x_{n-1}, x_n] \end{cases}$$

11.4.1. Lineal

Consiste en interpolar cada par de puntos con un polinomio de grado 1. En otras palabras, S_i es el polinomio interpolador de Lagrange en los puntos x_i y x_{i-1}. La desventaja de este tipo de interpolación es que en los extremos de los subintervalos no hay garantía de que S sea derivable (geométricamente la curva no es suave).

11.4.2. Cuadrática

Utilizamos $S_i(x) = a_i + b_i x + c_i x^2$ para ciertas constantes a_i, b_i y c_i. Si f es la función que estamos aproximando, estas constantes deben ser ajustadas de modo tal que

1. $S(x_i) = f(x_i)$ para todo $i = 0, \cdots, n$.
2. $S_{i+1}(x_{i+1}) = S_i(x_{i+1})$, para todo $i = 0, \cdots, n-2$.
3. $S'_{i+1}(x_{i+1}) = S'_i(x_{i+1})$, para todo $i = 0, \cdots, n-2$.

Las condiciones 1 y 2 aseguran que S esté bien definido y sea contínuo en $[x_0, x_n]$. La condición 3 asegura que sea derivable en (x_0, x_n). Una curva diferenciable definida por partes mediante polinomios se denomina *spline*.

Notemos que en total son $3n-1$ ecuaciones y $3n$ incógnitas, dejando un grado de libertad.

El inconveniente de esta interpolación es que muchas veces se desea fijar condiciones para $S'(x)$ en los extremos x_0 y x_n, y sin embargo no hay constantes suficientes para ello. Los polinomios cúbicos solucionan este problema.

11.4.3. Cúbica

Un spline cúbico para f es una función S que cumple las siguientes condiciones

1. $S(x) = \begin{cases} S_0(x) & \text{si } x \in [x_0, x_1] \\ \vdots & \\ S_{n-1}(x) & \text{si } x \in [x_{n-1}, x_n] \end{cases}$, con S_i un polinomio cúbico.
2. $S(x_i) = f(x_i)$ para todo $i = 0, \cdots, n$.
3. $S_{i+1}(x_{i+1}) = S_i(x_{i+1})$ para todo $i = 0, \cdots, n-2$.
4. $S'_{i+1}(x_{i+1}) = S'_i(x_{i+1})$ para todo $i = 0, \cdots, n-2$.
5. $S''_{i+1}(x_{i+1}) = S''_i(x_{i+1})$ para todo $i = 0, \cdots, n-2$.
6. Se satisface una de las siguientes condiciones frontera:
   - $S''(x_0) = S''(x_n) = 0$ (spline libre o natural).
   - $S'(x_0) = f'(x_0)$ y $S'(x_n) = f'(x_n)$ (spline sujeto).

Las condiciones 2 y 3 aseguran que S esté bien definido y sea contínuo en $[x_0, x_n]$. Las condiciones 4 y 5 aseguran que S sea dos veces derivable y, más aún, como es unión de polinomios, entonces las derivadas son contínuas. Las condiciones de frontera sujeta se utilizan cuando tengo esa información sobre la derivada de la función, y dan lugar a aproximaciones más exactas.

Estudiemos las ecuaciones que determinan las condiciones anteriores. Por conveniencia, vamos a considerar polinomios cúbicos de la forma

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3$$

La condición 2 equivale a $S_i(x_i) = f(x_i)$ y $S_{n-1}(x_n) = f(x_n)$. Como $S_i(x_i) = a_i$ entonces tenemos

$$a_i = f(x_i) \quad \text{para todo } i = 0, \cdots, n-1$$

$$a_{n-1} + b_{n-1}(x_n - x_{n-1}) + c_{n-1}(x_n - x_{n-1})^2 + d_{n-1}(x_n - x_{n-1})^3 = f(x_n)$$

La condición 3 equivale a $S_i(x_{i+1}) = f(x_{i+1})$. Entonces

$$a_i + b_i(x_{i+1} - x_i) + c_i(x_{i+1} - x_i)^2 + d_i(x_{i+1} - x_i)^3 = a_{i+1} \quad \text{para todo } i = 0, \cdots, n-2$$

Observar que $S_i'(x) = b_i + 2c_i(x - x_i) + 3d_i(x - x_i)^2$. Como $S_{i+1}'(x_{i+1}) = b_{i+1}$ entonces la condición 4 equivale a

$$b_i + 2c_i(x_{i+1} - x_i) + 3d_i(x_{i+1} - x_i)^2 = b_{i+1} \quad \text{para todo } i = 0, \cdots, n-2$$

Observar que $S_i''(x) = 2c_i + 6d_i(x - x_i)$. Como $S_{i+1}''(x_{i+1}) = 2c_{i+1}$ entonces la condición 5 equivale a

$$2c_i + 6d_i(x_{i+1} - x_i) = 2c_{i+1} \quad \text{para todo } i = 0, \cdots, n-2$$

Finalmente, si el spline es libre, la condición 6 equivale a

$$2c_0 = 0$$
$$2c_{n-1} + 6d_{n-1}(x_n - x_{n-1}) = 0$$

Definamos $a_n = f(x_n)$, $c_n = 0$ y $h_i = x_{i+1} - x_i$. Entonces las ecuaciones son

1. $a_i = f(x_i)$ para todo $0 \leq i \leq n$.
2. $a_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = a_{i+1}$ para todo $0 \leq i \leq n-1$.
3. $b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1}$ para todo $0 \leq i \leq n-2$.
4. $2c_i + 6d_i h_i = 2c_{i+1}$ para todo $0 \leq i \leq n-1$.
5. $c_0 = 0$.

De la ecuación 4 despejamos

$$d_i = \frac{c_{i+1} - c_i}{3h_i}$$

Sustituyendo esto último y la ecuación 1 en la ecuación 2 y despejando

$$\begin{aligned} b_i &= \frac{1}{h_i}\left(f(x_{i+1}) - f(x_i) - c_i h_i^2 - d_i h_i^3\right) \\ &= \frac{1}{h_i}\left(f(x_{i+1}) - f(x_i) - c_i h_i^2 - \frac{(c_{i+1} - c_i)h_i^3}{3h_i}\right) \\ &= \frac{f(x_{i+1}) - f(x_i)}{h_i} - c_i h_i - \frac{1}{3}(c_{i+1} - c_i)h_i \\ &= \frac{f(x_{i+1}) - f(x_i)}{h_i} - \frac{h_i}{3}\left(2c_i + c_{i+1}\right) \end{aligned}$$

Sustituyendo en la ecuación 3 y despejando

$$
\begin{aligned}
0 &= b_i - b_{i+1} + 2c_i h_i + 3d_i h_i^2 \\
&= \frac{f(x_{i+1}) - f(x_i)}{h_i} - \frac{h_i}{3}\left(2c_i + c_{i+1}\right) - \frac{f(x_{i+2}) - f(x_{i+1})}{h_{i+1}} + \frac{h_{i+1}}{3}\left(2c_{i+1} + c_{i+2}\right) \\
&+ 2c_i h_i + 3\frac{c_{i+1} - c_i}{3h_i} h_i^2 \\
&= \left[\frac{f(x_{i+1}) - f(x_i)}{h_i} - \frac{f(x_{i+2}) - f(x_{i+1})}{h_{i+1}}\right] - \frac{2}{3}h_i c_i - \frac{1}{3}h_i c_{i+1} + \frac{2}{3}h_{i+1}c_{i+1} + \frac{1}{3}h_{i+1}c_{i+2} \\
&+ 2h_i c_i + h_i c_{i+1} - h_i c_i \\
&= \left[\frac{f(x_{i+1}) - f(x_i)}{h_i} - \frac{f(x_{i+2}) - f(x_{i+1})}{h_{i+1}}\right] + \frac{1}{3}h_i c_i + \frac{2}{3}(h_i + h_{i+1})c_{i+1} + \frac{1}{3}h_{i+1}c_{i+2}
\end{aligned}
$$

Equivalentemente

$$
h_i c_i + 2(h_i + h_{i+1})c_{i+1} + h_{i+1}c_{i+2} = 3\left[\frac{f(x_{i+2}) - f(x_{i+1})}{h_{i+1}} - \frac{f(x_{i+1}) - f(x_i)}{h_i}\right]
$$

Esta ecuación, que vale para $i = 0, \cdots, n-2$, contiene toda la información de las demás (pues la obtuvimos a través de sustituciones sucesivas). Juntando estas $n-1$ ecuaciones con $c_0 = 0$ y $c_n = 0$ tenemos un sistema de $n+1$ ecuaciones y $n+1$ incógnitas

$$
\begin{pmatrix}
1 & 0 & 0 & 0 & & & & \\
h_0 & 2(h_0 + h_1) & h_1 & 0 & & & & \\
0 & h_1 & 2(h_1 + h_2) & h_2 & & & & \\
 & & & & \ddots & & & \\
 & & & & & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\
 & & & & & 0 & 0 & 1
\end{pmatrix}
$$

La matriz del sistema es estrictamente diagonal dominante (pues $h_i = x_{i+1} - x_i > 0$), por lo tanto es inversible. Luego, la solución es única, i.e., existe un único spline cúbico natural para $x_0, \cdots, x_n$. Más aún, como la matriz es tridiagonal puede ser almacenada y operada eficientemente. En particular, el costo de la eliminación gaussiana sobre esta matriz es $\mathcal{O}(n)$.

Procediendo en forma análoga se puede demostrar que si el spline cúbico es sujeto también es único. Más aún, también se obtiene un sistema tridiagonal estrictamente diagonal dominante.

12. Integración numérica

12.1. Problema

Dada una función $f : [a, b] \to R$, queremos calcular $\int_a^b f(x)dx$. Conociendo una primitiva F, entonces por la Regla de Barrow podemos calcular la integral y vale $\int_a^b f(x)dx = F(b) - F(a)$. En la mayoría de los casos calcular una primitiva de f es muy difícil o no es posible. Es por esto que queremos encontrar métodos numéricos que permitan aproximar la integral.

Dado que los polinomios se pueden integrar fácilmente y que además sabemos cómo aproximar una función via polinomios, vamos a utilizar la escritura

$$f(x) = P_n(x) + E_n(x)$$

donde P_n es el polinomio interpolador de Lagrange en $n + 1$ puntos en el intervalo $[a, b]$ y E_n es el error de la aproximación. Integrando:

$$\int_a^b f(x)dx = \int_a^b P_n(x)dx + \int_a^b E_n(x)dx$$

Estas fórmulas se llaman *fórmulas de cuadraturas* (geométricamente cuadran el área debajo de la curva).

12.2. Regla del trapecio ($n = 1$)

Tomamos $x_0 = a$, $x_1 = b$ y consideramos P_1 que interpola x_0 y x_1. Sea $h = b-a$ la longitud del intervalo de integración. Entonces

$$\int_a^b P_1(x)dx = \frac{h}{2}(f(x_0) + f(x_1))$$

$$\int_a^b E_1(x)dx = -\frac{h^3}{12}f''(\mu)$$

con $\mu \in (a, b)$.

12.3. Regla de Simpson ($n = 2$)

Tomamos $x_0 = a$, $x_1 = \frac{a+b}{2}$, $x_2 = b$ y consideramos P_2 que interpola x_0, x_1 y x_2. Sea $h = x_1 - x_0 = x_2 - x_1 = \frac{b-a}{2}$ la longitud de los subintervalos. Entonces

$$\int_a^b P_2(x)dx = \frac{h}{3}(f(x_0) + 4f(x_1) + f(x_2))$$

$$\int_a^b E_2(x)dx = -\frac{h^5}{90}f^{(4)}(\mu)$$

con $\mu \in (a, b)$.

12.4. Grado de precisión

Definición 12.1. Se llama grado de precisión de una fórmula de cuadratura al máximo entero positivo n tal que la fórmula es exacta para todo polinomio de grado menor o igual a n.

El grado de precisión se puede deducir de la fórmula del error de los métodos. En el caso de trapecio, el error involucra una derivada segunda, con lo cual el grado de precisión es 1 (es exacta para todo polinomio de grado 0 o 1 pero existen polinomios de grado 2 para los cuales no lo es). En el caso de Simpson, el grado de precisión es 3.

12.5. Reglas compuestas

Al intervalo de integración lo dividimos en subintervalos y en cada uno de ellos aplicamos alguno de los métodos conocidos.

12.5.1. Regla compuesta del trapecio

Si dividimos al intervalo $[a, b]$ en n subintervalos, entonces cada uno tendrá longitud $h = \frac{b-a}{n}$. La aproximación es ahora

$$\frac{h}{2}\left(f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)\right)$$

El error es

$$-\frac{b-a}{12}h^2 f''(\mu)$$

con $\mu \in (a, b)$.

12.5.2. Regla compuesta de Simpson

Recordemos que Simpson utilizaba tres puntos del intervalo para aproximar. Entonces, si dividimos al intervalo $[a, b]$ en n subintervalos, la regla de Simpson se aplicará sobre cada par consecutivo de ellos. Por lo tanto, necesitamos una cantidad n de subitervalos par. La aproximación es

$$\frac{h}{3}\sum_{k=0}^{n/2-1} (f(x_{2k}) + 4f(x_{2k+1}) + f(x_{2k+2}))$$

El error es

$$-\frac{b-a}{180}h^4 f^{(4)}(\mu)$$

con $\mu \in (a, b)$.

12.6. Métodos adaptativos

Supongamos que queremos integrar una función cuyo comportamiento es irregular. En cierto subintervalo, la función tiene una gran variación, lo cual obliga a utilizar una aproximación con una partición fina del subintervalo sobre la cual utilizar una regla compuesta. Sin embargo en otro subintervalo disjunto, la función tiene una variación muy pequeña, haciéndola apta para un método de aproximación sin demasiado refinamiento.

En este tipo de situaciones se utilizan métodos adaptativos, que analizan en cada subintervalo cuál es la precisión de una aproximación de la integral y en caso de no ser suficiente, utilizan una aproximación más fina partiendo en otros subintervalos.

Estudiemos el método basado en la regla de Simpson compuesta. Llamemos $S(x, y)$ a la aproximación de Simpson del intervalo $[x, y]$ para la función f. Supongamos que queremos integrar el intervalo $[a, b]$.

`Paso 1:` Tomamos dos subintervalos, cada uno de tamaño $h = \frac{b-a}{2}$, aplicando Simpson, obteniendose

$$\int_a^b f(x)dx = S(a, b) - \frac{h^5}{90}f^{(4)}(\mu)$$

`Paso 2:` Partimos cada subintervalo en otros dos de tamaño $\frac{h}{2}$. Aplicamos la regla compuesta de simpson en $\left[a, \frac{a+b}{2}\right]$ y $\left[\frac{a+b}{2}, b\right]$, obteniendose

$$\int_a^b f(x)dx = S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right) - \frac{1}{180}\left(\frac{h}{2}\right)^4 (b-a) f^{(4)}(\tilde{\mu})$$

Como $h = \frac{b-a}{2}$, el resto se puede reescribir como

$$-\frac{1}{16}\frac{h^5}{90} f^{(4)}(\tilde{\mu})$$

Supongamos que $f^{(4)}(\mu) \approx f^{(4)}(\tilde{\mu})$, entonces si igualamos las expresiones obtenidas en los pasos 1 y 2:

$$\begin{aligned}
&\int_a^b f(x)dx = S(a,b) - \frac{h^5}{90} f^{(4)}(\mu) = \int_a^b f(x)dx = S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right) - \frac{1}{16}\frac{h^5}{90} f^{(4)}(\mu) \\
&\Leftrightarrow -\frac{15}{16}\frac{h^5}{90} f^{(4)}(\mu) = S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right) - S(a,b) \\
&\Leftrightarrow -\frac{1}{16}\frac{h^5}{90} f^{(4)}(\mu) = \frac{1}{15}\left(S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right) - S(a,b)\right)
\end{aligned}$$

Es decir que el error al subdividir los intervalos es

$$\frac{1}{15}\left(S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right) - S(a,b)\right)$$

Si este valor es suficientemente chico, concluyo la subdivisión. En caso contrario, procedo recursivamente, volviendo al paso 1, sobre los intervalos $\left[a, \frac{a+b}{2}\right]$ y $\left[\frac{a+b}{2}, b\right]$

13. Referencias

- Clases de la materia Métodos Numéricos dictadas por la Dra. Isabel Mendez Diaz.
- R. Burden y J.D.Faires, *Análisis numérico*, International Thomson Editors, 1998.
- R. Duran, S. B. Lasalle y J. D. Rossi, *Elementos del Cálculo Numérico.*
- Apuntes de `www.cubawiki.com.ar`.
  - `http://www.cubawiki.com.ar/images/6/61/Metnum_apunte_jsackmann.pdf`.
  - `http://www.cubawiki.com.ar/images/7/7d/Metnum_overview.pdf`.
- D. Goldberg, *What Every Computer Scientist Should Know About Floating-Point Arithmetic*, ACM Computing Surveys, Volumen 23 Issue 1, marzo 1991, páginas 5-48.