Universidad Autónoma Metropolitana - *Iztapalapa*

Coloquio
del Departamento de Matemáticas

División de Ciencias Básicas e Ingenieria

Problemas inversos: deconvolución de imágenes

Mario Gerardo Medina Valdez

Taxco de Alarcón, Guerrero
Enero del 2009

2do Coloquio del Departamento de Matemáticas

Problemas inversos: aspectos matemáticos y computacionales de la deconvolución

Mario Medina Valdéz

Comité Organizador

Mario Pineda Ruelas
Roberto Quezada Batalla
Blanca Rosa Pérez Salvador
Luis Aguirre Castillo
Daniel Espinosa
Constancio Hernández García
Michaell Rivera Arce (Apoyo logístico)

Material didáctico del Departamento de Matemáticas de la Universidad Autonóma Metropolitana-Iztapalapa con el apoyo del Programa Integral de Fortalecimiento Institucional 3.3 (PIFI).

Problemas inversos: aspectos matemáticos y computacionales de la deconvolución

Mario Medina Valdéz

Departamento de Matemáticas, UAM-I

Casa abierta al tiempo

Universidad Autónoma Metropolitana

Contenido

Capítulo 1

Convolución y deconvolución

En muchas áreas de las ciencias e ingenierías es común el manejo de imágenes. Por citar un ejemplo, en la ingeniería biomédica una imagen puede ser obtenida a través de una tomografía computarizada o por rayos X. A través de estas imágenes el especialista médico espera obtener información que le permita detectar, por ejemplo, la existencia de tumores cancerígenos. En astronomía, las imágenes que podemos apreciar en periódicos, revistas de divulgación científica o especializadas en general no son las imágenes que obienen directamente los especialistas, sino que estas son mejoradas por procedimientos algorítmicos matemáticos para lograr las imágenes que apreciamos en tales medios y que alguna vez nos han producido una grata sorpresa.

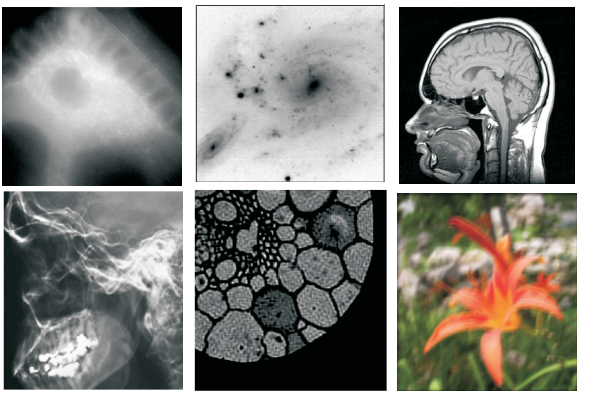

Figura 1.1. Distintas funciones de dispersión del punto.

Un caso muy conocido fueron las imágenes obtenidas por el observatorio espacial Hubble. Las primeras imágenes obtenidas estaban degradadas, eran imágenes borrosas. Esto era efecto de una aberración de los lentes del telescopio. Debido a la imposibilidad de corregir el desperfecto de los lentes se volvió necesario buscar la manera de mejorar la calidad de las imágenes. Fué imporante el conocimeinto

del sistema óptico del telescopio, así con este conocimiento y con las imágenes borrosas fué posible mejorar lacalidad de las imágenes. En muchas aplicaciones el conocimiento que se tiene de las causas que producen la degradadción de la o las imágenes es limitado o prácticamente nulo, por lo que el problema de recuperar lo que serían las imágenes originales se vuelve un problema más complejo.

El siguiente ejemplo nos muestra de manera sencilla una de las complicaciones que surgen al resolver los llamados problemas inversos. Pequeñas variaciones en los datos pueden producir grandes alteraciones en las soluciones.

EJEMPLO 1.0.1. Si consideramos el sistema de ecuaciones lineales

$$\begin{aligned} .780x + .563y &= .217 \\ .457x + .330y &= .127 \end{aligned} \tag{1.1}$$

cuya forma general está dada por

$$A\mathbf{x} = \mathbf{b}.$$

Si calculamos la solución por métodos clásicos con aproximación a tres dígitos en la aritmética, entonces la solución resulta ser $x = 1.71$ y $y = -1.98$. Al sustituir esta solución aproximada en el sistema (1.1) obtenemos

$$\begin{aligned} .780 * (1.71) + .563 * (-1.98) - .217 &= 0.00206 \\ .457 * (1.71) + .330 * (-1.98) - .217 &= 0.00107 \end{aligned}$$

El residual de la solución ($\|A\mathbf{x} - \mathbf{b}\|$) aproximada que se ha calculado pudiera parecer pequeño. Pero, por otra parte, la solución exacta del sistema de ecuaciones lineales está dada por $x = 1$ y $y = -1$, como puede verificar fácilmente el lector. ¿Qué es lo que ocurre en este ejemplo? La solución "aproximada" que hemos hallado se encuentra "lejos" de l solución exacta del sistema de ecuaciones lineales. ¿Hay algún problema con el algoritmo usado para calcular la solución aproximada?. No, el problema no se encuentra en el algoritmo sino en la naturaleza misma del sistema de ecuaciones lineales, el cual es un sistema *mal condicionado*, lo que podemos decir al respecto es que pequeños cambios en los coeficientes del sistema producirán cambios sensibles en las soluciones. Lo que este ejemplo muestra es que no podemos esperar que un algoritmo dado funcione correctamente en un problema que resulta ser mal condicionado.

En estas notas daremos una introducción a un tema fascinante, los problemas inversos, en particular al problema de la recuperación de imágenes degradadas. Esta degradación puede manifestarse a través de la obención de una imagen borrosa, la cual puede presentar o no ruido. Estos dos tipos de distorsión de una imagen (ver Figura 1), el

que se mamifiesta por el hecho de que la imagen se vea borrosa se debe a un proceso determinístico asociado a las llamadas funciones de dispersión del punto que a su vez se encuentran asociadas al mecanismo de obtención de la image (e.g. cámara fotográfica, telescopio o microscopio), mientras que la distorsión de la imagen debida al ruido se debe a una distorsión estocástica asociado al mecanismo mediante el cual una imagen se almacena. Presentaremos algunos aspectos matemáticos asociados a este problema en particular y a los problemas inversos en general.

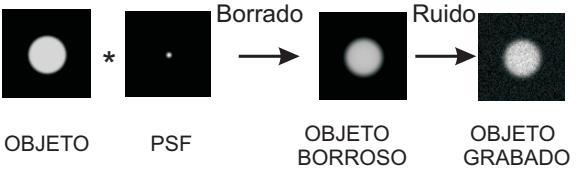

Figura 1.2. Mecanismo de obtención de una imagen degradada.

Con el fin de simplificar un poco la idea de como se puede lograr disminuir la degradación de una imagen, ver Figura 1, supongamos que esta no presenta ruido. Como se ha dicho, el degradado está asociado a una función de dispersión del punto. Para atenuar la degradación se hace uso de la función de dispersión del punto y de un proceso llamado regularización. En muchos casos, en la práctica, el conocimiento que se tiene de la función de dispersión del punto es poca o prácticamente nula, por lo que el mejorar la calidad de la image obtenida se vuelve un problema mucho más complejo. Es de notar que en los últimos años este problema de recuperación de una imagen degradada ha recibiodo gran atención de parte de la comuniad científica, en gran parte por las innumerables aplicaciones que presenta y por otra debido al enorme desarrollo del hardware computacional.

1.1. Integral de Fredholm de primera especie

Dadas dos funciones $h, f : \mathbb{R} \to \mathbb{R}$ se define su convolución como la función dada por

$$g(s) = (h * f)(s) = \int_{-\infty}^{\infty} h(s-t)f(t)dt, \quad s \in \mathbb{R} \tag{1.2}$$

PROPOSICIÓN 1.1.1. *La convolución satisface las siguientes propiedades*

(1) $(h * f)(s) = (f * h)(s)$, *(conmutatividad).*
(2) $((h_1 * f_1) * f_2)(s) = (h * (f_1 * f_2))(s)$, *(asociatividad)*

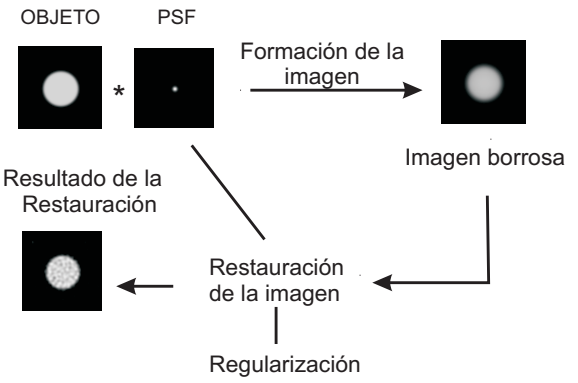

Figura 1.3. Proceso de restauración de una imagen.

(3) $(h * (f_1 + f_2))(s) = (h * f_1)(s) + (h * f_2)(s)$, *(distributividad).*

EJERCICIO 1.1. La prueba de estas propiedades se deja como ejercicio al lector.

Observemos que la primera de estas propiedades nos muestra que, desde un punto de vista matemático, las funciones f_1 y f_2 desempeñan el mismo papel, pero desde el punto de vista de las aplicaciones, las interpretaciones de ambas, como veremos son distintas.

Un problema asociado a la integral de convolución y que llamaremos *problema directo* es el de calcular la convolución de dos funciones dadas En estas notas estaremos interesados en los llamados *Problemas de Deconvolución*, uno de ellos consiste en determinar la función $f(t)$ a partir de las funciones $g(s)$ y $h(t)$. Otro problema de deconvolución, llamado *deconvoluión ciega*, consiste en determinar las funciones $h(t)$ y $f(t)$ a partir solo de la función $g(s)$. Estos son problemas inversos asociados a la integral de convolución (1.2).

Muchos problemas pueden modelarse como problemas de deconvolución, como veremos a continuación.

EJEMPLO 1.1.2 (Problema inverso de calor). Supongamos que se desea calcular la temperatura f (como función del tiempo t) de una cara inaccesible de una pared solo a partir de mediciones de temperatura realizadas en la cara accesible de la pared. En este caso la función h de la ecuación (1.2) está dada por

$$h(s-t) = \frac{(s-t)^{-\frac{3}{2}}}{2\kappa\sqrt{\pi}} \exp(-\frac{1}{4\kappa(s-t)}),$$

donde el parámetro κ está asociado a las características de conducción de calor de la pared.

EJEMPLO 1.1.3 (Diferenciación). Consideremos una función $f \in \mathcal{C}^1[0,1]$ diferenciable, con derivada continua, un parámetro $\delta \in (0,1)$ y cualquier número natural n mayor que 1. A partir de f podemos definir una función $f_{\delta,n}$, cercana a f en la norma $\|\cdot\|_\infty$. Tomemos

$$f_{\delta,n}(x) = f(x) + \delta \sin(\frac{nx}{\delta})$$

donde $x \in [0,1]$.

Entonces

$$f'_{\delta,n}(x) = f'(x) + n\cos(\frac{nx}{\delta}).$$

Además,

$$\|f - f_{\delta,n}\|_\infty = \|\delta \sin(\frac{nx}{\delta})\|_\infty = \delta$$

y

$$\|f' - f'_{\delta,n}\|_\infty = \|n\cos(\frac{nx}{\delta})\|_\infty = n.$$

Observemos que la derivada no depende continuamente de los datos con respecto a la norma uniforme. La derivada f' resuelve la ecuación integral

$$(F(x))(s) := \int_0 tx(s)ds = f(t) - f(0)$$

debido al Teorema Fundamental del cálculo, ecuación soluble en $\mathcal{C}[0,1]$ solo si $f \in \mathcal{C}^1[0,1]$.

En este ejemplo, el problema directo es calcular f a partir de x; es decir, debemos integrar. Lo que se puede observar es que la integración suaviza los errores altamente oscilatorios. '¿Qué es lo que hace que uan función pueda deivarse? Debemos excluir los errores en los datos de frecuencia arbtrariamente grandes, lo que se puede lograr si conocemos una cota para la segunda derivada de f. Lo que podemos observar de la gráficas es que el operador F es un operador lineal, inyctivo y continuo cuya inversa es no acotada. Pero si restringimos el operador al conjunto

$$\left\{x \in \mathcal{C}^1[0,1] \mid \|x'\|_\infty + \|x'\|_\infty \leq c\right\}$$

el cual puede mostrarse que es un conjunto compacto, en tal caso la inversa sería continua. En vase a esto es que si deseamos restaurar la *estabilidad* podríamos suponer cotas a priori para la primera y segunda derivadas de f.

Sean f una función cualquiera, la cual deseamos derivar, y f_δ una versión de f la cual contiene ruido de manera que

$$\|f - f_\delta\|_\infty \leq \delta$$

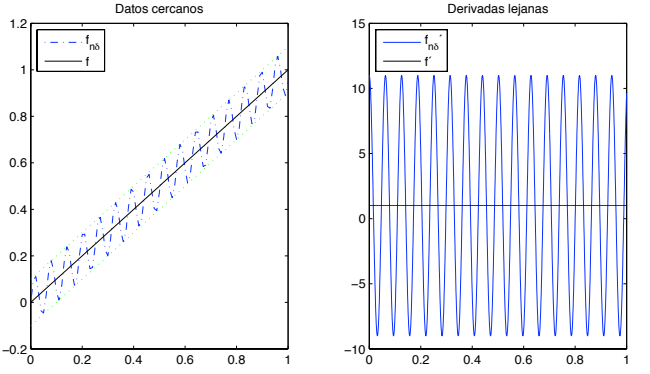

Figura 1.4. Alta frecuencia

Si usamos deferencias centradas con tamaño de paso h y si $f \in \mathcal{C}^2[0,1]$, la aproximación mediante la serie de Taylor nos da la expansión

$$\frac{f(x+h) - f(x-h)}{2h} = f'(x) + O(h).$$

En caso de que $f \in \mathcal{C}^3(x)$ la expansión está dada por la expresión

$$\frac{f(x+h) - f(x-h}{2h} = f'(x) + O(h^2).$$

De esta manera, la precisión en el método de diferencias centradas depende de la suavidad de los datos exactos. Asimismo, en lugar de calcular f' calculamos

$$\frac{f_\delta(x+h) - f_\delta(x-h)}{2h} \sim \frac{f(x+h) - f(x-h)}{2h} + \frac{\delta}{h}$$

De esta forma, el error total se comporta como

$$O(h^\nu) + \frac{\delta}{h}$$

donde ν es igual a 1 o 2, si $f \in \mathcal{C}^2[0,1]$ o $f \in \mathcal{C}^3[0,1]$.

Para un nivel fijo de error δ tendríamosque si h es cada vez más pequeño entonces el error total aumenta debido al error, si h fuese demasiado grande, entonces la aproximación del error sería demasiado grande. Hay un parámetro h_0 de discretización óptima.

A coninuación estimaremos el comportamiento asintótico de h_0: Si elegimos al parámetro h como potencia de δ, digamos δ^k entonces es posible minimizar el término $O(h^\nu) + \frac{\delta}{h}$ tomando $k = \frac{1}{3}$ o $k = \frac{1}{2}$. Con estas elecciones, el error total es del orden $O(\sqrt(\delta))$ o $O(\sqrt(\delta^{3/2}))$, para $f \in \mathcal{C}^2([0,1])$ o $f \in \mathcal{C}^3([0,1])$, respectivamente. Aún en el mejor caso

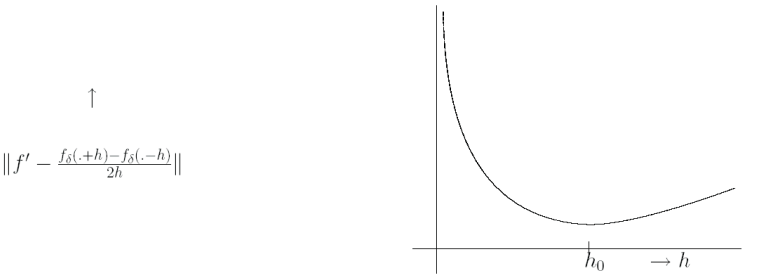

Figura 1.5. Derivadas

posible (observemos que en la condición $O(h^\nu) + \frac{\delta}{h}$ no es posible tomar $\nu > 2$), con un valor óptimo de h la m,ejor convergencia que obtenemos es del orden de $O(\delta^{3/2})$, donde δ denota el error en los datos. En suma, hay una pérdida intrínseca de información.

Otros aspectos que podemos resaltar de este ejemplo es que muestra varios de las características que son típicas de los *problemas mal condicionados.*

- Presenta una amplificación de los errores con frecuencias altas.
- Es posible restaurar la estabilidad haiendo uso de información a priori.
- Presenta dos errores de naturaleza distinta, uno debido al error de aproximación y el otro debido a la propagación de los errores de los datos.
- Presenta un valor óptino de un parámetro de discretización cuya elección depende solo de información a priori.
- Aún en circunstancias óptimas hay pérdida de información.

Los ejemplos anteriores son casos especiales de un tipo especial de ecuaciones integrales, llamadas integrales de Fredholm de primera especie.

DEFINICIÓN 1.1.4. Una integral de Fredholm de primera especie es una integral de la forma

$$\int_a^b K(s,t)f(t)dt = g(s),\ s \in [c,d]. \tag{1.3}$$

La función K es llamado el núcleo de la ecuación integral y es una función de las variables s y t.

En el caso de problemas de deconvolución la función g es una función conocida o de la cual se tienen mediciones en una muestra

de valores discretos de s. La función K también es conocida, y está asociada al modelo mateático del problema subyaente.

Observemos que los problemas de deconvolución citadas anteriormente no son sino casos especiales de una ecuación integral de Fredholm de primera especie dond el núcleo K depende de la diferencia $s - t$, es decir,

$$K = K(s, t) = K(s - t).$$

1.2. Modelo discreto de una integral de Fredholm

Al considerar el problema de aproximar numéricamente una integral se pueden considerar distintos métodos de cuadratura: punto medio, trapecio, Simpson, solo por citar algunas. En general, una integral

$$\int_a^b \phi(t)dt$$

puede aproximarse por una sumatoria de la forma

$$\sum_{l=1}^{n} \omega_l \phi(t_l),$$

donde los coeficientes ω_l son pesos para cada valor $\phi(t_l)$ y dependen del método de cuadratura usado. En el caso del método del trapecio sabemos que el área del trapecio mostrado en la figura 1.2 está dada por

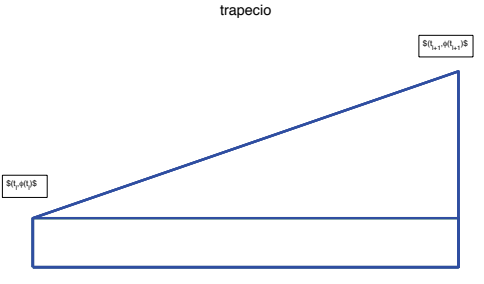

Figura 1.6. Area de un trapecio

$$\begin{aligned} Area &= \frac{1}{2}(base) \times (altura) \\ &= (t_{l+1} - t_l)\phi(t_l) + \frac{1}{2}(t_{l+1} - t_l)\phi(t_{l+1}) \\ &= \frac{1}{2}(\phi(t_l) + \phi(t_{l+1}))(t_{l+1} - t_l) \\ &= \frac{1}{2}(\phi(t_l) + \phi(t_{l+1}))h, \end{aligned} \tag{1.4}$$

donde $h = \frac{b-a}{n-1} = t_{l+1} - t_l$, $l = 1, 2, \ldots, n-1$. De esta forma la integral se puede aproximar como

$$\begin{aligned} \int_a^b \phi(t)dt &\simeq \sum_{l=1}^{n-1} \frac{1}{2}(\phi(t_l) + \phi(t_{l+1}))h \\ &= \frac{1}{2}\phi(t_1)h + \phi(t_2)h + \cdots + \phi(t_{n-1})h + \frac{1}{2}\phi(t_n)h. \end{aligned} \tag{1.5}$$

Po lo que,

$$\omega_l = \begin{cases} \frac{1}{2}h, & \text{si } l = 1, n; \\ h, & \text{si } l = 2, 3, \ldots, n-1. \end{cases}$$

Con esta herramienta obtendremos una aproximación al problema de deconvolución de una ecuación integral de Fredholm de primera especie (1.3). Para ello tomemos una partición $a = t_1 < t_2 < \cdots < t_n = b$ del intervalo $[a, b]$, donde $t_l = a + (l-1)\frac{b-a}{n-1}$. Entonces, para cada $s \in [c, d]$, el valor $g(s)$ se aproxima numéricamente mediante

$$g(s) \simeq \sum_{l=1}^{n-1} w_l K(s, t_l) f(t_l)$$

Si además contamos con información de la función $g(s)$ en una cantidad finita de puntos del intervalo $[c, d]$, digamos en los puntos $c = s_1 < s_2 < \cdots < s_m = d$ del intervalo $[c, d]$, donde $s_k = c + (k-1)\frac{d-c}{m-1}$, entonces para cada $k = 1, 2, \ldots, m$

$$g(s_k) \simeq \sum_{l=1}^{n-1} w_l K(s_k, t_l) f(t_l).$$

Por lo que obtenemos un sistema de ecuaciones lineales

$$\begin{pmatrix} \omega_1 K(s_1, t_1) & \omega_2 K(s_1, t_2) & \ldots & \omega_n K(s_1, t_n) \\ \omega_1 K(s_2, t_1) & \omega_2 K(s_2, t_2) & \ldots & \omega_n K(s_2, t_n) \\ \vdots & \vdots & \ddots & \vdots \\ \omega_1 K(s_m, t_1) & \omega_2 K(s_m, t_2) & \ldots & \omega_n K(s_m, t_n) \end{pmatrix} \begin{pmatrix} f(t_1) \\ f(t_2) \\ \vdots \\ f(t_n) \end{pmatrix} = \begin{pmatrix} g(s_1) \\ g(s_2) \\ \vdots \\ g(s_m) \end{pmatrix}.$$

Si definimos

$$A = \begin{pmatrix} \omega_1 K(s_1,t_1) & \omega_2 K(s_1,t_2) & \dots & \omega_n K(s_1,t_n) \\ \omega_1 K(s_2,t_1) & \omega_2 K(s_2,t_2) & \dots & \omega_n K(s_2,t_n) \\ \vdots & \vdots & \ddots & \vdots \\ \omega_1 K(s_m,t_1) & \omega_2 K(s_m,t_2) & \dots & \omega_n K(s_m,t_n) \end{pmatrix},$$

$$\mathbf{x} = \begin{pmatrix} f(t_1) \\ f(t_2) \\ \vdots \\ f(t_n) \end{pmatrix},$$

y

$$\mathbf{b} = \begin{pmatrix} g(s_1) \\ g(s_2) \\ \vdots \\ g(s_m) \end{pmatrix},$$

entonces tenemos la ecuación

$$A\mathbf{x} = \mathbf{b}.$$

En el caso de considerar la regla del trapecio, la matriz A está dada por

$$A = \begin{pmatrix} \frac{1}{2}hK(s_1,t_1) & hK(s_1,t_2) & \dots & hK(s_1,t_{n-1}) & \frac{1}{2}hK(s_1,t_{n-1}) \\ \frac{1}{2}hK(s_2,t_1) & hK(s_2,t_2) & \dots & hK(s_2,t_n) & \frac{1}{2}hK(s_1,t_{n-1}) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{2}hK(s_m,t_1) & hK(s_m,t_2) & \dots & hK(s_m,t_{n-1}) & \frac{1}{2}hK(s_m,t_{n-1}) \end{pmatrix},$$

De esta forma tenemos que la versión discreta del problema de deconvolución asociado a la ecuación (1.3) está dada por el sistema de ecuaciones lineales

$$A\mathbf{x} = \mathbf{b}.$$

EJERCICIO 1.2. Utilice el método de cuadratura conocido como el método del punto medio para determinar la versión discreta del problema de deconvolución donde se tome la misma cantidad de muestras para la funcoión g igual a la cantidad de muestras que se desean ontener para la función f y obtenga la matriz A correspondiente, muestre que tal matriz es una matriz simétrica.

EJERCICIO 1.3. Considere el conjunto $\mathcal{C}[0,1]$ de funciones continuas definidas en el inervalo $[0,1]$. Dado un núcleo continuo $K : I \times I \to \mathbb{R}$ y un elemento de $\phi \in \mathcal{C}[0,1]$ desarrolle un programa en Matlab que

calcule la integral

$$\Gamma(\phi)(s) = \int_0^1 K(s,t)\phi(t)dt, s \in [0,1].$$

usando la regla trapezoidal, rectangular o de punto medio.

1.2.1. Derivación. Problema Directo: Dada una función continua $x \in \mathcal{C}[0,1]$, determinar su antiderivada y tal que $y(0) = 0$. Es decir, hallar

$$y(t) = \int_0^t x(s)ds, \qquad t \in [0,1]$$

Problema inverso: Dada una función $y \in \mathcal{C}^1[0,1]$ tal que $y(0) = 0$, encontrar su derivada $x = y'$.

Esto lo podemos interpretar como el resolver la ecuación integral $Kx = y$, donde K es el operador $K : \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$ definido por

$$(Kx)(t) := \int_0^t x(s)ds, \qquad t \in [0,1], \quad x \in \mathcal{C}([0,1]).$$

En este problema consideramos la norma del supremo en $(\mathcal{C}([0,1]), \|\cdot\|)$, es decir, $\|x\| = \max\{|x(t)||t \in [0,1]\}$.

La solución de la ecuación $Kx = y$ está dada por la antiderivada $x = y'$, siempre que $y(0) = 0$ y la función y sea de clase $\mathcal{C}^1$. Supongamos que x es la solucioón exacta de la ecuación $Kx = y$, y consideremos una perturbación $\tilde{y}$ de y en $(\mathcal{C}([0,1]), \|\cdot\|_\infty)$, esta perturbación no necesariamente es una función diferenciable. Más aún, la solución del problema perturbado no necesariamente está cercana a la solución del problema original.

Para ver esto, tomemos $\tilde{y}(t) = y(t) + \delta \sin(t/\delta^2)$, con δ suficientemente pequeño.

1.3. Deconvolución Bidimensional

Este tema tiene incidencia directa en el manejo y procesamiento de imágenes, donde estas son representadas por matrices de dimensiones grandes. De la misma forma como lo hicimos en el caso en dimensión uno. Consideraremos la discretización de la convolución de dos funciones de dos variables.

En el caso bidimensional una integral de Fredholm toma la forma

$$g(x,y) = \int_a^b \int_c^d K(x,y,x',y')f(x',y')dx'dy'. \tag{1.6}$$

Consideremos el caso especial, en el cual el kernel K toma la forma

$$K(x,y,x',y') = h(x-x')k(y-y')$$

es decir, el kernel (PSF) separa las variables $x - x'$ y $y - y'$. De esta forma la integra doble de la convolución bidimensional (1.6), debido al Teorema de Fubini puede reescribirse como

$$\begin{aligned} g(x,y) &= \int_a^b \int_c^d h(x-x')k(y-y')f(x',y')dx'dy' \\ &= \int_a^b h(x-x')\left\{\int_c^d k(y-y')f(x',y')dy'\right\}dx'. \end{aligned} \tag{1.7}$$

De manera semejante a como lo hicimos en el caso unidimensional, es posible discretizar el caso de dos variables. Paa esto es necesario usar alguno de los métos de aproximación de una integal; por ejemplo, podemos usar el método del punto medio. Hacemos uso de este puesto que es el más sencillo de ellos.

Primero consideremos la discretización de la integral interior $\int_c^d k(y-y')f(x',y')dy'$. Para esto, tomemos una partición del intervalo $[c,d]$: $c = y'_0 < y'_1 < \cdots < y'_n = d$ donde $y'_{i+1} - y'_i = \frac{d-c}{n}$, con $i = 0,\ldots,n-1$. Sea $\overline{y'_i} = \frac{y'_{i+1}-y'_i}{2}$ el punto medio del intervalo $[y'_i, y'_{i+1}]$, $i = 0,\ldots,n-1$. En consecuencia, la integral mencionada puede aproximarse por la expresión

$$\begin{aligned} \sum_{i=0}^{n} k(y-\overline{y'_i})f(x',\overline{y'_i})(y'_{i+1}-y'_i) &= \sum_{i=0}^{n} k(y-y'_i)f(x',y'_i)(y'_{i+1}-y'_i) \\ &= \frac{1}{n}\sum_{i=0}^{n} k(y-y'_i)f(x',y'_i)(d-c) \end{aligned}$$

En microscopía confocal surge ejemplo de una convolución bidimensional. Esta técnica de microscipía arroja mejores resultados que los obtenidos por micoscopía de luz convencional. En este mejoramiento se considera que el objeto se encuentra iluminado uniformemente y que el lente que colecta la luz es simplemente una apertura de radio d, por lo que el objeto bidimensional f está relacionado con la imagen g a través de una convolución unidimensional donde el núcleo K se separa como producto de las funciones

$$h(z) = k(z) = \frac{\sin(\pi dz)}{\pi dz}.$$

Este tipo de núcleos separables son típicos en problemas de restauración de imágenes.

1.4. Funciones de dispersión del punto

Hemos visto que los núcleos, que reciben el nombre de *funciones de dispersión del punto* (PSF, Point Spread Functions en inglés) de una

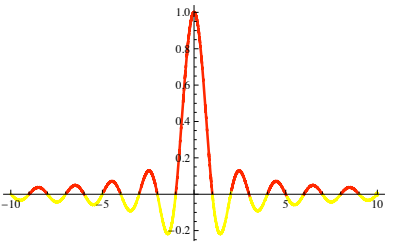

Figura 1.7. Función $h(z) = \frac{\sin(\pi z)}{\pi z}$.

ecuación integral de tipo Fredholm determina el degradado por borrado de una imagen. Una PSF es una función matemática que describe la distorsión en términos de la trayectoria que toma una fuente puntual de luz (u otras ondas) a través del instrumento con el cual se obtiene la imagen. A continuación presentaremos algunos de los núcleos más comunes que aparecen durante el proceso de restauración de una imagen digital.

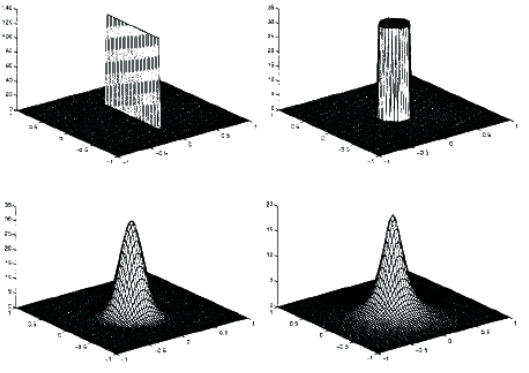

Figura 1.8. Distintas funciones de dispersión del punto.

Para grabar una imagen con una cámara digital lo que se usa es el número de fotones que pegan en una arreglo de sensores bidimensional, por lo que es posible considerar a una imagen digital como un muestreo de una función continua $f(x', y')$, $f : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$, la cual es una representación exacta de la imagen. Ya hemos hablado anteriormente cerca de los valores que puede tomar $f(x', y')$, los cuales pueden ser números entre 0 y 1, enteros entre 0 y 255. Nosotros tomaremos el caso de que estos valores pueden ser calesquiera números reales.

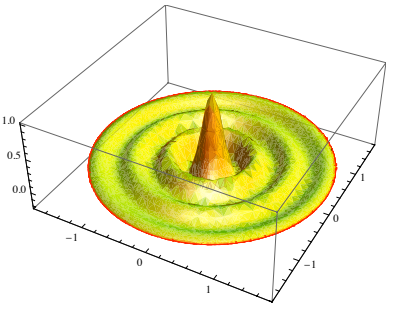

Figura 1.9. Función de dispersión de punto $sinc(4\pi\sqrt{x^2+y^2})$

Una cámara es un conjnunto de sensores, al igual que nuestros ojos, la diferencia entre ambos es el procesamiento.

El emborronamiento de una imagen e debe a distintos fenómenos. En distintos sistemas de imagenología la PSF es invariante con respecto a traslaciones; en el siguiente sentido, la imagen $K(x, y, x', y')$ de una fuente puntual localizada en el punto (x', y') es trasladada por (x', y') de la imagen $K(x, y, 0, 0)$ de una fuente puntual localizada en el origen de el plano objetivo. Matemáticamente, $K(x, y, x', y') = K(x, y, 0, 0)$. De esta forma, se tiene que $K(x, y, x', y')$ depende solamente de la diferencia o traslación $(x, y)-(x', y')$ y en en consecuencia escribiremos $K(x - x', y - y')$ en lugar de $K(x, y, x', y')$. Núcleos con esta propiedad son llamados *espacialmente invariantes*. De esta forma, para conocer un núcleo espacialemnte invariante es suficiente con detectar la imagen de solamente una fuente puntual, digamos una fuente puntual que se encuentra localizada en el centro de la imagen.

En vista de lo anterior, la representación obtenida es de la forma

$$g(x, y) = \int K(x - x', y - y')f(x', y')dx'dy'. \tag{1.8}$$

Usando la notación clásica asociada a la convolución

$$g = K * f. \tag{1.9}$$

Pero nos referiremos específicamente en estas notas a degradación debida a núcleos espacialmente invariantes que son separables y de tipo convolución. Es decir, núcleos de la forma

$$K(x, y, x', y') = h(x - x')k(y - y'). \tag{1.10}$$

El que el núcleo tenga esta forma significa que la degradación es la mismo en todas partes en la imagen y esta además se separa en sus componentes horizontal y vertical.

Figura 1.10. Imagen original.

Un tipo de PSF es la conocida como gaussina, una aproximación discreta de esta función de dispersión del punto está dada por la matriz

$$\frac{1}{16}\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix} = \mathbf{X}^T \cdot \mathbf{X}, \tag{1.11}$$

donde

$$\mathbf{X} = \frac{1}{4}\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

la cual se muestra en la Figura 1.4.

Uno de los problemas más complicados en la recuperación de una imagen es el de determinar la PSF que origina la degradación de esta. A continuación daremos las principales características de las funciones de dispersión del punto asociadas a degradaciones debidas a movimientos lineales y a imágenes fuera de foco. Es posible considerar estas degradaciones por separado o considerarlas simultáneamente.

1.5. Degradación por movimiento

Primero recordaremos el proceso de formción de una imagen en una cámara. Un sistema fundamental de una cámara cosnsist de un lente convexo el cual tiene una longitud focal la cual denotaremos por

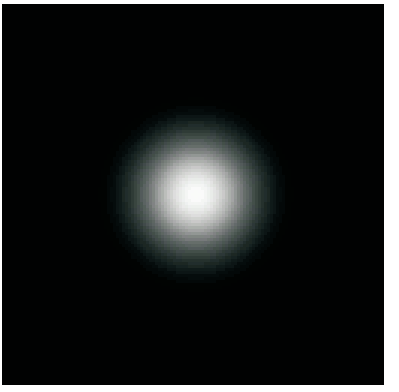

Figura 1.11. PSF Gaussiana.

f. La relación entre la distancia que tiene un punto de una imagen que desamos tomar de manera que se encuentre enfocada

$$\frac{1}{p}+\frac{1}{q}=\frac{1}{f} \tag{1.12}$$

donde p es la distancia entre el lente convexo y la posición que estamos enfocando con la cámara, q es la distancia entre la posición enfocada en el plano de la imagen y el lente convexo. Si el objeto se mueve a una diatancia d paralela al plano de la imagen, entonces tenemos

$$\frac{d}{p}=\frac{d'}{q} \tag{1.13}$$

donde d' es el desplazamiento de la posición correspondiente en el plano de la imagen.

Si suponemos en la ecuación (1.13) que el tiempo de exposición de la cámara es pequeño, el objeto en el plano de la imaen tiene la misma intensidad que después del deplazamniento d'. Por otra parte, la PSF asociada a un movimiento lineal puede modelarse como una recta que se mueve en un plano,

$$w(x',y')=\begin{cases}\frac{1}{|\overline{\mathbf{v}}|T}, & \text{si } -\frac{T\mathbf{v}_x}{2}\le x'\le\frac{T\mathbf{v}_x}{2},\ y'=\frac{\mathbf{v}_y}{\mathbf{v}_x};\\ 0, & \text{en cualquier otr caso.}\end{cases} \tag{1.14}$$

donde $\overline{\mathbf{v}}$ es la velocidad del objeto en el plano imagen, $\mathbf{v}_x$ y $\mathbf{v}_y$ son las componentes de este vector velocidad.

De esta forma el modelo para el efecto de borrado debido a un movimiento lineal está dado por

$$g(x,y)=\int\!\!\int f(x',y')w(x-x',y-y')dx'dy', \tag{1.15}$$

donde $g(x, y)$ es la imagen observada con degradación debida a un movimiento lineal y $f(x', y')$ es la imagen *real*.

Figura 1.12. Degradación debida movimiento horizontal.

En el caso de que el movimiento sea horizontal y el tiempo de exposición sea pequeño, el kernel toma la forma

$$K(x, y, x'y') = h_L(x - x') = \begin{cases} \frac{1}{2L}, & \text{si } |x - x'| \leq L; \\ 0, & \text{en cualquier otro caso.} \end{cases} \tag{1.16}$$

En el caso de movimeinto vertical tenemos

$$K(x, y, x'y') = h_L(y - y') = \begin{cases} \frac{1}{2L}, & \text{si } |y - y'| \leq L; \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Figura 1.13. Degradación debida a un movimiento vertical de la cámara.

1.6. Imágenes fuera de foco

El efecto de borrado debido a que una imagen se encuentra desenfocada en consecuencia de la imagen se encuentra fuera del plano de la imagen. De esta forma lo que queda en el plano de la imaen es un círculo con una mayor intensidad en el centro. Este efecto se modelo como una función con distribución gaussiana. En suma, la degradación de una imagen fuera de foco está modelada por

$$g_{ff}(x,y) = \int\int f(x',y') \left[\frac{1}{2\pi\sigma^2} \exp\left[-\frac{(x-x')^2 + (y'-y)^2}{2\sigma^2} \right] \right] dx'dy' \tag{1.17}$$

donde $g_{ff}(x,y)$ es la imagen observada que se encuentra fuera de foco, σ es la varianza para la función de distribución gaussiana y $f(x',y')$ es la imagen *real*. Observemos que el núcleo está dado por la función

$$K(x,x',y,y') = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(x-x')^2 + (y'-y)^2}{2\sigma^2} \right],$$

ver Figura 1.6. Por otra parte, el núcleo es separable pues puede reescriirse como un producto de la forma (1.10)

$$K(x,x',y,y') = \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-x')^2}{2\sigma^2} \right] \right] \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-y')^2}{2\sigma^2} \right] \right],$$

donde cada uno de los factores es una traslación de la función

$$h(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{z^2}{2\sigma^2} \right],$$

ver Figura 1.6. Además el efecto es el mismo, independientemente de

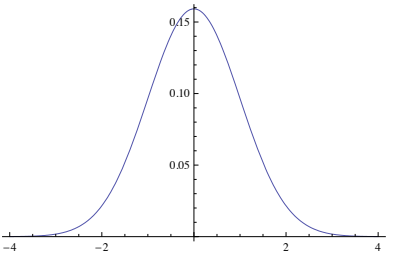

Figura 1.14. Gráfica de $h(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{z^2}{2\sigma^2} \right]$.

el punto que estemos considerando, como lo muestra la Figura 1.6.

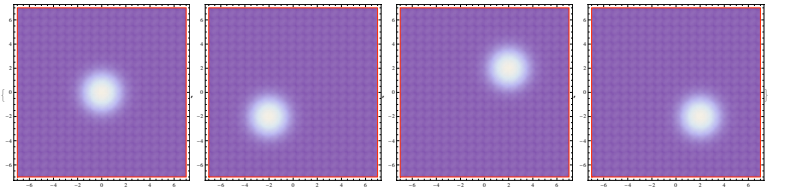

Figura 1.15. Núcleo gaussiano invariante bajo traslaciones.

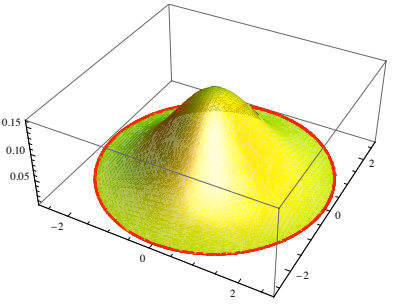

Figura 1.16. Núcleo gaussiano.

Figura 1.17. Degradación debida aque el objeto se encuentra fuera de foco.

1.7. Aspectos computacionales con Matlab

```
%%=========================================%%
%%%%    DECONVOLUCION UNIDIMENSIONAL    %%%%%%%
%%=========================================%%

% A continuacion mostraremos las dificultades que se
```

```
% presentan al intentar resolver el problema de
% deconvolucion x=Ab usando la solucion directa x=A\b

%%% Primero construiremos una seal unidimensional, la cual
%%% posteriormente intentaremos reconstruir

% La seal que reconstruiremos conta de 64 muestras

N = 64;
x = zeros(1,N);
x(round(N/4):round(3*N/4)) = 1;
x = x(:);
% La seal se considera como vector vertical, no horizontal

%----------------
% Veamos la seal
%----------------
figure(1)
clf
plot(1:N,x,'k','linewidth',2)
title('Seal x, por reconstruir','Fontsize',16)
axis([1 64 -.2 1.2])
axis square
drawnow
print -depsc -tiff Conv1_signalOriginal
pause
```

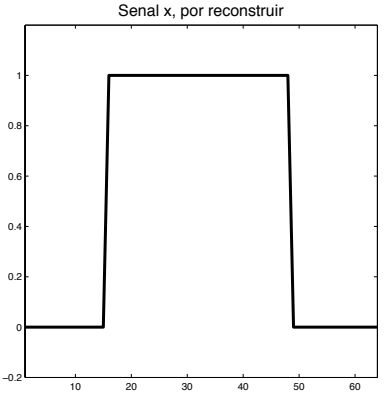

Figura 1.18. Señal por reconstruir.

```
%-----------------------------------
% Construccion de la PSF
% Construccion de la matriz sparse A
%
%-----------------------------------
M   = 3;
len = 2*M+1;
psf = ones(1,len)/len;
%psf = conv2(psf,psf,'same');
%psf = [.5 2 .5];
psf = psf/sum(sum(psf));
A   = convmtx(psf,N);
A   = A(:,(1+M):(end-M));

%-----------------------------------------
% Veamos la geometria de la matriz sparse A
%-----------------------------------------
figure(2)
clf
spy(A)
title('Entradas no nulas de la matriz
sparse A','Fontsize',16)
print -depsc -tiff Conv2_sparseA
pause
```

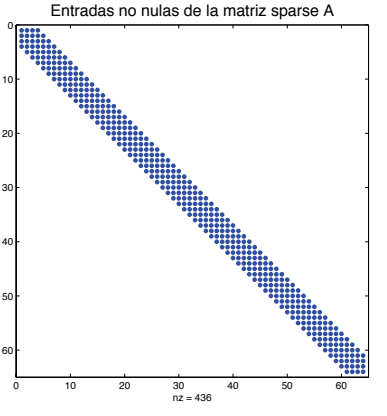

Figura 1.19. Geometría de la matriz sparse A..

```
%-----------------------------------------------
%%  Construccion de las seales de salida:
%%  Ideal:      b=Ax
%%  con ruido   br=b+0.02*randn(size(b))
%-----------------------------------------------

b=Ax;
br = b+0.02*randn(size(b));

%-----------------------------------------------
% Veamos la seal ideal de salida b=Ax
%-----------------------------------------------

figure(4)
plot(1:N,b,'r','linewidth',2)
title('seal de salida ideal b=Ax','Fontsize',16)
axis([1 64 -.2 1.2])
axis square
drawnow
print -depsc -tiff Conv3_SignalIdeal
pause
```

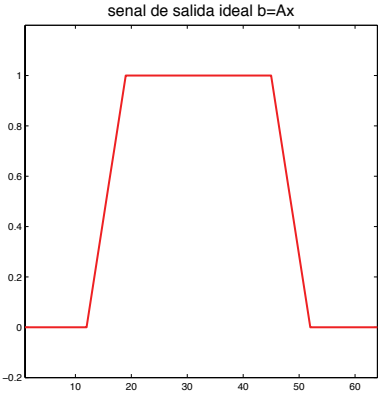

Figura 1.20. Señal ideal $b = Ax$.

```
%-------------------------------------------------------------
% Veamos simultneamente las seales ideal de salida b y br
%-------------------------------------------------------------
```

```
figure(7)
plot(1:N,b,'r','linewidth',4)
set(gca,'xticklabel',{})
set(gca,'yticklabel',{})
hold on
plot(1:N,br,'b','linewidth',2)
title('Seales salida b y b+ruido','Fontsize',16)
axis([1 64 -.2 1.2])
axis square
legend('datos ideales','datos con ruido')
drawnow
print -depsc -tiff Conv5_signal_b_br
pause
```

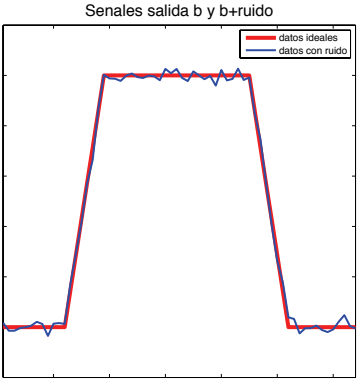

Figura 1.21. Señales b y br.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Problemas que aparecen al usar la reconstruccion directa %
% Reconstruccion:                                          %
%        de "x" a partir de datos ideales b                %
%        de "x" a partir de datos con ruido br           %
% usando el comando backslash de Matlab "\" e.g.  x=A\b    %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

rec = A\b;
recr= A\br;

%-------------------------------------
```

```
% Veamos "x" y su reconstruccion "A\b"
%-------------------------------------
figure(8)
clf
subplot(2,1,1)
plot(1:N,x,'k','linewidth',2)
set(gca,'xticklabel',{})
set(gca,'yticklabel',{})
box off
title('Seal x','Fontsize',16)
axis([1 64 -.2 1.2])
axis square
subplot(2,1,2)
plot(1:N,rec,'r','linewidth',2)
set(gca,'xticklabel',{})
set(gca,'yticklabel',{})
box off
title('Reconstruccion de x usando x=A\b','Fontsize',16)
axis([1 64 -.2 1.2])
axis square
drawnow
print -depsc -tiff Conv8_signal_x_rec_b
pause
```

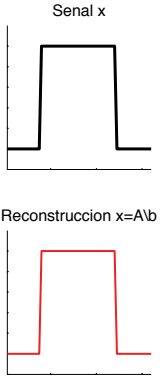

Figura 1.22. Señales x y br.

```
%-------------------------------------
% Veamos "x" y su reconstruccion "A\br"
```

```
%-------------------------------------


figure(10)
plot(1:N,x,'k','linewidth',4)
set(gca,'xticklabel',{})
set(gca,'yticklabel',{})
hold on
plot(1:N,recr,'b','linewidth',2)
set(gca,'xticklabel',{})
set(gca,'yticklabel',{})
box off
title('Seales "x" y "x=A\br, br=b+ruido','Fontsize',16)
axis([1 64 -.2 1.2])
axis square
print -depsc -tiff Conv6_x_rec_br
```

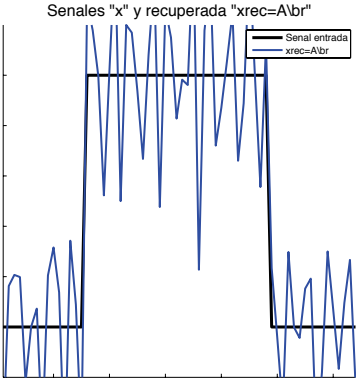

Figura 1.23. Señales x y br.

Capítulo 2

Problemas inversos en espacios de dimensión finita

El ágebra lineal tiene una fuerte influencia sobre el estudio de los problemas inversos. Normalmente se usan modelos continuos para modelar problemas inversos, más, aún, estos problemas inversos están definidos en espacios de Hilbert de dimensión infinta, pero para poder simularlos es necesario proceder a una discretización d esos modelos. La herramienta fundamental para trabajar con esas versiones dsicretas de un problema inversos es precisamente el álgebra lineal. Recordemos que el matemático francés Jacques Solomon Hadamard (1865-1963) estableció primeramente dos características que debía satisfacer un problema para que estuviese **bien condicionado**:

- Existencia: El rpoblema siempre debe tener solución
- Unicidad: La solución del problema debe ser única

posteriormente la anadió una tercera característica para que un problema estuviese bien condicionado

- Estabilidad: las soluciones dependen continuamente de las condiciones iniciales.

De un tiempo para acá han aparecido gran cantidad de problemas, los cuales no satisfacen alguna de estas tres condiciones. Es decir, o ono tienen solución, o la solución no es única o la solución no depende continuamente de las condiciones iniciales. Estos problemas son llamados **problemas mal condicionados**. Los **problemas inversos** caen en esta esfera de los problemas mal condicionados.

En el ámbito del ágebra lineal podemos determinar sin gran dificultad problemas que no resulen estar bien condicionados usando ejemplos de sistemas de ecuaciones lineales.

Ejercicio: Determinar un sistema de ecuaciones $Ax = b$ el cual tenga al menos dos soluciones.

Ejercicio: Determinar un sistema de ecuaciones $Ax = b$ el cual no tenga soluciones.

El hecho de encontar un sistema de ecuaciones lineales $Ax = b$ el cual no sea estable, tiene que ver con el hecho de que la matriz A de coficientes del sistema esté muy mal condicionada. Lo que coloquialmente podemos decir, como que la matriz estémuy cerca de no ser invertible (en el caso de matrices cuadradas).

EJEMPLO 2.0.1. Consideremos el sistema de ecuaciones lineales

$$\begin{aligned} x + 3y &= a \\ 2x + 6y &= 2a \end{aligned}$$

donde $a \in \mathbb{R}$. Observemos que este problema lo podemos reescribir matricialmente como

$$A\mathbf{x} = \mathbf{b},$$

donde $\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}$ y $\mathbf{b} = \begin{pmatrix} a \\ 2a \end{pmatrix}$.

2.1. Aspectos básicos de álgebra lineal

Recordemos unos aspectos básicos del álgebra lineal:

Algunas veces será necesario hacer referencia a renglones o columnas específicas de una matriz o a una submatriz de la misma, recordemos que la entrada (i, j) de la matriz corresponde al elemento de la matriz que se encuentra simultáneamente en el renglón i-ésima y la columna j-ésima, el cual denotamos por $A_{i,j}$. Si en lugar de esto queremos referirnos al renglón i-ásimo de la matriz $A \in \mathbb{R}^{m\times n}$, lo haremos con la notación $A(:, i)$, mientras que si deseamos hacer referencia a la columna j-ésima lo haremos usando la notación $A(:, j)$. ESta notación es la usada en *Matlab*.

EJEMPLO 2.1.1. Si $A = \begin{pmatrix} 1 & -2 & 4 \\ -9 & 5 & 12 \\ 0 & 3 & 23 \\ 15 & -6 & 0 \end{pmatrix}$, entonces $A(2,:) = (-9\ 5\ 12)$ y $A(:, 3) = \begin{pmatrix} 4 \\ 12 \\ 23 \\ 0 \end{pmatrix}$.

Observemos que usando notación vectorial es posible reescribir un sistema de ecuaciones lineales en forma vectorial, como lo muestra el siguiente ejemplo

EJEMPLO 2.1.2. El sistema de ecuaciones lineales

$$\begin{aligned} 2x + 4y - 3z &= 0 \\ 3x - 8y + z &= -7 \end{aligned} \tag{2.1}$$

puede escribirse como

$$x\begin{pmatrix}2\\3\end{pmatrix}+y\begin{pmatrix}4\\-8\end{pmatrix}+z\begin{pmatrix}-3\\1\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}$$

Recordemos que si $A\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^m)$ y la describimos mediante sus vectores columna $\mathbf{a_1},\ldots,\mathbf{a_n}$, (en notación de Matlab las columnas están dadas por $A(:,1),\ldots,A(:,n)$) los cuales son elementos de $\mathbb{R}^m$; es decir,

$$A=(\mathbf{a_1}\ \mathbf{a_2}\ \ldots\ \mathbf{a_n})=(A(:,1)\ A(:,2)\ \ldots\ A(:,n)),$$

donde

$$\mathbf{a}_j=A(:,j)=\begin{pmatrix}a_{1j}\\x_{2j}\\\vdots\\x_{mj}\end{pmatrix}$$

y consideramos un vector

$$x=\begin{pmatrix}x_1\\x_2\\\vdots\\x_x\end{pmatrix}\in\mathbb{R}^n,$$

entonces

$$Ax=(\mathbf{a_1}\ \mathbf{a_2}\ \ldots\ \mathbf{a_n})\begin{pmatrix}x_1\\x_2\\\vdots\\x_x\end{pmatrix}=x_1\mathbf{a_1}+x_2\mathbf{a_2}+\cdots+x_n\mathbf{a_n}=\sum_{j=1}^{n}x_jA(:,j).$$

Asimismo, si tomamos el producto matricial directo dde la parte izquierda de la igualdad

$$Ax=b$$

donde A y x están dada como anteriormente, entonces tenemos que

$$b_q=\sum_{k=1}^{n}a_{qk}x_k.$$

Al considerar matrices, se tiene que si $A\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^m)$ y $B\in\mathcal{L}(\mathbb{R}^m,\mathbb{R}^p)$, con $B=(\mathbf{b_1}\ \mathbf{b_2}\ \ldots\ \mathbf{b_p})$, o en notación de Matlab $B=(B(;,1)\ B(;,2)\ \ldots\ B(;,p))$, con $\mathbf{b_j}=B(;,j)\in\mathbb{R}^m$, entonces

$$AB=A(\mathbf{b_1}\ \mathbf{b_2}\ \ldots\ \mathbf{b_p})=(A\mathbf{b_1}\ A\mathbf{b_2}\ \ldots\ A\mathbf{b_p})\in\mathbb{R}^{m\times p}.$$

equivalentemente

$$AB=A(B(;,1)\ B(:,2)\ \ldots\ B(:,p))=(AB(:,1)\ AB(:,2)\ \ldots\ AB(:,p)).$$

EJEMPLO 2.1.3. Sean

$$A = \begin{pmatrix} 1 & 2 \\ 3 & -2 \\ -4 & 0 \end{pmatrix} \text{ y } B = \begin{pmatrix} -1 & 2 \\ 3 & 5 \end{pmatrix}$$

entonces

$$\begin{aligned} AB &= (AB(:,1), AB(:,2)) \\ &= \left(\begin{pmatrix} 1 & 2 \\ 3 & -2 \\ -4 & 0 \end{pmatrix} \begin{pmatrix} -1 \\ 3 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 \\ 3 & -2 \\ -4 & 0 \end{pmatrix} \begin{pmatrix} 2 \\ 5 \end{pmatrix} \right) \\ &= \left((-1) \begin{pmatrix} 1 \\ 3 \\ -4 \end{pmatrix} + 3 \begin{pmatrix} 2 \\ -2 \\ 0 \end{pmatrix} \quad 2 \begin{pmatrix} 1 \\ 3 \\ -4 \end{pmatrix} + 5 \begin{pmatrix} 2 \\ -2 \\ 0 \end{pmatrix} \right) \\ &= \begin{pmatrix} 5 & 12 \\ -9 & -4 \\ 4 & -8 \end{pmatrix} \end{aligned}$$

DEFINICIÓN 2.1.4. Dada una matriz $A \in \mathbb{R}^{m\times n}$ y $A = (a_{ij})_{i=1,\dots,m;j=1,\dots,n}$ definimos la matriz transpuesta de A como la matriz $A^t = (\tilde{a}_{ij})$ donde $\tilde{a}_{ij} = a_{ji}$.

Una propiedad de las matrices transpuestas es que $(AB)^T = B^T A^T$, como lo muestran los siguientes cálculos:

$$(AB)^t = \left[A(\mathbf{b_1}\ \mathbf{b_2}\ \dots\ \mathbf{b_p})\right]^t = \begin{pmatrix} \mathbf{b_1}^t A^t \\ \mathbf{b_2}^t A^t \\ \vdots \\ \mathbf{b_p}^t A^t \end{pmatrix} = \begin{pmatrix} \mathbf{b_1}^t \\ \mathbf{b_2}^t \\ \vdots \\ \mathbf{b_p}^t \end{pmatrix} A^t = B^t A^t.$$

DEFINICIÓN 2.1.5. Una matriz $A \in \mathbb{R}^{n\times n}$ es simétricas si $A^T = A$.

DEFINICIÓN 2.1.6. Una matriz $A \in \mathbb{R}^{m\times n}$ se dice que es una matriz diagonal si $A(i,j) = 0$ para $i \neq j$.

EJEMPLO 2.1.7. La matriz

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & -65 & 0 \\ 0 & 0 & 7 \\ 0 & 0 & 0 \end{pmatrix}$$

es una matriz diagonal.

DEFINICIÓN 2.1.8. Para dos vectores $\mathbf{x} \in \mathbb{R}^m$ y $\mathbf{y} \in \mathbb{R}^n$ definimos su **producto exterior** como

$$\begin{aligned}\mathbf{x}\mathbf{y}^t :&= \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \\ &= (y_1 \ y_2 \ \dots \ y_n) \\ &= \begin{pmatrix} x_1y_1 & x_1y_2 & \dots & x_1y_n \\ x_2y_1 & x_2y_2 & \dots & x_2y_n \\ \vdots & \ddots & \ddots & \vdots \\ x_my_1 & x_my_2 & \dots & x_my_n \end{pmatrix} \in \mathbb{R}^{m\times n}\end{aligned}$$

PROPOSICIÓN 2.1.9. *Si $\alpha \in \mathbb{R}^n$, entonces*

$$(\mathbf{x}\mathbf{y}^t)\alpha = \mathbf{x}(\mathbf{y}^t\alpha) \in \langle\{\mathbf{x}\}\rangle$$

Esto significa que la matriz tiene rango igual a 1. Más generalmente, para $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k \in \mathbb{R}^m$ y $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k \in \mathbb{R}^n$ se tiene que la matriz A dada por

$$A = \sum_{j=1}^{k} \mathbf{u}_j\mathbf{v}_j^t$$

es una matriz con rango a lo más igual a k, ya que

$$A\mathbf{x} = \sum_{j=1}^{k} \mathbf{u}_j(\mathbf{v}_j^t\mathbf{x}) \in \langle\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\}\rangle$$

TEOREMA 2.1.10. *Sean $U = [\mathbf{u_1}, \mathbf{u_2}, \dots, \mathbf{u_n}] \in \mathbb{R}^{m\times n}$ con $\mathbf{u_k} \in \mathbb{R}^m$ y $V = [\mathbf{v_1}, \mathbf{v_2}, \dots, \mathbf{v_n}] \in \mathbb{R}^{p\times n}$ con $\mathbf{v_i} \in \mathbb{R}^p$. Entonces*

$$UV^t = \sum_{i=1}^{n} \mathbf{u_i}\mathbf{v_i}^t \in \mathbb{R}^{m\times p}.$$

DEFINICIÓN 2.1.11. Si $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r\} \subset \mathbb{R}^n$, definimos el **espacio generado** por esta familia de vectores como el subespacio de $\mathbb{R}^n$ dado por

$$\langle\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d\}\rangle := \{\mathbf{b} \mid \mathbf{b} = \sum_{j=1}^{d} \alpha_j\mathbf{v}_j\}.$$

DEFINICIÓN 2.1.12. Decimos que una matriz $U = (\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_n) \in \mathbb{R}^{n\times n}$ es una mtriz ortogonal si

(1) Los vectores columna son mutuamente ortogonales. Es decir

$$\mathbf{u}_j \perp \mathbf{u}_k$$

para $j \neq k$.

(2) Los vectores columna son de norma unitaria. Es decir,

$$\|\mathbf{u}_j\| := \sqrt{\mathbf{u}_j^t \mathbf{u}_j} = 1$$

para $j = 1, 2 \ldots, n$.

PROPOSICIÓN 2.1.13. *Dada una matriz $U \in \mathbb{R}^{n\times n}$ ortogonal, $U^tU = Id$.*

DEMOSTRACIÓN 2.1.14. Mediante cálculos directos obtenemos

$$\begin{aligned} U^tU &= \begin{pmatrix} \mathbf{u}_1^t \\ \mathbf{u}_2^t \\ \vdots \\ \mathbf{u}_n^t \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_n \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{u}_1^t\mathbf{u}_1 & \mathbf{u}_1^t\mathbf{u}_2 & \ldots & \mathbf{u}_1^t\mathbf{u}_n \\ \mathbf{u}_2^t\mathbf{u}_1 & \mathbf{u}_2^t\mathbf{u}_2 & \ldots & \mathbf{u}_2^t\mathbf{u}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}_n^t\mathbf{u}_1 & \mathbf{u}_n^t\mathbf{u}_2 & \ldots & \mathbf{u}_n^t\mathbf{u}_n \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 \end{pmatrix} \\ &= I \end{aligned}$$

PROPOSICIÓN 2.1.15. *Toda matriz ortogonal $U \in \mathbb{R}^{n\times n}$ preserva la norma.*

DEMOSTRACIÓN 2.1.16.

$$\|U\mathbf{x}\|^2 = (U\mathbf{x})^t(U\mathbf{x}) = \mathbf{x}^tU^tU\mathbf{x} = \mathbf{x}^t\mathbf{x} = \|\mathbf{x}\|^2.$$

PROPOSICIÓN 2.1.17. *Si $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k \in \mathbb{R}^n$ son mutualmente ortogonales y $k \leq n$, entonces* $\dim < \{\mathbf{u}_1, \ldots, \mathbf{u}_k\} >= k$.

PROPOSICIÓN 2.1.18. *Si U es una matriz ortogonal, entonces para cada par de vectores $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ se tiene que $\mathbf{x}^T\mathbf{y} = U\mathbf{x}^T\mathbf{U}\mathbf{y}$.*

Como consecuencia inmediata del resultado anterior tenemos que la familia de vectores dada por $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ forma una base ortogonal del subespacio $< \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\} >$ de $\mathbb{R}^m$.

DEFINICIÓN 2.1.19. Dados dos subespacios M, N de $\mathbb{R}^n$, decimos que estos subespacios son mutualmente ortogonales si para cualquier pareja de elementos $\mathbf{x} \in M, \mathbf{y} \in N$ se tiene que $\mathbf{x}^t\mathbf{y} = 0$. En este caso escribimos $M \perp N$.

DEFINICIÓN 2.1.20. Dada $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, definimos el **rango** de A como el subconjunto de $\mathbb{R}^m$ dado por

$$\mathcal{R}(A) := \{A\mathbf{x} \,|\, \mathbf{x} \in \mathbb{R}^n\}.$$

Debemos notar que

$$\mathcal{R}(A) =< \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\} > .$$

DEFINICIÓN 2.1.21. Dado M un subespacio de $\mathbb{R}^n$, el **complemento ortogonal de**M está dado por el conjunto

$$M^{\perp} := \{\mathbf{y} \in \mathbb{R}^n \mid \mathbf{y} \perp \mathbf{x}, \forall \mathbf{x} \in M\}.x$$

Si $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ son las transformaciones lineales de $\mathbb{R}^n$ en $\mathbb{R}^m$, y si A denota indistintamente a una transformación lineal entre estos espacios euclideanos o a la representación matricial de la transformación lineal con respecto a las bases canónicas (naturales) de $\mathbb{R}^n$ y $\mathbb{R}^m$, respectivamente, entonces tenemos los siguientes resultados.

TEOREMA 2.1.22. *Sea $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. Entonces*

(1) $\mathcal{N}(A)^{\perp} = \mathcal{R}(A^t)$. *Esta afirmación sólo se satisface para espacios vectoriales de dimensión finita.*
(2) $\mathcal{R}(A)^{\perp} = \mathcal{N}(A^t)$. *Esta afirmación sólo se satisface para espacios vectoriales de dimensión finita.*

DEMOSTRACIÓN 2.1.23. (1) Sea $x \in \mathcal{N}(A)$, entonces $Ax = 0$, de esta forma, para todo y se tiene que $y^tAx = 0$, pero $y^tAx = (A^ty)^tx$. POr lo que $Ax = 0$ si y sólo si $(x, A^t, y) = 0$. Es decir, si x es ortogonal a todo elemento de la forma A^ty, es decir x es elemento del complemento ortogonal del conjunto $\mathcal{R}(A^t)$. Pero como todas las afirmaciones son equivalencias, entonces se satisface la igualdad.
(2) Se deja como ejercicio al lector.

DEFINICIÓN 2.1.24. Para $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ definimos el **núcleo derecho** como

$$\{v \in \mathbb{R}^n \,|\, Av = 0\}.$$

Mientras que el **núcleo izquierdo** es el conjunto

$$\{w \in \mathbb{R}^m \,|\, w^tA = 0\}.$$

Observemos que el núcleo derecho no es más que el conjunto $\mathcal{N}(A)$, mientras que el núcleo izquierdo está dado por $\mathcal{N}(A^t)$.

TEOREMA 2.1.25. *Sea $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. Entoncesç*

(1)
$$\mathbb{R}^n = \mathcal{N}(A) \oplus \mathcal{R}(A^t).$$
Es decir, para todo $v \in \mathbb{R}^n$, existen únicos elementos $x \in \mathcal{N}(A)$ y $y \in \mathcal{N}(A)^\perp = \mathcal{R}(A^t)$ tales que $v = x + y$.

(2)
$$\mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{N}(A^t).$$
Es decir, para todo $w \in \mathbb{R}^m$ existen únicos elementos $x \in \mathcal{R}(A)$ y $y \in \mathcal{N}(A^t)$ tales que $w = x + y$.

DEFINICIÓN 2.1.26.

2.2. Subspacios fundamentales

Consideremos $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$.

DEFINICIÓN 2.2.1. Definimos $ran(A) := \dim \mathcal{R}(A)$. Este el llamado usualmente el **rango por columnas** de A (máximo número de columnas linealmente independientes). El **rango por renglones** de A está dado por el número $\dim \mathcal{R}(A^t)$ (máximo número de renglones linealmente independientes). El $coran(A)$ está dado por el número $\dim \mathcal{N}(A)$.

TEOREMA 2.2.2. *Para $A \in \mathbb{R}^{m\times n}$, $\dim \mathcal{R}(A) = \dim \mathcal{N}(A^\perp)$.*

DEMOSTRACIÓN 2.2.3.

COROLARIO 2.2.4. *Para $A \in \mathbb{R}^{m\times n}$, $\dim \mathcal{N}(A) + \dim \mathcal{R}(A) = n$.*

DEMOSTRACIÓN 2.2.5. Sea $A \in \mathbb{R}^{m\times n}$, es trivialmente claro que $\dim \mathcal{N}(A) + \dim(\mathcal{N}(A))^\perp = n$. Pero, sabemos, por el resultado anterior, que $\dim(\mathcal{N}(A))^\perp = \dim \mathcal{R}(A)$. Por lo tanto, tenemos la igualdad deseada.

TEOREMA 2.2.6. *Si $A, B \in \mathbb{R}^{n\times n}$, entonces*

(1) $0 \leq ran(A+B) \leq ran(A) + ran(B)$.
(2) $ran(A) + ran(B) - n \leq ran(AB) \leq \min\{ran(A), ran(B)\}$
(3) $null(B) \leq null(AB) \leq null(A) + null(B)$.
(4) *Si B es una matriz invertible, entonces $ran(AB) = ran(BA) = ran(A)$ y $\mathcal{N}(BA) = \mathcal{N}(A)$.*

DEMOSTRACIÓN 2.2.7.

TEOREMA 2.2.8. *Sean $A \in \mathbb{R}^{m\times n}$ y $B \in \mathbb{R}^{n\times p}$. Entonces*

(1) $\mathcal{R}(AB) \subseteqq \mathcal{R}(A)$.
(2) $\mathcal{N}(AB) \subseteqq \mathcal{N}(B)$.
(3) $\mathcal{R}((AB)^t) \subseteqq \mathcal{R}(B^t)$.
(4) $\mathcal{N}((AB)^t) \subseteqq \mathcal{N}(A^t)$.

DEMOSTRACIÓN 2.2.9.

TEOREMA 2.2.10. *Para $A \in \mathbb{R}^{m\times n}$ se tiene que*

(1) $\mathcal{R}(A^t) = \mathcal{R}(AA^t)$.
(2) $\mathcal{R}(A^t) = \mathcal{R}(A^tA)$.
(3) $\mathcal{N}(A) = \mathcal{N}(A^tA)$.
(4) $\mathcal{N}(A^t) = \mathcal{N}(AA^t)$.

DEMOSTRACIÓN 2.2.11.

TEOREMA 2.2.12. *Para $A \in \mathbb{R}^{m\times n}$ se tiene que*

(1) *A es suprayectiva si y sólo si $ran(A) = m$ (A tiene renglones linealmente independientes, se dice que tiene rango maximal por renglones; o equivalentemente AA^t es no singular).*
(2) *A es inyectiva si y sólo si $ran(A) = n$ (A tiene columnas linealmente independientes, se dice que tiene rango maximal por columnas; o equivalentemente A^tA es no singular).*

DEMOSTRACIÓN 2.2.13.
(1) Si A es suprayectiva, entonces $\mathcal{R}(A) = \mathbb{R}^m$, es decir, $\dim \mathcal{R}(A) = m = ran(A)$. Inversamente, si $\mathbf{y} \in \mathbb{R}^m$ y definimos $\mathbf{x} = A^t(AA^t)^{-1}\mathbf{y}$, entonces $A\mathbf{x} = y$, y en consecuencia A es suprayectiva.
(2) Si A es inyactiva, entonces $\mathcal{N}(A) = \{\mathbf{0}\}$; de esta forma $\dim(\mathcal{N}(A))^\perp = n$. Por un teorema anterior $\dim(\mathcal{N}(A))^\perp = \dim \mathcal{R}(A)$, en conclusión obtenemos la identidad buscada. Inversamente, sea A tal que $ran(A) = n$, es decir, tal que A^tA es invertible. Tomemos $\mathbf{u}$ y $\mathbf{v}$ taes que $A\mathbf{u} = A\mathbf{v}$, entonces $A^tA\mathbf{u} = A^tA\mathbf{v}$, pero, por hipótesis AA^t es invertible, entonces $\mathbf{u} = \mathbf{v}$. Por lo tanto, A es inyectiva.

DEFINICIÓN 2.2.14. Una transformación lineal $T : X \to Y$ es biyactiva si y sólo si es inyectiva y suprayectiva.

Observemos que T es no singular si y sólo si $ran(A) = n$.

DEFINICIÓN 2.2.15. Se dice que una transformación lineal $T : X \to Y$ es **invertible por la derecha** si existe una transformación $T^{-1}_{Der} : Y \to X$ tal que $TT^{-1}_{Der} = Id_Y$.

Se dice que una transformación lineal $T : X \to Y$ es **invertible por la izquierda** si existe una transformación $T^{-1}_{Izq} : Y \to X$ tal que $T^{-1}_{Izq}T = Id_X$.

TEOREMA 2.2.16. *Sea $T : X \to Y$ una transformación lineal. Entonces*

(1) *T es invertible por la derecha si y sólo si es suprayactiva.*
(2) *T es invertible por la izquierda si y sólo si es inyectiva.*

DEMOSTRACIÓN 2.2.17.

Como consecuencia inmediata de este resultado se tiene que una transformación lineal es invertible si y sólo si tiene inversas derecha e izquierda; es decir, si es suprayectiva e inyectiva, por lo que $T_{Izq}^{-1} = T_{Der}^{-1} = T^{-1}$.

TEOREMA 2.2.18. *Sea $T : X \to X$ una transformación lineal.*

(1) *Si T tiene una inversa derecha T_{Der}^{-1} tal que $TT_{Der}^{-1} = Id_Y$, entonces T es invertible.*
(2) *Si T tiene una inversa izquierda T_{Izq}^{-1} tal que $T_{Izq}^{-1}T = Id_X$, entonces T es invertible.*

DEMOSTRACIÓN 2.2.19. (1) Antes que nada observemos que la siguiente sucesión de igualdades se satisfacen

$$\begin{aligned} T(T_{Der}^{-1} + T_{Der}^{-1}T - I) &= TT_{Der}^{-1} + TT_{Der}^{-1}T - T \\ &= I + IT - T \\ &= I \end{aligned}$$

De esta forma $(T_{Der}^{-1} + T_{Der}^{-1}T - I)$ es inversa derecha de T, pero la inversa derecha es única, por lo que $(T_{Der}^{-1} + T_{Der}^{-1}T - I) = T_{Der}^{-1}$. Por lo tanto $T_{Der}^{-1}T = I$ y en consecuencia T_{Der}^{-1} es inversa izquierda.

(2) Se deja como ejercicio al lector.

2.3. Seudoinversa de Moore--Penrose.

Consideremos vdos espacios vectoriales X y Y de dimensión finita y una transformación lineal T entre estos espacios vectoriales. De esta forma, tenemos definidos los subespacios $\mathcal{N}(T)$ y $\mathcal{R}(T)$ de X y Y, respectivamente. Es posible definir la transformación

$$\tilde{T} : \mathcal{N}(T)^{\perp} \to \mathcal{R}(T)$$

por

$$\tilde{T}(x) = T(x), \quad x \in \mathcal{N}(T)^{\perp}.$$

PROPOSICIÓN 2.3.1. *$\tilde{T}$ es biyectiva.*

DEMOSTRACIÓN 2.3.2. Se deja como ejercicio.

Del resultado anterior tenemos definida la inversa

$$\tilde{T}^{-1} : \mathcal{R}(T) \to \mathcal{N}(T)^{\perp}.$$

A continuación haremos uso de esta transformación para describir la inversa de Moore--Penrose de T.

DEFINICIÓN 2.3.3. La **pseudo--inversa de Moore--Penrose** está dada por la transformación,

$$T^{\dagger} : Y \to X$$

defnida por

$$T^{\dagger}y = \tilde{T}^{-1}(y_1)$$

donde y está descompuesta de manera única como $y = y_1 + y_2$, con $y_1 \in \mathcal{R}(T)$ y $y_2 \in \mathcal{R}(T)^{\perp}$.

TEOREMA 2.3.4. *Sea $A \in M_r^{m\times n}$. Entonces $G = A^{\dagger}$ si y sólo si*

P1: $AGA = A$.
P2: $GAG = G$.
P2: $(AG)^t = AG$.
P3: $(GA)^t = GA$.

Más aún , la pseudo--inversa siempre existe y es única.

En el caso de matrices cuadradas no--singulares, es decri, con determinante no nulo, l inversa de la matriz satisface todas las propiedades establecidas por Penrose en 1955. Asimismo, si consideramos la inversa derecha o izquierda de la matriz rectangular, estas deben satisfacer no menos de tres de estas propiedades.

PROPOSICIÓN 2.3.5.
(1) *Si A es suprayectiva (renglones independientes)(A es invertible por la derecha)entonces $A^{\dagger} = A^t(AA^t)^{-1}$.*
(2) *Si A es inyectiva (columnas independientes)(A es invertible por la izquierda)entonces $A^{\dagger} = (A^tA)^{-1}A^t$.*
(3) *Para cualquioer escalar α, $\alpha^{\dagger} = \begin{cases} \alpha^{-1}, \text{si } \alpha \neq 0; \\ 0, \text{si } \alpha = 0. \end{cases}$*

A continuación se presenta una serie de resul tados concernientes a la pseudo--inversa de Moore--Penrose.

TEOREMA 2.3.6. *Si $A \in \mathbb{R}^{m\times n}$ y $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ son matrices ortogonales, entonces*

$$(UAV)^{\dagger} = V^tA^{\dagger}U^t.$$

DEMOSTRACIÓN 2.3.7. Verificar cada una de las condiciones de Penrose.

TEOREMA 2.3.8. *Sea $S \in \mathbb{R}^{n\times n}$ una matriz simétrica tal que $U^tSU = D$, donde U es una matriz ortogonal y D una matriz diagonal. Entonces $S^{\dagger} = UD^dagU^t$, donde $D^{\dagger}$ es una matriz duagonal.*

TEOREMA 2.3.9. *Si $A \in \mathbb{R}^{m\times n}$, entonces*
(1) $A^{\dagger} = (A^tA)^{\dagger}A^t = A^t(AA^t)^{\dagger}$.
(2) $(A^t)^{\dagger} = (A^{\dagger})^t$.

DEMOSTRACIÓN 2.3.10.

Estos últimos dos resultados dan una idea bastante clara de que, al menos teóricamente, es posible calcular la pseudo--inversa de Moore--Pentose para cualquier matriz. Esto debido a que los productos AA^t y A^tA son matrices simétricas.

De la misma manera, que, en general, el producto de matrices no es conmutatuvo, tampoco, lo es la pseudo--inversa de Moore--Penrose. Para mostrar esto, basta considerar las matrices $A = [01]$ y $B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

Bajo ciertas condiciones necasrias y suficientes la anterior propiedad se satisface.

TEOREMA 2.3.11. $(AB)^\dagger = B^\dagger A^\dagger$ *su y sólo si se satisfacen las dos condiciones siguientes:*

(1) $\mathcal{R}(BB^tA^t) \subset \mathcal{R}(A^t)$, *y*
(2) $\mathcal{R}(A^tAB) \subset \mathcal{R}(B)$.

TEOREMA 2.3.12. $(AB)^\dagger = B_1^\dagger A_1^\dagger$, *donde* $B_1 = A^\dagger AB$ *y* $A_1 = AB_1B_1^\dagger$.

TEOREMA 2.3.13. *Si* $A \in \mathbb{R}_r^{n\times r}$ *y* $B \in \mathbb{R}_r^{r\times m}$, *entonces* $(AB)^\dagger = B^\dagger A^\dagger$.

DEFINICIÓN 2.3.14. Decimos que una matriz A es **normal** si $AA^t = A^tA$.

Toda matriz simétrica, anti--simétrica u ortogonal es una matriz normal.

TEOREMA 2.3.15. *Para* $A \in \mathbb{R}^{m\times n}$,

(1) $A^\dagger A = A$.
(2) $(AA^t)^\dagger = A^\dagger(A^t)^\dagger$, $(AA^t)^\dagger = (A^t)^\dagger A^\dagger$.
(3) $\mathcal{R}(A^\dagger) = \mathcal{R}(A^t) = \mathcal{R}(A^\dagger A) = \mathcal{R}(A^tA)$.
(4) $\mathcal{N}(A^\dagger) = \mathcal{N}(AA^\dagger) = \mathcal{N}((AA^t)^\dagger) = \mathcal{N}(AA^t) = \mathcal{N}(A^t)$.
(5) *Si A es normal, entonces* $A^kA^\dagger = A^\dagger A^k$ *y* $(A^k)^\dagger = (A^\dagger)^k$ *para todo número natural k.*

TEOREMA 2.3.16. *Si* $A \in \mathbb{R}^{n\times p}$, $B \in \mathbb{R}^{n\times m}$. *Entonces* $\mathcal{R}(B) \subsetneqq \mathcal{R}(A)$ *si y sólo si* $AA^\dagger B = B$.

DEMOSTRACIÓN 2.3.17. Supongamos que $\mathcal{R}(B) \subsetneqq \mathcal{R}(A)$, donde $A \in \mathbb{R}^{n\times p}$ y $B \in \mathbb{R}^{n\times m}$. Sea $\mathbf{x} \in \mathbb{R}^m$, entonces $B\mathbf{x} \in \mathcal{R}(B) \subsetneqq \mathcal{R}(A)$, por lo que existe $\mathbf{y} \in \mathbb{R}^p$ tal que $A\mathbf{y} = B\mathbf{x}$, por lo que

$$B\mathbf{x} = A\mathbf{y} = AA^\dagger A\mathbf{y} = AA^\dagger B\mathbf{x}.$$

Por otra parte, supongamos que $AA^\dagger B = B$ y tomemos $\mathbf{y} \in \mathcal{R}(B)$, entonces existe $\mathbf{x} \in \mathbb{R}^m$ tal que $B\mathbf{x} = \mathbf{y}$, de esta forma

$$= B\mathbf{x} = AA^\dagger B\mathbf{x} = A(A^\dagger B\mathbf{x}) \in \mathcal{R}(A).$$

2.4. Descomposición en valores singulares

En cursos de métodod numéricos es muy común encontrarse con distintos tipos de descomposiciones para ciertos tipos de matrices. Tales descomposiciones o factorizaciones permiten resolver sistemas de ecuaciones lineales: De entre este tipo de factorizaciones podemos mencionar lass llamadas factorizaciones LU, QR, entre otras. Una de las características es que estas descompocisiones no existen para todo tipo de matrices, sino, como lo mencionamos, para tipos especiales de ellas.

A continuación mencionaremos una factorización que existe para cualquier matriz.

Una **descompocisión en valores singulares** de una matriz $A \in \mathbb{R}^{m\times n}$ es una factorización de la forma

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

donde $U \in \mathbb{R}^{m\times m}$ y $V \in \mathbb{R}^{n\times n}$ son matrices ortogonales (es decir, $UU^t = I_{m\times m}$ y $VV^t = I_{n\times n}$) y Σ es una matriz que en sus entradas fuera de la diagonal son nulas, mientras que las entradas de la diagponal son elementos no negativos. Estos elementos sobre la diagonal son llamados los **valores singulares** de A, mientras que los vectores columna de las natrices U y V son llamados los **vectores singulares izquierdos** y **vectores singulares derechos** de la matriz A.

La gama de aplicaciones de la decomposición en valores singulares de una matriz es bastante rica. Una de las aplicaciones más impactanes de esta descomposición es al comprimir imágenes fotográficas digitales.

Antes de continuar, y para tener una mayor claridad al establecer el siguiente resultado, mencionesmos algunos aspectos de matrices diagonales.

En el caso de que $m > n$; es decir, que la cantidad de renglones sea mayor a la cantidad de columnas, entonces,

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots \\ \vdots & 0 & \ddots & \dots \\ 0 & & \dots & \sigma_n \\ \vdots & \dots & \dots & 0 \\ 0 & \dots & 0 & 0 \end{pmatrix} = diag(\sigma_1, \dots, \sigma_n).$$

En el caso que $m \leq n$, tenemos

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & 0 & \vdots & 0 \\ \vdots & & \sigma_m & \dots & 0 \end{pmatrix} = diag\,(\sigma_1, \dots, \sigma_m)\,.$$

De esta forma, es posible escribir,

$$\Sigma = diag\left(\sigma_1, \dots, \sigma_{\min\{m,n\}}\right).$$

donde los elementos sobre la diagonal satisfagan las desigualdaes

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{m,n\}} \geq 0.$$

TEOREMA 2.4.1. *Si $A \in \mathbb{R}^{m\times m}$ Entonces existen matrices ortogonales $U \in \mathbb{R}^{m\times m}$ y $V \in \mathbb{R}^{n\times n}$ tales que*

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

donde $\Sigma = \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}$, $S = diag\{\sigma_1, \sigma_2, \dots, \sigma_p\}$ y los elementos diagonales son tales que $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > 0$. Más aún, la matriz A puede escribirse como

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} \begin{bmatrix} V_1^t \\ V_2^t \end{bmatrix} = U_1 S V_1^t. \tag{2.2}$$

Las submatrices son tales que $U_1 \in \mathbb{R}^{m\times r}$, $U_2 \in \mathbb{R}^{m\times(m-r)}$, $V_1 \in \mathbb{R}^{n\times r}$ y $V_2 \in \mathbb{R}^{n\times(n-r)}$, donde $r \leq \min\{m,n\}$. Además los subbloques de la matriz Σ deben tener dimensiones compatibles con las demás matrices y submamtrices.

Antes de escribr la demostración de Teorema de Descomposición en Valores Singulares, detengámonos un poco en el mismo.

Si escribimos las matrices ortogonales U y V en términos de sus vectores columna,

$$U = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_m \end{pmatrix},$$

y

$$V = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{pmatrix}$$

entonces la descompición en valores singulares de la matriz A puede escribirse como

$$A = \sum_{j=1}^{k} \sigma_j \mathbf{u}_j \mathbf{v}_j^t.$$

DE esta forma, la matriz A se escribe como la suma de matrices de rango uno.

2.4.1. Descripción geométrica. Cuando consideramos una matriz $A \in \mathbb{R}^{m\times n}$ de rango r, y la vemos como transformación lineal, $\mathbb{R}^n$ (dominio de A) y $\mathbb{R}^m$ (contradominio) pueden descomponerse como

$$\mathbb{R}^n = \mathcal{R}(A^t) \oplus \mathcal{N}(A)$$

(la primera componente es el llamado *espacio--renglón*) y

$$\mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{N}(A^t)$$

las componentes de $\mathbb{R}^n$ tienen dimensiones $\dim \mathcal{R}(A^t) = r$ y $\dim \mathcal{N}(A) = n - r$; mientras que las componentes de $\mathbb{R}^m$ ienen dimensiones $\dim \mathcal{R}(A) = r$, y $\dim \mathcal{N}(A^t) = m - r$. Estos custro subespacios son los subespacios fundamentales asociados a una transformación lineal. Además, si consideramos un vector $\mathbf{x} \in \mathbb{R}^n$ (dominio de la transformación lineal), a este vector lo podemos decomponer en sus componentes en el espacio--renglón y su componente en el núcleo $\mathcal{N}(A)$ de la transformación.

Para el caso particular de matrices en $A \in \mathbb{R}^{2\times 2}$ se tiene que la imagen de una circunferencia es un elipsoide. Por otra pate si tenemos la descomposición en valores singulares de A y su sistema propio asociado está dado por el par de ternas

$$\{(\mathbf{v}_1, \mathbf{u}_1, \alpha_1), (\mathbf{v}_2, \mathbf{u}_2, \alpha_2)\}$$

las cuales son tales que por una parte

$$\mathbf{v}_1 \perp \mathbf{v}_2,\ \mathbf{u}_1 \perp \mathbf{u}_2$$

y por otra

$$A\mathbf{v}_1 = \alpha_1 \mathbf{u}_1\ A\mathbf{v}_2 = \alpha_2 \mathbf{u}_2.$$

La razón entre los valores propios singulares $\frac{\alpha_2}{\alpha_1}$ mide el grado de deformación, ya sea alargamiento o achatamiento en las direccioes de los vectores propios izquierdos.

2.4.2. Descomposición en valores singulares y el mal-condicionamiento. Dado el problema

$$\mathbf{y} = A\mathbf{x}_* + \epsilon \in \mathbb{R}^m,\ \mathbf{x}_* = \mathbf{x}\text{verdadero}$$

donde $\epsilon \in \mathbb{R}^m$ es un vector de error. Queremos dar una estimación del valor de $\mathbf{x}$ a partir de los datos. Habrá algún problema para resolver esta cuestión?.

Si buscamos resolver la ecuación $\mathbf{y} = A\mathbf{x}$. Si tomamos un elemento $\mathbf{x}_0 \in \mathcal{N}(A)$, entonces por la linealidad de A tendremos que

$$A\mathbf{x} = A(\mathbf{x} + \mathbf{x}_0).$$

De esta forma, si el núcleo de A es no trivial, entonces la solución al problema no necesariamente es única.

La descomposición en valores singulares $A = U\Sigma V^t$ de A nos provee de una herramienta que nos permite caracterizar a $\mathcal{N}(A)$.

TEOREMA 2.4.2.

$$\mathcal{N}(A) = \langle\{\mathbf{v}_{p+1}, \mathbf{v}_{p+2}, \ldots, \mathbf{v}_n\}\rangle.$$

DEMOSTRACIÓN 2.4.3. Como sabemos,

$$\Sigma = diag\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$$

donde $k = \min\{m, n\}$ y además, los valores singulares están escritos de manera no creciente

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_r > \alpha_{r+1} = \cdots = \alpha_k = 0.$$

Como V está dado por

$$V = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \ldots & \mathbf{v}_n \end{pmatrix} \in \mathbb{R}^{n\times n}$$

donde

$$A\mathbf{v}_j = \begin{cases} \alpha_j\mathbf{u}_j, & \text{si } j \leq r; \\ \mathbf{0}, & \text{si } p+1 \leq j \leq n. \end{cases} \tag{2.3}$$

De lo anterior se tiene claramente que

$$\mathcal{N}(A) = \langle\{\mathbf{v}_{p+1}, \mathbf{v}_{p+2}, \ldots, \mathbf{v}_n\}\rangle.$$

A continuación daremos una descripción del rango $\mathcal{R}(A)$ de la transformación A.

TEOREMA 2.4.4. *Si la ecuación*

$$\mathbf{y} = A\mathbf{x}$$

tiene solución, entonces

$$\mathcal{R}(A) = \langle\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}\rangle.$$

DEMOSTRACIÓN 2.4.5. Supongamos que $\mathbf{y} \in \mathbb{R}^m$ y que para algún vector $\mathbf{x} \in \mathbb{R}^n$ se tiene que

$$A\mathbf{x} = \mathbf{y}.$$

Como $\mathbf{x} \in \mathbb{R}^n$, entonces lo podemos escribir como combinación lineal de los vectores columna de la matriz V; es decir, existen $x_i \in \mathbb{R}$, $i = 1, \ldots, n$ tales que

$$\mathbf{x} = \sum_{i=1}^{n} x_i\mathbf{v}_i,$$

donde $x_i = \mathbf{v}_i\mathbf{x}$.

Un cálculo directo muestra que

$$\begin{aligned}\mathbf{y} &= A\mathbf{x} \\ &= A(\sum_{i=1}^{n} x_i \mathbf{v}_i) \\ &= \sum_{i=1}^{n} x_i A\mathbf{v}_i \\ &= \sum_{i=1}^{n} x_i \alpha_i \mathbf{u}_i \\ &= \sum_{i=1}^{r} x_i \alpha_i \mathbf{u}_i.\end{aligned}$$

Estas igualdades muestran, en consecuencia, que

$$\mathbf{y} \in \mathcal{R}(A) =< \{\mathbf{u}_1, \ldots, \mathbf{u}_r\} > .$$

COROLARIO 2.4.6. *Si* $\mathbf{y}$ *tiene alguna componente no nula en el complemento ortogonal de* $\mathcal{R}(A)$ *entonces la ecuación*

$$\mathbf{y} = A\mathbf{x}$$

no tiene solución.

DEMOSTRACIÓN 2.4.7. Si $\mathbf{y}$ tiene alguna componente no nula en el complemento ortogonal de $\mathcal{R}(A)$; es decir, si para algún $i \in r+1, r+2, \ldots, n$ se tiene que $\mathbf{u}_i^t \mathbf{y} \neq 0$, entonces la ecuación

$$\mathbf{y} = A\mathbf{x}$$

no tiene solución.

Ahora consideraremos la relación de la descomposición en valores singulares de A y el mal--condicionamiento del problema

$$\mathbf{y} = A\mathbf{x}.$$

Supongamos que $\mathbf{y} \in \mathcal{R}(A)$; Los cálculos realizados en la demostración del último teorema muestran que si $\mathbf{y} = \sum_{i=1}^{r} y_i \mathbf{u}_i$ y $\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{v}_i$, entonces la ecuación $\mathbf{y} = A\mathbf{x}$ implica que

$$\alpha_i x_i = y_i, \ 1 \leq i \leq r$$

y, en consecuencia,

$$\begin{aligned}\mathbf{x} &= \sum_{i=1}^{r} x_i \mathbf{v}_i \\ &= \sum_{i=1}^{r} \left(\frac{y_i}{\alpha_i}\right) \mathbf{v}_i \\ &= \frac{1}{\alpha_1} \sum_{i=1}^{r} \left(\frac{\alpha_1}{\alpha_i}\right) y_i \mathbf{v}_i.\end{aligned}$$

Resumiendo

$$\mathbf{x} = \frac{1}{\alpha_1} \sum_{i=1}^{r} \left(\frac{\alpha_1}{\alpha_i}\right) y_i \mathbf{v}_i$$

En el caso de que $\frac{\alpha_r}{\alpha_1} \ll 1$, entonces, los errores al determinar $\mathbf{y}$ son amplificados por un factor igual a $\frac{\alpha_1}{\alpha_r}$ dando lugar a estimaciones ruidosas o de plano inútiles del vector solución $\mathbf{x}$.

2.5. Truncamiento de DSV y solución de sistemas de ecuaciones lineales.

Desde los cursos elementales de álgebra lineal es una sensación el considerar a los sistemas de ecuaciones lineales de la forma

$$A\mathbf{x} = \mathbf{b}$$

donde $A \in \mathbb{R}^{m\times n}$, y $\mathbf{b} \in \mathbb{R}^m$, como una parte simple y quizá hasta inocente de los métodos matemáticos. El problema es estimar al vector $\mathbf{x} \in \mathbb{R}^n$ que resuelva el sistema de ecuaciones lineales.

A continuación usaremos la herramienta que representa el poder descomponer la matriz A en la forma $U\Sigma V^t$ para estudiar la solución al sistema de ecuaciones lineales propuesto.

Si sustuimos la matriz A por su descomposicioón obtenemos que

$$U\Sigma V^t\mathbf{x} = \mathbf{b}.$$

Recordemos que la matriz $U \in \mathbb{R}^{m\times m}$ es una matriz ortogonal, es decir, $UU^t = I_d \in \mathbb{R}^{m\times m}$. De esta forma al multipliar la igualdad anterior por U^t a la izquierda, obtenemos

$$U^tU\Sigma V^t\mathbf{x} = U^t\mathbf{b},$$

de donde

$$\Sigma V^t\mathbf{x} = U^t\mathbf{b}.$$

Si escribimos $V^t\mathbf{x} = \hat{\mathbf{x}}$ y $U^t\mathbf{b} = \hat{\mathbf{b}}$, obbtenemos

$$\Sigma\hat{\mathbf{x}} = \hat{\mathbf{b}}.$$

A continuación consideraremos dos posibilidades: cuando $m \leq n$ y cuando $m > n$.

Primero consideremos el caso cuando $m \leq n$. Supongamos, como lo hemos hecho a lo largo de esta parte de las noas, que

$$\alpha_1 \geq \alpha_2 \cdots \geq \alpha_r > \alpha_{r+1} = \cdots = \alpha_n = 0.$$

ASÍ, si

$$\Sigma = \begin{pmatrix} \alpha_1 & 0 & \dots & 0 & \dots & 0 \\ & \alpha_2 & & & & \\ & & \ddots & & & \vdots \\ & & & \alpha_r & & \\ 0 & & 0 & 0 & & 0 \\ & & & & \ddots & 0 \\ & \dots & & & & \\ 0 & \dots & 0 & 0 & & 0 \end{pmatrix}$$

entonces la igualdad

$$\Sigma \hat{\mathbf{x}} = \hat{\mathbf{b}}$$

se traduce en el sistema

$$\begin{pmatrix} \alpha_1 & 0 & \dots & 0 & \dots & 0 \\ & \alpha_2 & & & & \\ & & \ddots & & & \vdots \\ & & & \alpha_r & & \\ 0 & & 0 & 0 & & 0 \\ & \dots & & & \ddots & 0 \\ 0 & \dots & 0 & 0 & & 0 \end{pmatrix} \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \\ \vdots \\ \\ \vdots \\ \hat{x}_n \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_1 \\ \vdots \\ \\ \vdots \\ \hat{b}_m \end{pmatrix}$$

de donde claramente se tiene que

$$\hat{x}_j = \frac{1}{\alpha_j} \hat{b}_j, \ \ j = 1, 2, \dots, r$$

y $\hat{x}_j$ toma valores arbitrarios para $j \geq k$.

Más generalmente, si escribimos, por una parte

$$V = [V_1 \, V_2],$$

donde V_1 y V_2 tienen r y $n - r$ columnas y por otra parte

$$U = [U_1 \, U_2],$$

donde U_1 y U_2 tienen r y $m - r$ columnas, entonces la solución general está dada por

$$\mathbf{x} = \sum_{j=1}^{r} \frac{1}{\alpha_j} \hat{b}_j \hat{v}_j + \sum_{j=r+1}^{n} \hat{x}_j v_j.$$

Si escribimos

$$\sum_{j=r+1}^{n} \hat{x}_j v_j \equiv \mathbf{x}_0.$$

entonces tenemos que

$$\mathbf{x} = V_1 \begin{pmatrix} \frac{1}{\alpha_1} & & & \\ & \frac{1}{\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\alpha_r} \end{pmatrix} U_1^t \mathbf{b} + \mathbf{x}_0,$$

con $\mathbf{x}_0 \in \mathcal{N}(A)$.

En los casos cuando $m = n$ y $m = n = r$ se tiene que la solución al problema está determinada de manera única, y en tales casos se tiene además que la solución está dada por

$$\mathbf{x} = V \begin{pmatrix} \frac{1}{\alpha_1} & & & \\ & \frac{1}{\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\alpha_r} \end{pmatrix} U^t \mathbf{b} = A^{-1}\mathbf{b}.$$

Por otra parte, si en la solución general tomamos $\mathbf{x}_0 \equiv 0$, entonces obtenemos la solución de *Norma Mínima*,

$$\begin{aligned} \mathbf{x} &= V_1 \begin{pmatrix} \frac{1}{\alpha_1} & & & \\ & \frac{1}{\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\alpha_r} \end{pmatrix} U_1^t \mathbf{b} \\ &= V \begin{pmatrix} \frac{1}{\alpha_1} & & & & & \\ & \frac{1}{\alpha_2} & & & & \\ & & \ddots & & & \\ & & & \frac{1}{\alpha_r} & & \\ & & & & 0 & \\ & & & & & 0 \end{pmatrix} U^t \mathbf{b} \end{aligned}$$

donde

$$V \begin{pmatrix} \frac{1}{\alpha_1} & & & & & \\ & \frac{1}{\alpha_2} & & & & \\ & & \ddots & & & \\ & & & \frac{1}{\alpha_r} & & \\ & & & & 0 & \\ & & & & & 0 \end{pmatrix} U^t \equiv A^\dagger$$

donde $A^\dagger$ es la pseudoinversa de la matriz A.

A continuación consideraremos el caso cuando $m > n$. Es decir, cuando la matriz Σ tiene la forma

$$\Sigma = \begin{pmatrix} \alpha_1 & & & \\ & \alpha_2 & & \\ & & \ddots & \\ & & & \alpha_n \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{pmatrix}$$

donde, como lo hemos considerado hasta el momento, se satisfacen las desigualdades

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_r > \alpha_{r+1} = \cdots = \alpha_n = 0.$$

La ecuación $\Sigma\hat{\mathbf{x}} = \hat{\mathbf{b}}$ está descrita por el sistema

$$\begin{pmatrix} \alpha_1 & & & \\ & \alpha_2 & & \\ & & \ddots & \\ & & & \alpha_n \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \vdots \\ \hat{x}_n \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \vdots \\ \hat{b}_m \end{pmatrix}$$

Y las soluciones al sistema están dadas por

$$\alpha_j \hat{x}_j = \hat{b}_j, \ 1 \leq j \leq r,$$

mientras que las entradas últimas entradas del vector $\mathbf{b}$ están dadas por $\hat{b}_j = 0$, con $j = r+1, \dots, m$.

El problema de cuando es posible satisfacer este último sistema de ecuaciones lineales no depende de $\hat{\mathbf{x}}$. Solo podemos determinar los primeros términos de la solución definidos por

$$\hat{x}_j = \frac{\hat{b}_j}{\alpha_j}, \ 1 \leq j \leq r.$$

Ahoa bien, al definir el vector $\mathbf{x}$ como

$$\mathbf{x} = V \begin{pmatrix} \frac{1}{\alpha_1} & & & & & & \\ & \frac{1}{\alpha_2} & & & & & \\ & & \ddots & & & & \\ & & & \frac{1}{\alpha_r} & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{pmatrix} U^t \mathbf{b}$$

Observemos que

$$V \begin{pmatrix} \frac{1}{\alpha_1} & & & & & & \\ & \frac{1}{\alpha_2} & & & & & \\ & & \ddots & & & & \\ & & & \frac{1}{\alpha_r} & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{pmatrix} U^t = A^\dagger.$$

De esta forma se satisface la ecuación

$$A\mathbf{x} = \mathbf{b}$$

de la mejor manera posible, minimizando la norma de la solución.

Recordemos que tenemos una matriz $A \in \mathbb{R}^{m\times n}$. Supongamos que $m \geq n$ y $\mathbf{b} \in \mathbb{R}^n$ es un vector dado.

DEFINICIÓN 2.5.1. El problema de **lineal de mínimos cuadrados** consiste en determinar un elemento del conjunto

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid \rho(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\| \text{es mínimo.}\}$$

De esta forma, en el caso $m > n$, la solución que hemos hallado es la solución de mínimos cuadrados la cual a su vez tiene norma mínima.

En todos y cada uno de los casos, la pseudo--inversa $A^\dagger$ de A está dada por

$$A^\dagger = V\Sigma^\dagger U^t \in \mathbb{R}^{n\times m}$$

donde

$$\Sigma^\dagger = diag\left\{\frac{1}{\alpha_1}, \frac{1}{\alpha_2}, \ldots, \frac{1}{\alpha_r}, 0, \ldots, 0\right\}$$

y α_r es el valor singular no nulo más peuqnho.

Antes de continuar, nos detendremos un poco para considerar el problema lineal de mínimos cuadrados y algunas de sus característcas.

2.6. Mínimos cuadrados

El problema de mínimos cuadrados surgió originalmente de la necesidad de ajustar un modelo lineal a un conjunto de observaciones. Con el fin de reducir la influencia de los errores en las observaciones puede resultar tentadora la idea de considerar una mayor cantidad de observaciones de manera que este número de obsevaciones sea mayor que el número de incógnitas existentes en el modelo. De esta manera tendríamos un sistema de ecuaciones lineales sobredeterminado modelado donde se busca determinar un vector $\mathbf{x} \in \mathbb{R}^n$ de manera que Ax, donde $A \in \mathbb{R}^{m\times n}$ sea la mejor aproximación posible al vector $\mathbf{b}$. Pero,

¿qué significa la expresión ‘‘mejor aproximación"?. En este sentido entendermos a un vector $\mathbf{x}$ que minimize el problema

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2, \quad A \in \mathbb{R}^{m\times n}, \ \mathbf{b} \in \mathbb{R}^m, \tag{2.4}$$

donde $\|\mathbf{x}\|_2$ representa la norma euclideana.

Al vector $\mathbf{x}$ que sea solución del problema de minimización anterior se le llama una solución del problema lineal de mínimos cuadrados $A\mathbf{x} = \mathbf{b}$. Si consideramos el vector

$$\mathbf{r}_\mathbf{x} = A\mathbf{x} - \mathbf{b} \tag{2.5}$$

llamado **vector residual**, donde $\mathbf{r} = (r_1, \ldots, r_m)$, entonces una solución de mínimos cuadrados minimiza la función

$$f(\mathbf{r}) = \|\mathbf{r}\|_2^2 = \sum_{i=1}^{m} r_i^2.$$

En el caso de que el rango de la matriz A sea menor a n, entonces la solución por mínimos cuadrados no es única. Sin embargo, de entre todas estas soluciones posibles del problema de minimzación existe sólo una solución que también minimize $\|\mathbf{x}\|_2$.

Antes de continuar, recordemos que una noción que resulta de gran utilidad en muchas situaciones es la de proyección de un vector $\mathbf{x}$ sobre otro vector $\mathbf{y}$ o sobre un subespacio W del espacio euclideano $\mathbb{R}^n$. En el primer caso, se tiene que la proyeción está la podemos obtener a partir de la igualad

$$\mathbf{x}^T\mathbf{y} = \|\mathbf{x}\|_2\|\mathbf{y}\|_2 \cos\theta$$

donde θ es el ángulo entre los vectores $\mathbf{x}$ y $\mathbf{y}$. Además si denotamos por $\mathbf{p}$ al vector que resulta de proyectar $\mathbf{x}$ sobre $\mathbf{y}$ entonces

$$\cos\theta = \frac{\|\mathbf{p}\|_2}{\|\mathbf{x}\|_2},$$

y en consecuencia

$$\|\mathbf{p}\|_2 = \frac{\mathbf{x}^T\mathbf{x}}{\|\mathbf{y}\|_2}.$$

Como los vectores $\mathbf{y}$ y $\mathbf{p}$ apuntan en la misma dirección, entonces

$$\mathbf{p} = proy_\mathbf{y}\mathbf{x} = \frac{\mathbf{x}^T\mathbf{y}}{\mathbf{y}^T\mathbf{y}}\mathbf{y}.$$

En el caso de querer obtener la proyección sobre un subespacio W de $\mathbb{R}^n$, donde $\{\mathbf{w}_1, \ldots, \mathbf{w}_l\}$ es una baso ortogonal de W, entonces la proyección de $\mathbf{x}$ sobre W se obtiene como la suma de las proyecciones de $\mathbf{x}$ sobre cada una de los elementos de la base de W; es decir,

$$\mathbf{p} = proy_W\mathbf{x} = \sum_{i=1}^{l} \frac{\mathbf{x}^T\mathbf{w}_i}{\mathbf{w}_i^T\mathbf{w}_i}\mathbf{w}_i.$$

Una de las propiedades interesantes de la proyección de un vector sobre un subespacio vectorial de $\mathbb{R}^n$ es que tal proyección es el punto de W más cercano al punto $\mathbf{x}$.

¿Qué tien que ver esta noción de proyección ortogonal con mínimos cuadrados lineales?. Supongamos que tenemos un sistema de ecuaciones lineales

$$A\mathbf{x} = \mathbf{b},$$

el cual es inconsistente. Si no es posible resolver el sistema de ecuaciuones lineales, ¿sería posible hallar una solución aproximada?. Una forma de resolver esta cuestión consiste en halla un vector $\bar{\mathbf{x}}$ de manera que minimizemos la distancia entre todas las imágenes $A\mathbf{x}$ y el vector $\mathbf{b}$, distancia dada por

$$\|A\mathbf{x} - \mathbf{b}\|.$$

Una solución que minimize este problema es lo que hemos llamado una solución por mínimos cuadrados. Una dificultad para hallar esta solución es que de acuerdo a lo comentado anteriormente es necesario conocer una bse ortogonal de vectores para $A(\mathbb{R}^n)$. Lo que podemos hacer en esta situación es resolver el problema

$$A\mathbf{x} = \mathbf{p} = pro\mathcal{y}_{A(\mathbb{R}^n)}\mathbf{b}. \tag{2.6}$$

De esta forma, el vector $A\mathbf{x} - \mathbf{b}$ es ortogonal a $A(\mathbb{R}^n)$. De esta forma este vector diferencia es ortogonal a cada uno de los vectores columna de la matriz A y en consecuencia

$$A^T(A\mathbf{x} - \mathbf{b}) = 0,$$

equivalentemente

$$A^T A\mathbf{x} = A^T\mathbf{b}. \tag{2.7}$$

De hecho, es claro que esta condición es necesaria y suficiente para que el vector $\mathbf{x}$ sea solución por mínimos cuadrados de la ecuación $A\mathbf{x} = \mathbf{b}$.

Observemos que $\mathbf{x} \in \mathcal{X}$ si y sólo si $A^t\mathbf{r} = 0$, donde $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ es el vector residual asociado con $\mathbf{x}$. Como $A^t\mathbf{r} = A^t(\mathbf{b} - A\mathbf{x}) = A^t\mathbf{b} - A^tA\mathbf{x}$, entonces la ecuación $A^t\mathbf{r} = 0$ puede reescribirse como $A^t\mathbf{b} = A^tA\mathbf{x}$. Esta última ecuación es llamada la **ecuación normal**. Dicho de otra forma, $\mathbf{x} \in \mathcal{X}$ si y sólo si $\mathbf{x}$ es solución de la ecuación normal.

Observemos que minimizar la distancia euclideana $\|A\mathbf{x} - \mathbf{b}\|$ es equivalente a determinar un vector $\mathbf{x}$ para el cual el vector $\mathbf{p} = A\mathbf{x}$ es el vector más cercano al vector $\mathbf{b}$. Para logar esto, se debe satisfacer que el vector residual $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ debe ser ortogonal al rango $\mathcal{R}(A)$ de A.E Es decir, si $A\mathbf{y}$ es cualquier elemento del rango de A entonces

$$\begin{aligned} 0 &= (A\mathbf{y})^t(\mathbf{b} - A\mathbf{x}) \\ &= \mathbf{y}^t A^t(\mathbf{b} - A\mathbf{x}) \\ &= \mathbf{y}^t(A^t\mathbf{b} - A^tA\mathbf{x}) \end{aligned}$$

pero al ser arbitrario el vector $\mathbf{y}$, entonces se debe satisfacer que $A^t\mathbf{b} - A^tA\mathbf{x} = 0$; es decir $A^t\mathbf{b} = A^tA\mathbf{x}$.

En el caso paricular de que A sea una matriz de rango completo (por columnas), entonces se tiene que $\mathbf{x} = (A^tA)^{-1}A^t\mathbf{b}$.

2.6.1. Factorización QR.

TEOREMA 2.6.1. *Sea $A \in \mathbb{R}^{m\times n}$. A puede factorizarse como un producto de la forma*

$$A = QR, \tag{2.8}$$

donde Q es una matriz ortogonal y R es una matriz triangular superior

Es común tener la situación de sistemas de ecuaciones lineales sobredeterminados, es decir, con más ecuaciones que incógnitas, es decir, donde $m > n$ con rango maximal por columnas. En tal caso las matrices R y Q toman, respectivamente, las formas dadas por

$$R = \begin{pmatrix} R_1 \\ \mathbf{0} \end{pmatrix}, \text{ con } R \in \mathbb{R}^{n\times n} \tag{2.9}$$

y

$$Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}, \text{ con } Q_1 \in \mathbb{R}^{m\times n} \text{ y } Q_2 \in \mathbb{R}^{m\times(m-n)}. \tag{2.10}$$

TEOREMA 2.6.2. *Sean $A \in \mathbb{R}^{m\times n}$ con $m > n$ y $ran(A) = n$ y $A = QR$ su factorización QR. Entonces*

- *Los vectores columna de Q_1 forman una base ortogonal de $\mathfrak{R}(A)$.*
- *Los vectores columna de Q_2 forman una base ortogonal de $\mathfrak{N}(A^T)$.*
- *La matriz R_1 es no singular.*

A continuación resolveremos el problema de mínimos cuadrados en este caso, es decir, queremos minimizar la distancia

$$\|A\mathbf{x} - \mathbf{b}\|_2.$$

Si Q es cualquier matriz ortogonal entonces Q y Q^T preservan la distancia euclideana, por lo que equivalentmente desamos calcular

$$\min \|Q^T(A\mathbf{x} - \mathbf{b})\|_2$$

Pero si $A = QR$, entonces

$$Q^T(A) = Q^T(QR) = (Q^TQ)R = IdR = R$$

por lo que

$$\min \|Q^T(A\mathbf{x}-\mathbf{b})\|_2 = \min \|R\mathbf{x}-Q^T\mathbf{b})\|_2 = \min \left\| \begin{array}{c} R_1\mathbf{x}-Q_1^T\mathbf{b} \\ \mathbf{0x}-Q_2^T\mathbf{b} \end{array} \right\|_2,$$

donde, el la última igualdad hemos hecho uso de las presentaciones (2.9) y (2.10) de R y Q, respectivamente.

Observemos que, independientemente del vector $\mathbf{x}$, siempre tendremos un error no nulo debido a la existencia de la componente $\mathbf{0x}-Q_2^T\mathbf{b}$ en el problema de mínimos cuadrados, la cual no es posible minimizar; pero lo que si podemos es resolver la ecuación

$$R_1\mathbf{x} = Q_1^T\mathbf{b}.$$

De esta forma, es posible minimizar el problema de mínimos cuadrados resolviendo este º'ultimo sistema de ecuaciones lineales. Una ventaja de usar la descomposición QR es debida al hecho de que es más sencillo resolver este sistema de ecuaciones lineales que resolver las ecuaciones normales, las cuales son casi siempre más mal condicionadas que el sistema a resolver obtenido mediante la descomposición QR de la matriz A.

2.6.2. Solución de mínimos cuadrados usando TDVS. Sea $A \in \mathbb{R}^{m\times n}$ con descomposición en valores singulares dada por

$$A = U\Sigma V^T = U_1\Sigma V_1^T. \qquad \text{Por Teorema 2.2.}$$

En esta situación tenemos que

$$\begin{aligned} \|A\mathbf{x}-\mathbf{b}\|_2^2 &= \|U\Sigma V^T\mathbf{x}-\mathbf{b}\|_2^2 && (2.11)\\ &= \|\Sigma V^T\mathbf{x}-U^T\mathbf{b}\|_2^2 && \text{Debido a que } U \text{ es ortogonal} \\ & && (2.12)\\ &= \|\Sigma\mathbf{z}-\mathbf{d}\|_2^2 && \text{donde } \mathbf{z}=V^T\mathbf{x} \text{ y } \mathbf{d}=U^T\mathbf{b} \\ & && (2.13)\\ &= \left\| \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix} - \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{pmatrix} \right\|_2^2 \\ &= \begin{pmatrix} S\mathbf{z}_1-\mathbf{d}_1 \\ -\mathbf{d}_2 \end{pmatrix} \\ &= \|S\mathbf{z}_1-\mathbf{d}_1\|_2^2 + \|-\mathbf{d}_2\|_2^2 \end{aligned}$$

Es decir,

$$\|A\mathbf{x}-\mathbf{b}\|_2^2 = \|S\mathbf{z}_1-\mathbf{d}_1\|_2^2 + \|-\mathbf{d}_2\|_2^2. \qquad (2.14)$$

Observemos que si tomamos $\mathbf{z}_1 = S^{-1}\mathbf{d}_1$ entonces podemos minimizar la última expresión del conjunto de igualdades anteriores, y

en consecuencia, hemos minimizado $\|A\mathbf{x} - \mathbf{b}\|$, siendo el mínimo dado por $\|\mathbf{d}_2^2\|$, donde la componente $\mathbf{z}_2$ de $\mathbf{z}$ es arbitraria.

Antes de continuar, observemos que de la ecuación (2.13) se tiene que

$$\begin{aligned}\mathbf{d} &= U^T\mathbf{b}\\ &= \begin{pmatrix} U_1^T\mathbf{b} \\ U_2^T\mathbf{b} \end{pmatrix}\\ &= \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{pmatrix}.\end{aligned}$$

Como consecuencia de estas igualdades y de la ecuación (2.13) tenemos que

$$\begin{aligned}\mathbf{x} &= V\mathbf{z}\\ &= \begin{pmatrix} V_1 & V_2 \end{pmatrix}\begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix}\\ &= V_1\mathbf{z}_1 + V_2\mathbf{z}_2\\ &= V_1S^{-1}\mathbf{d}_1 + V_2\mathbf{z}_2\\ &= V_1S^{-1}U_1^T\mathbf{b} + V_2\mathbf{z}_2\end{aligned}$$

Es de notar que al ser arbitraria la segunda componente $\mathbf{z}_2$ del vector $\mathbf{z}$ entonces $V_2\mathbf{z}_2$ también resulta ser vector arbitario, el cual está contenido en $\mathcal{R}(V_2) = \mathcal{N}(A)$. De esta forma hemos logrado escribir el vector $\mathbf{x}$ en la forma

$$\mathbf{x} = A^\dagger\mathbf{b} + \left(Id - A^\dagger A\right)\mathbf{y}$$

donde el vector $\mathbf{y}$ es un vector arbitrario.

Finalmente, mínimo del residual de mínimos cuadrados está dado por

$$\|\mathbf{d}_2\|_2 = \|U_2^T\mathbf{b}\|_2.$$

De los argumentos anteriores tenemos un resultado que no podemos dejar pasar desapercibido.

PROPOSICIÓN 2.6.3. *Si* $\mathbf{x}$ *es solución del problema de mínimos cuadrados* $\|A\mathbf{x} - \mathbf{b}\|$ *entonces*

$$\mathbf{x} = A^\dagger\mathbf{b} + \left(Id - A^\dagger A\right)\mathbf{y}. \tag{2.15}$$

Asimismo, hemos obtenido distintas condiciones euivalentes para que el vector residual del problema de mínimos cuadrados se anulen, las cuales están dadas por

(1) El vector $\mathbf{b}$ es ortogonal a todos los vectores de U_2.

(2) El vector $\mathbf{b}$ es ortogonal a todos los vectores contenidos en $\mathcal{R}(A)^{\perp}$.
(3) El vector $\mathbf{b}$ está contenido en $\mathcal{R}(A)$.

De las igualdades

$$\begin{aligned} \| \left(Id - AA^{\dagger}\right) \mathbf{b}\|_2^2 &= \|U_2 U_2^T \mathbf{b}\|_2^2 \\ &= \mathbf{b} U_2 U_2^T U_2 U_2^T \mathbf{b} \\ &= \mathbf{b} U_2 U_2^T \mathbf{b} \\ &= \|U_2^T \mathbf{b}\|_2^2 \end{aligned}$$

se tiene que $\| \left(Id - AA^{\dagger}\right) \mathbf{b}\|$ es otra expresión para el vector residial del problema de mínimos cuadrados.

El *** no resuelve el problema de determinar $\mathbf{x}$ en $A\mathbf{x} = \mathbf{b}$ a partir de $\mathbf{b}$ cuando este vector contiene ruido. Una solución posible es usar un truncamiento de la descomposición en valores de la matriz de coeficientes A. (TSVD por sus siglas en inglés).

Para ver esto, consideremos que $\mathbf{x} \in \mathbb{R}^n$ es la solución verdadera del problema con ruido; decir, que $\mathbf{x}$ es tal que se satisface la igualdad

$$\mathbf{b} = A\mathbf{x} + \epsilon$$

donde $\mathbf{b}_0 = A\mathbf{x}$ es la señal sin ruido y ϵ es el ruido.

Supongamos que la descomposición en valores singulares de A está dada por

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

con $\Sigma = diag\{\alpha_1, \alpha_2, \ldots, \alpha_{min\{m,n\}}, 0, \ldots, 0\}$.

Sea $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ una base de V y escribamos al vector $\mathbf{x}$ en términos de esta base

$$\mathbf{x} = \sum_{j=1}^{n} \hat{x}_j \mathbf{v}_j$$

donde $\hat{x}_j = \mathbf{v}_j^t \mathbf{x}$ lo podemos ver como la componente del vector $\mathbf{x}$ en la dirección del elemento $\mathbf{v}_j$ de la base determinada por V.

DEFINICIÓN 2.6.4. Dado el vector $\mathbf{x}$ definimos el **truncamiento de orden** k de la descomposición en valores singulares como el vector dado por

$$\hat{\mathbf{x}}^{(k)} = \sum_{j=1}^{k} \frac{1}{\alpha_j} (\mathbf{u}_j^t \mathbf{b}) \mathbf{v}_j.$$

Observemos que la parte derecha de la igualdad dada en la definición anterior no es más que otra forma de escribir el producto matricial

$$[\mathbf{v}_1\, \mathbf{v}_2\, \ldots\, \mathbf{v}_k] \begin{pmatrix} \frac{1}{\alpha_1} & & & \\ & \frac{1}{\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\alpha_k} \end{pmatrix} [\mathbf{u}_1\, \mathbf{u}_2\, \ldots\, \mathbf{u}_k]^t\, \mathbf{b}.$$

Asimismo, si denotamos $V_k = [\mathbf{v}_1\, \mathbf{v}_2\, \ldots\, \mathbf{v}_k]$ y $U_k = [\mathbf{u}_1\, \mathbf{u}_2\, \ldots\, \mathbf{u}_k]^t$, entonces podemos escribir el mismo producto como

$$V_k \begin{pmatrix} \frac{1}{\alpha_1} & & & \\ & \frac{1}{\alpha_2} & & \\ & & \ddots & \\ & & & \frac{1}{\alpha_k} \end{pmatrix} U_k.$$

Sustituyendo $\mathbf{b} = A\mathbf{x} + \epsilon$ en esta expresión, tenemos que

$$\begin{aligned} \hat{\mathbf{x}}^{(k)} &= \sum_{j=1}^{k} \frac{1}{\alpha_j} (\mathbf{u}_j^t \mathbf{b}) \mathbf{v}_j \\ &= \sum_{j=1}^{k} \frac{1}{\alpha_j} (\mathbf{u}_j^t (A\mathbf{x} + \epsilon)) \mathbf{v}_j \\ &= \sum_{j=1}^{k} \frac{1}{\alpha_j} (\mathbf{u}_j^t (A\mathbf{x})) \mathbf{v}_j + \sum_{j=1}^{k} \frac{1}{\alpha_j} (\mathbf{u}_j^t \epsilon) \mathbf{v}_j \end{aligned}$$

Por otra parte tenemos que

$$\mathbf{u}_j^t A\mathbf{x} = \mathbf{u}^t \sum_{l=1}^{r} \mathbf{u}_l \alpha_l \left(\mathbf{v}_l^t \mathbf{x}\right) = \alpha_j \hat{x}_j;$$

por lo que obtenemos una expresón final para $\hat{\mathbf{x}}^{(k)}$, el truncamiento de orden k del vector $\mathbf{x}$, expresión dada por

$$\hat{\mathbf{x}}^{(k)} = \sum_{j=1}^{k} \hat{x}_j \mathbf{v}_j + \sum_{j=1}^{k} \frac{1}{\alpha_j} (\mathbf{u}_j^t \epsilon) \mathbf{v}_j.$$

Podemos usar esta expresión para $\hat{\mathbf{x}}^{(k)}$ con el fin de calcular la distancia de $\hat{\mathbf{x}}^{(k)}$ al vector $\mathbf{x}$. Para lograr esto calculemos primero $\mathbf{x} - \hat{\mathbf{x}}^{(k)}$:

$$\mathbf{x} - \hat{\mathbf{x}}^{(k)} = \sum_{j=k+1}^{n} \hat{x}_j \mathbf{v}_j + \sum_{j=1}^{k} \frac{1}{\alpha_j} (\mathbf{u}_j^t \epsilon) \mathbf{v}_j.$$

Además observemos que por una parte $\sum_{j=k+1}^{n} \hat{x}_j \mathbf{v}_j \in < \{\mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\} >$ y por otra $\sum_{j=1}^{k} \frac{1}{\alpha_j} (\mathbf{u}_j^t \epsilon) \mathbf{v}_j \in < \{\mathbf{v}_1, \ldots, \mathbf{v}_k\} >$, por lo que calcular la norma de esta diferencia es simple,

$$\begin{aligned} \|\mathbf{x} - \hat{\mathbf{x}}^{(k)}\|^2 &= \| \sum_{j=k+1}^{n} \hat{x}_j \mathbf{v}_j \|^2 + \| \sum_{j=1}^{k} \frac{1}{\alpha_j} (\mathbf{u}_j^t \epsilon) \mathbf{v}_j \|^2 \\ &= \sum_{j=k+1}^{n} \hat{x}_j^2 + \sum_{j=1}^{k} \frac{1}{\alpha_j^2} \left(\mathbf{u}_j^t \epsilon\right)^2. \end{aligned}$$

Ahora procedamos a analizar cada uno de estos dos sumandos. Por un lado, el primer sumando es completamente independiente del ruido ϵ. Además este primer término es llamado **bias** asociada a la estimación $\hat{\mathbf{x}}^{(k)}$, y se satisface que

$$\sum_{j=k+1}^{n} \hat{x}_j^2 \to 0 \text{ cuando } k \to n.$$

Con respecto al segundo término, este resulta ser una reconstrucción del ruido ϵ, reconstrucción que aumenta al tomar k cada vez más cercano a n. Si los últimos valores singulares son grandes, esta reconstrucción se sale de control, es decir, puede crecer mucho.

Una situación afortunada sería que pudiésemos elegir un valor de $k = k_{opt}$ de tal forma que

$$k_{opt} = \min_k \|\mathbf{x} - \hat{\mathbf{x}}^{(k)}\|^2.$$

Pero, nos encontramos con una situación problemática, para resolverlo necesitamos conocimeinto del valor verdadero $\mathbf{x}$. Ante esto, solo podemos hacer uso de soluciones aproximadas.

2.7. Principio de Discrepancia de Morozov

Uno de estos métodos es el llamado **Principio de Discrepancia de Morozov**. Con el fin de establecerlo, consideremos que se tiene una esimación de la norma del ruido, digamos que

$$\|\epsilon\| \lesssim \eta,$$

es decir, el verdadero valor $\mathbf{x}$ es tal que

$$\|A\mathbf{x} - \mathbf{b}\| \lesssim \eta,$$

y este es el único dato que conocemos a priori.

De esta forma, si hallamos una estimación $\hat{\mathbf{x}}$ de $\mathbf{x}$ de tal manera que $\|A\hat{\mathbf{x}} - \mathbf{b}\| \lesssim \eta$, entonces tomando en consideración nuestra información obtenida a priori, tenemos que esta estimación $\hat{\mathbf{x}}$ es factible. Por otra parte, al aumentar el valor de k en la estimación $\hat{\mathbf{x}}^{(k)}$ esto nos llevará a la obtención de ruido, por lo que se hace necesario hallar una k lo más pequegna posible.

El Principio de Discrepancia de Morozov nos dice que debemos elegir la k mº'as pequegna de tal forma que las siguientes igualdades seab satisfechas:

(1) $\|A\hat{\mathbf{x}}^{(k)} - \mathbf{b}\| \leq \eta$, y
(2) $\|A\hat{\mathbf{x}}^{(k+1)} - \mathbf{b}\| > \eta$.

De esta forma, lo que establece el principio de discrepancia de Morozov es toma k como el mínimo entero positivo para el cual la norma de la discrepancia $A\hat{\mathbf{x}}^{(k)} - \mathbf{b}$ es menor que la norma de los datos.

Inmediatamente surge la incógnita de si es posible elegir k de este modo. La respuesta es afirmativa en el caso de que $\mathbf{b} \in \mathcal{R}(A)$ como se muestra a continuación.

Recordemos que por una parte

$$A\mathbf{x} = \sum_{j=1}^{r} \mathbf{u}_j \alpha_j \left(\mathbf{v}_j^t \mathbf{x}\right),$$

y por otra

$$\mathcal{R}(A) = < \{\mathbf{u}_1, \ldots, \mathbf{u}_r\} > .$$

De esta forma, como $\mathbf{b} \in \mathcal{R}(A)$, entonces

$$\mathbf{b} = \sum_{j=1}^{r} \hat{b}_j \mathbf{u}_j.$$

Por lo que

$$\hat{\mathbf{x}}^{(r)} = \sum_{j=1}^{r} \frac{1}{\alpha_j} \hat{b}_j \mathbf{v}_j$$

(VERIFICAR ESTA IGUALDAD!!!!!!!!!!!) y en conscuencia,

$$A\hat{\mathbf{x}}^{(r)} = \sum_{j=1}^{r} \mathbf{u}_j \alpha_j \left(\mathbf{v}_j^t \left(\sum_{l=1}^{r} \frac{1}{\alpha_l} \hat{b}_l \mathbf{v}_l \right) \right) = \sum_{j=1}^{r} \mathbf{u}_j \alpha_j \frac{1}{\alpha_j} \hat{b}_j = \sum_{j=1}^{r} \hat{b}_j \mathbf{u}_j$$

para la norma de la discrepancia para el entero positivo k se tiene que

$$\|A\hat{\mathbf{x}}^{(k)} - \mathbf{b}\|^2 = \sum_{j=k+1}^{r} \hat{b}_j^2 \to 0 \text{ cuando } k \to r.$$

2.8. Método de la curva--L

Es otro método heurístico para elegir un buen valor del entero positivo k.

Observemos que la sucesión

$$\delta_k = \|A\hat{\mathbf{x}}^{(k)} - \mathbf{b}\|$$

de las normas de las discrepancias es una sucesión monótonamente decreciente; es decir,

$$\delta_0 = \|\mathbf{b}\| \geq \delta_1 \geq \cdots \geq \delta_r = 0.$$

Además, al considerar la sucesión ν_k de las normas de los k truncamientos $\hat{\mathbf{x}}^{(k)}$, se tiene que

$$\nu_k = \|\hat{\mathbf{x}}^{(k)}\| = \| \sum_{j=1}^{k} \frac{\hat{b}_j}{\alpha_j} \mathbf{v}_j \|^2,$$

y tal sucesión es una sucesión creciente. M'as aún,

$$\|\hat{\mathbf{x}}^{(k)}\|^2 = \sum_{j=1}^{k} \left(\hat{x}_j + \frac{\mathbf{u}_j^t \epsilon}{\alpha_j} \right),$$

por lo que si $\alpha_j \to 0$ entonces se tiene que el término debido al ruido comienza a dominar.

Al usar la función

$$k \to (\log \nu_k, \log \delta_k)$$

y graficar los pares

$$(\log \nu_k, \log \delta_k)$$

obtenemnos una gráfica en forma de letra "L"; de ahí el nombre de método de la curva--L.

Lo que ocurre en la "esquina" de la curva "L", si es que realmente teemos una curva de esta forma, es que la discrepancia no disminuye más y el ruido $\frac{\mathbf{u}_j^t \epsilon}{\alpha_j}$ toma el control. De esta forma este método lo que nos

sugiere es que hay que detenernos al momento de llegar a la esquina de la letra “L".

2.9. Aspectos computacionales con Matlab

```
%**************************************
%======================================
%      Deconvolucion 1-dimensional
%      Usando truncamiento de la
%  Descomposicion en Valores Singulares
%======================================
%**************************************

%%%% Consideraremos una seal de entrada "x" y una matriz A
%%%% a partir de las cuales tendremos una de salida "b",
%%%% El problema que plantearemos es el de la reconstruccin de
%%%% la seal original "x"


%=========================================
% %%CONSTRUCCION DE UNA SEAL UNIDIMENSIONAL
%=========================================

%La seal tendr N muestras, N puede tomar los valores 64, 128,...
N=64;
x = zeros(1,N);
x(round(N/4):round(3*N/4)) = 1;
%%% x(round(3*N/4):round(7*N/8)) = 2;
x = x(:); % Haremos "x" un vector vertical en lugar de uno horizontal

% SE PROCEDE A GRAFICAR LA SEAL
%------------------------------
figure(1)
clf
plot(1:N,x,'k','linewidth',2)
title('Seal por reconstruir x','Fontsize',16)
axis([1 N -.2 1.2])
set(gca,'ytick',[0 1 2])
set(gca,'xtick',[1 round(N/2) N])
axis square
drawnow
print -depsc -tiff DEConv_1_seal_original
pause
```

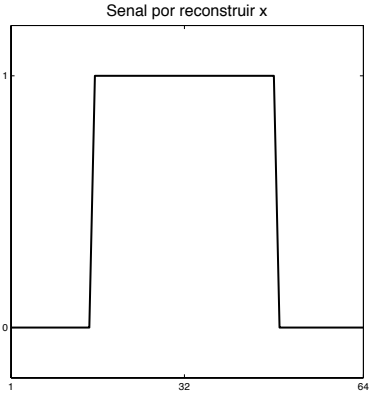

Figura 2.1. Señal original, por reconstruir.

```
%=========================================
% %%CONSTRUCCION DE LA MATRIZ (SPARSE) A
%=========================================
% Matriz A
M   = 4; % Width of PSF can vary, e.g. 1,2,3,...
tmp = linspace(0,1,M+1);
t   = [-fliplr(tmp(2:end)),tmp];
psf = exp(-4*t.^2);
psf = psf/sum(psf);
len = 2*M+1;
A   = convmtx(psf,N);
A   = A(:,(1+M):(end-M));
A
tamagnoA=size(A)

% Calculo de la DVS
%------------------
[U,D,V] = svd(A);
tamagnoD=size(D)
D

% Veamos la PSF
figure(3)
plot(t,psf,'k')
```

```
hold on
plot(t,psf,'k+')
title('Funcion de Dispersion del punto','Fontsize',16)
axis square
print -depsc -tiff DEConv_2_PSF
pause
```

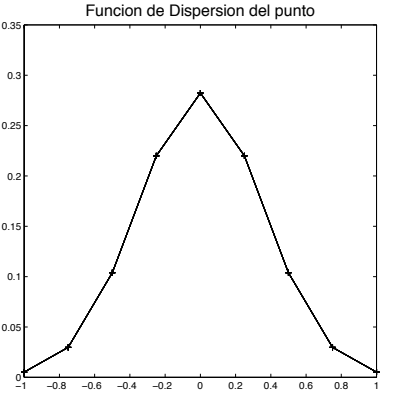

Figura 2.2. Función de dispersi'on del punto $h(z) = \exp(-4z^2)$.

```
%Geometria sparse de A
figure(4)
spy(A)
[row,col] = size(A);
axis([1 col 1 row])
set(gca,'xtick',[1 round(col/2) col])
set(gca,'ytick',[1 round(row/2) row])
axis square
title('Entradas no nulas de A','Fontsize',16)
print -depsc -tiff DEConv_3_sparseA
pause

%Grafica de los valores singulares de A
figure(5)
semilogy(diag(D))
xlim([1 N])
set(gca,'xtick',[1 round(N/2) N])
axis square
title('Valores singulares de A','Fontsize',16)
```

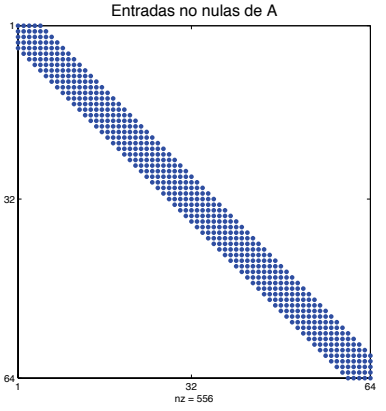

Figura 2.3. Geometría de la matriz sparse A.

```
drawnow
print -depsc -tiff DEConv_4_ValoresSingulares_A
pause
```

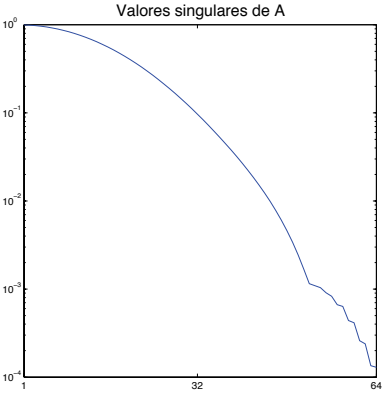

Figura 2.4. Valores singulares de A.

```
% Construyamos una version de "b" con ruido; i.e. ~b=b+ruido
b          = A*x;
nivelruido = 0.01; % Noise level can vary, e.g. 0.001, 0.1
br         = b + nivelruido*randn(size(b));
```

```
% Veamos b y b+ruido
figure(6)
clf
plot(1:N,b,'k','linewidth',2)
hold on
plot(1:N,br,'r','linewidth',2)
legend('b','br=b mas ruido')
axis([1 N -.2 1.2])
set(gca,'ytick',[0 1 2])
set(gca,'xtick',[1 round(N/2) N])
axis square
title('seal salida "b" y seal de salida con ruido "br"','Fontsize',16)
drawnow
pause
print -depsc -tiff DEConv_5_b_br
```

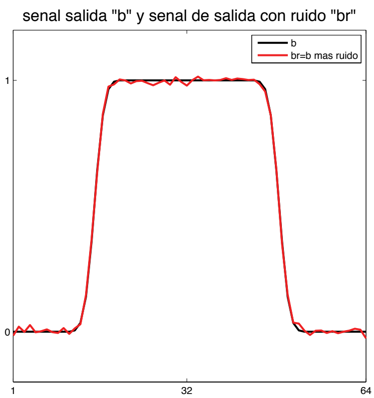

Figura 2.5. Señales de salida b y salida con ruido br.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Inversion usando DVS

% Reconstruccion "directa" de x:  x=A\Amr

recdir = A\br;
% Reconstruccion a partir de datos con ruido usando valores singulares
% grandes
% se descartan valores singulares menores que una tolerancia dada
tolerancia = .10;
```

```
rec       = pinv(A,tolerancia)*br;

% VEAMOS LA RECONSTRUCCION

figure(8)
plot(1:N,recdir,'r','linewidth',2)
title('Reconstruccion "directa" de seal con ruido x=A\ mr','Fontsize',
axis([1 N -.2 1.2])
set(gca,'ytick',[0 1 2])
set(gca,'xtick',[1 round(N/2) N])
set(gca,'xticklabel',{})
set(gca,'yticklabel',{})
axis square
print -depsc -tiff DEConv_6a_segnal_ruido_ReconstruccionDirecta
```

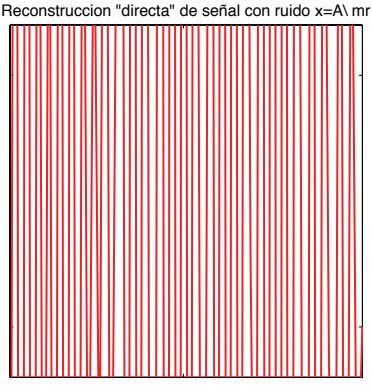

Figura 2.6. Reconstrucción $x = A \cdot br$.

```
figure(9)
plot(1:N,rec,'r','linewidth',2)
title(' Reconstruccion usando Truncamiento DVS','Fontsize',16)
axis([1 N -.2 1.2])
set(gca,'ytick',[0 1 2])
set(gca,'xtick',[1 round(N/2) N])
set(gca,'xticklabel',{})
set(gca,'yticklabel',{})
axis square
print -depsc -tiff DEConv_6b_segnal_TruncSVD
```

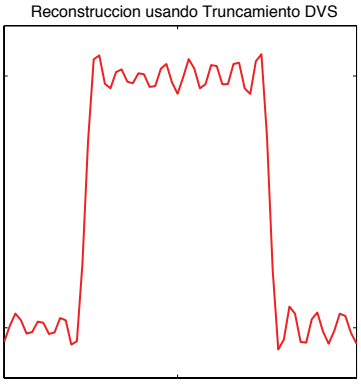

Figura 2.7. Solución de $x = A \cdot br$ usando DVS.

```
% VISTAZO A LOS VECTORES SINGULARES (vectores columna de la matriz ortogo

figure(10)
subplot(2,2,1)
plot(1:N,V(:,1),'k','linewidth',2)
title('Primer vector singular','Fontsize',16)
axis([1 N -.2 .4])
set(gca,'ytick',[0 1 2])
set(gca,'xtick',[1 round(N/2) N])
set(gca,'xticklabel',{})
set(gca,'yticklabel',{})
axis square
subplot(2,2,2)
plot(1:N,V(:,15),'k','linewidth',2)
title('Quinveavo Vector Singular','Fontsize',16)
axis([1 N -.2 .4])
set(gca,'ytick',[0 1 2])
set(gca,'xtick',[1 round(N/2) N])
set(gca,'xticklabel',{})
set(gca,'yticklabel',{})
axis square
subplot(2,2,3)
plot(1:N,V(:,end-15),'k','linewidth',2)
title('113avo. vector singular','Fontsize',16)
axis([1 N -.2 .4])
set(gca,'ytick',[0 1 2])
```

```
set(gca,'xtick',[1 round(N/2) N])
set(gca,'xticklabel',{})
set(gca,'yticklabel',{})
axis square
subplot(2,2,4)
plot(1:N,V(:,end),'k','linewidth',2)
title('ltimo vector singular','Fontsize',16)
axis([1 N -.2 .4])
set(gca,'ytick',[0 1 2])
set(gca,'xtick',[1 round(N/2) N])
set(gca,'xticklabel',{})
set(gca,'yticklabel',{})
axis square
print -depsc -tiff DEConv_7_vectores_propios
```

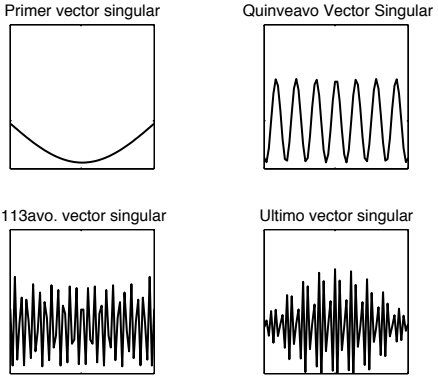

Figura 2.8. Algunos vectores propios de A.

Capítulo 3

Métodos Iterativos

TEOREMA 3.0.1. *Si $A \in R^{n\times n}$ es una matriz simétrica entonces A tiene una descomposición de la forma*

$$A = Q\Lambda Q^{-1} = Q\Lambda Q^T$$

donde Q es una matriz real ortogonal cuyos vectores columna son los vectores propios de A y Λ es una matriz diagonal real cuyos elementos de la diagonal son los valores propios de A.

Una *forma cuadrática* en $\mathbb{R}^n$ es una función

$$f : \mathbb{R}^n \to \mathbb{R}$$

de la forma

$$f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} \tag{3.1}$$

donde A es una matriz simétrica. Diremos que la forma cuadrática es *positiva definida* si

$$f(\mathbf{x}) \geq 0 \tag{3.2}$$

para toda $\mathbf{x}$ y el único punto donde f se anula es el origen. Siremos que la forma cuadrática es *positiva semidefinida* si la desigualdad (3.2) se satisface para cualquier vector $\mathbf{x}$. De manera smejante se dice que una matriz simérica $A \in \mathbb{R}^{n\times n}$ es una *matriz positiva definida* si la forma cuadraática asociada $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ es positiva definida. Diremos que la forma cuadrática es *negativa semidefinida* si la forma cuadrática $-f(\mathbf{x})$ es positiva semidefinida.

Sea A una matriz simétrica positiva definida, entonces el conjunto solución de la desigualad

$$(\mathbf{x} - \mathbf{q})A(\mathbf{x} - \mathbf{q}) \leq \epsilon$$

es una región elipsoidal centrada en el punto $\mathbf{q}$.

En este caso, es posible diagonalizar la matriz A, obteniendo una descomposición para esta de la forma

$$A = P\Lambda P^{-1},$$

donde los vectores columna de P se obtienen al normalizar los vectores propios de la matriz A y los elementos de la matriz diagonal Λ son los valores propios de A.

TEOREMA 3.0.2. *Una matriz A simétrica es positiva definida si y sólo si sus valores propios son no negativos.*

La conocida factorización de Choleski nos provee de otra forma de saber si una matriz simétrica es positiva definida.

TEOREMA 3.0.3. *(***Descomposición de Choleski***) Sea A una matriz simétrica positiva definida. Entonces A puede descomponerse en la forma*

$$A = R^T R,$$

donde R es una matriz triangular superior no singular. Más aún, una matriz tiene una factorización de esta forma solo si es positiva definida.

3.1. Subespacios de Krylov

El principal problema al que le hemos dedicado nuestra atención ha sido el aparentemente simple sistema de ecuaciones lineales

$$A\mathbf{x} = \mathbf{b}, \tag{3.3}$$

donde deseamos determinar la existencia del vector $\bar{\mathbf{x}}$ el cual satisfaga la ecuación (3.3), donde $A \in \mathbb{R}^{m\times n}$ y $\mathbf{b} \in \mathbb{R}^{n\times 1}$.

Dado un vector $\mathbf{x} \in \mathbb{R}^n$ definimos su *vector residual* asociado o simplemente *residual* como el vector

$$\mathbf{r}_{\mathbf{x}} = \mathbf{b} - A(\mathbf{x}). \tag{3.4}$$

En caso de no haber lugar a confusión alguna, denotaremos al residual como $\mathbf{r}$.

Para definir los *subespacios de Krylov* es necesario contar con una primera aproximación $\mathbf{x}_1$ de la solución $\bar{\mathbf{x}}$ ecuación (3.3). Para esta primera aproximación consideramos su residual

$$\mathbf{r}_1 = \mathbf{b} - A(\mathbf{x}_1). \tag{3.5}$$

Supongamos que este residual es un vector no nulo, ya que en caso contrario tendríamos la solución $\bar{\mathbf{x}} = \mathbf{x}_1$ de la ecuación (3.3).

Definimos el *espacio de Krylov de orden k* asociado al residual $\mathbf{r}_1$ y a la matriz A como el subespacio vectorial de $\mathbb{R}^n$ dado por

$$\mathfrak{K}_k(\mathbf{r}_1, A) := < \left\{ \mathbf{r}_1, A\mathbf{r}_1, A^2\mathbf{r}_1, \ldots, A^{k-1}\mathbf{r}_1 \right\} > . \tag{3.6}$$

Este subespacio se ha obtenido generando iteraciones de la forma $A^j(\mathbf{r}_1)$, $j = 0, \ldots, n-1$, es decir con productos de A^j con $\mathbf{r}_1$. La idea de

usar los subespacios de Krylov es buscar en este subespacio una nueva aproximación $\mathbf{x}_2$ de la solución $\bar{\mathbf{x}}$ de la ecuación (3.3).

En el caso de que el residual $\mathbf{r}_1$ sea un vector propio de A entonces $\dim \mathfrak{K}_k(\mathbf{r}_1, A) = 1$. Si $K \subset \mathbb{R}^n$ es un subespacio invariante $m-$dimensional, entonces $\dim \mathfrak{K}_k(\mathbf{r}_1, A) \leq m$.

3.2. Método del Gradiente Conjugado

De la experiencia en cursos de Métodos o Análisis Numérico un método iterativo no produce directamente la solución $\bar{\mathbf{x}}$ del sistema (3.3), sino que construye una sucesión de vectores $\mathbf{x}_k \in \mathbb{R}^n$ que se espera converja a una estimación eazonable para la solución de (3.3). El *Método del Gradiente Conjugado* es uno ejemplo de este tipo de métodos iterativos. Además, este método es el método más clásico donde se usa la técnica de los subespacios de Krylov.

Consideremos una matriz $A \in \mathbb{R}^{n\times n}$ simétrica (i.e. $A = A^T$) y positiva definida ($\mathbf{x}^T A\mathbf{x} > 0$ para $\mathbf{x} \neq \mathbf{0}$). Del Teorema de Descomposición en Valores Singulares 2.2 para el caso de una matriz simétrica tenemos que A tiene una representación de la forma

$$A = U\Sigma U^T,$$

donde $U \in \mathbb{R}^{n\times n}$ es una matriz ortogonal, $\Sigma = Diag\{d_1, d_2, \ldots, d_n\}$. Asimismo, la descomposición en valores singulares de A es también la descomposición en valores propios de A, pues si

$$U = [\mathbf{u}_1, \ldots, \mathbf{u}_n],$$

donde $\mathbf{u}_i$ son los vectores columna de U, entonces

$$A\mathbf{u}_j = d_j\mathbf{u}_j.$$

Observemos que

$$0 < \mathbf{u}_j^T A\mathbf{u}_j = \mathbf{u}_j^T(d_j\mathbf{u}_j) = d_j\mathbf{u}_j^T\mathbf{u}_j = d_j,$$

lo que muestra que todos los valores propios (valores singulares) son positivos.

La matriz A induce una norma

$$\|\cdot\|_A : \mathbb{R}^n \to \mathbb{R}$$

en $\mathbb{R}^n$, definida por

$$\|\mathbf{x}\|_A = \mathbf{x}^T A\mathbf{x}. \tag{3.7}$$

Como el conjunto $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ es una base ortogonal de $\mathbb{R}^n$, entonces todo elemento $\mathbf{x} \in \mathbb{R}^n$ tiene una descomposición de la forma

$$\mathbf{x} = \sum_{j=1}^{n} \tilde{x}_j\mathbf{u}_j, \text{ con } \tilde{x}_j = \mathbf{u}_j^T\mathbf{x}.$$

Así,

$$\begin{aligned}
\|\mathbf{x}\|_A^2 &= \mathbf{x}^T A\mathbf{x} \\
&= (\sum_{j=1}^{n} \tilde{x}_j \mathbf{u}_j)^T A(\sum_{k=1}^{n} \tilde{x}_k \mathbf{u}_k) \qquad (3.8)\\
&= (\sum_{j=1}^{n} \tilde{x}_j \mathbf{u}_j)^T (\sum_{k=1}^{n} A(\tilde{x}_k \mathbf{u}_k)) \qquad (3.9)\\
&= (\sum_{j=1}^{n} \tilde{x}_j \mathbf{u}_j)^T (\sum_{k=1}^{n} \tilde{x}_k A(\mathbf{u}_k)) \\
&= (\sum_{j=1}^{n} \tilde{x}_j \mathbf{u}_j)^T (\sum_{k=1}^{n} \tilde{x}_k d_k \mathbf{u}_k) \\
&= \sum_{j=1}^{n} d_j \tilde{x}_j^2,
\end{aligned}$$

Por lo que podemos pensar a la norma

$$\|\mathbf{x}\|_A^2 = \mathbf{x}^T A\mathbf{x} = \sum_{j=1}^{n} d_j \tilde{x}_j^2 \qquad (3.10)$$

como una norma euclideana con *pesos* en el marco ortogonal de coordenadas $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ para $\mathbb{R}^n$.

Si $\bar{\mathbf{x}}$ es la solución *verdadera* de la ecuación (3.3) tendríamos que $\bar{\mathbf{x}} = A^{-1}\mathbf{b}$. Todo funcionaría sin ninguna dificultad si la medición del vector $\mathbf{b}$ no incluyese errrores, pero en la práctica cotidiana este no es el caso.

Dado un vector $\mathbf{x} \in \mathbb{R}^n$ definamos el *vector error* como el vector dado por

$$\mathbf{e} = \bar{\mathbf{x}} - \mathbf{x}. \qquad (3.11)$$

Observemos que

$$A\mathbf{e} = A(\bar{\mathbf{x}} - \mathbf{x}) = A(\bar{\mathbf{x}}) - A(\mathbf{x}) = \mathbf{b} - A(\mathbf{x}) = \mathbf{r}.$$

En consecuencia podemos definir una función

$$\phi : \mathbb{R}^n \to [0, \infty)$$

dada por

$$\phi(\mathbf{x}) = \|\mathbf{e}\|_A^2, \qquad (3.12)$$

la cual mide el tamaño del error. Es claro que

$$\phi(\mathbf{x}) = \|\mathbf{e}\|_A^2 = \mathbf{e}^T A\mathbf{e} = \mathbf{e}^T\mathbf{r} = (A^{-1}\mathbf{r})^T\mathbf{r} = \mathbf{r}^T(A^{-1})^T\mathbf{r} = \mathbf{r}^T A\mathbf{r}.$$

Además, es inmediato que $\phi(\mathbf{x}) = \mathbf{0}$ si y sólo si $\mathbf{x} = \bar{\mathbf{x}}$.

De esta forma, si deseamos determinar $\bar{\mathbf{x}}$ entonces para resolver la ecuación (3.3) será suficiente con minimizar la función ϕ. Por otra parte, aún cuando para evaluar $\phi(\mathbf{x})$ es necesario conocer el vector $\bar{\mathbf{x}}$, para minimizar ϕ no necesitamos evaluar esta función.

Para construir el método del gradiente conjugado supongamos que tenemos una primera aproximación $\mathbf{x}_1$ de la solución $\bar{\mathbf{x}}$ de la ecuación (3.3). Su correspondiente residual está dado por $\mathbf{r}_1 = \mathbf{b} - A\mathbf{x}_1$. Supongamos que $\mathbf{r}_1 \neq \mathbf{0}$.

3.2.1. Direcciones de búsqueda. Definamos la primera *dirección de búsqueda* como la determinada por el residual asociado a la primera aproximación de la solución de (3.3); es decir, tomemos

$$\mathbf{s}_1 = \mathbf{r}_1. \tag{3.13}$$

A continuación buscaremos minimizar la función ϕ en la dirección de $\mathbf{s}_1$. Con el fin de lograr esto consideremos la restricción de ϕ a la recta

$$l_1(t) = \mathbf{x}_1 + t\mathbf{r}_1$$

Para minimizar ϕ restringida a l_1 debemos minimizar la composición $(\phi \circ l_1)(t)$. De la definición de ϕ tenemos que

$$\begin{aligned}
\phi(\mathbf{x}_1 + t\mathbf{s}_1) &= ([\mathbf{x}_1 + t\mathbf{s}_1] - \bar{\mathbf{x}})^T A([\mathbf{x}_1 + t\mathbf{s}_1] - \bar{\mathbf{x}}) \\
&= ([\mathbf{x}_1 - \bar{\mathbf{x}}] + t\mathbf{s}_1)^T A([\mathbf{x}_1 - \bar{\mathbf{x}}] + t\mathbf{s}_1) \\
&= [\mathbf{x}_1 - \bar{\mathbf{x}}]^T A[\mathbf{x}_1 - \bar{\mathbf{x}}] + [\mathbf{x}_1 - \bar{\mathbf{x}}]^T A[t\mathbf{s}] + [t\mathbf{s}_1]^T A[\mathbf{x}_1 - \bar{\mathbf{x}}] + [t\mathbf{s}_1]^T A[t\mathbf{s}_1] \\
&= [\mathbf{x}_1 - \bar{\mathbf{x}}]^T A[\mathbf{x}_1 - \bar{\mathbf{x}}] + 2t[\mathbf{s}]^T A[\mathbf{x}_1 - \bar{\mathbf{x}}] + [t\mathbf{s}_1]^T A[t\mathbf{s}_1] \\
&= [\mathbf{x}_1 - \bar{\mathbf{x}}]^T A[\mathbf{x}_1 - \bar{\mathbf{x}}] + 2t[\mathbf{s}]^T A[\mathbf{x}_1 - \bar{\mathbf{x}}] + t^2[\mathbf{s}_1]^T A[\mathbf{s}_1].
\end{aligned}$$

Como

$$A(\mathbf{x}_1 - \bar{\mathbf{x}}) = A(\mathbf{x}_1) - A(\bar{\mathbf{x}}) = A(\mathbf{x}_1) - \mathbf{b} = -\mathbf{r}_1,$$

entonces

$$\phi(\mathbf{x}_1 + t\mathbf{s}_1) = [\mathbf{x}_1 - \bar{\mathbf{x}}]^T A[\mathbf{x}_1 - \bar{\mathbf{x}}] - 2t\mathbf{s}_1^T\mathbf{r}_1 + t^2[\mathbf{s}_1]^T A[\mathbf{s}_1]$$

De esta forma, la derivada de la composición está dada por

$$\frac{d(\phi \circ l_1)(t)}{dt} = 2t\mathbf{s}_1^T A\mathbf{s}_1 - 2\mathbf{s}_1^T\mathbf{r}_1,$$

y esta se anula para

$$t = t_1 = \frac{\mathbf{s}_1^T\mathbf{r}_1}{\mathbf{s}_1^T A\mathbf{s}_1}.$$

Definimos la segunda aproximación $\mathbf{x}_2$ de $\bar{\mathbf{x}}$ por

$$\mathbf{x}_2 = \mathbf{x}_1 + t_1\mathbf{s}_1,$$

cuyo residual asociado está dado por

$$\mathbf{r}_2 = \mathbf{b} - A\mathbf{x}_2 = \mathbf{b} - A(\mathbf{x}_1 + t_1\mathbf{s}_1) = (\mathbf{b} - A\mathbf{x}_1) - t_1 A\mathbf{s}_1 = \mathbf{r}_1 - t_1 A\mathbf{s}_1$$

Observemos que $\mathbf{r}_2$ es ortogonal a la primera dirección de búsqueda $\mathbf{s}_1$:

$$\mathbf{s}_1^T\mathbf{r}_2 = \mathbf{s}_1^T(\mathbf{x}_1 + t_1\mathbf{s}_1) = \mathbf{s}_1^T(\mathbf{r}_1 - t_1 A\mathbf{s}_1) = \mathbf{s}_1^T\mathbf{r}_1 - t_1\mathbf{s}_1^T A\mathbf{s}_1 = 0.$$

A continuacón supongamos que se cuenta con una segunda dirección de búsqueda $\mathbf{s}_2$ y procedamos a minimizar la función ϕ restringida a la recta

$$l_2(t) = \mathbf{x}_2 + t\mathbf{s}_2.$$

Es decir, hay que determinar el valor de t para el cual la derivada de la composición $\phi \circ l_2$ se anula.

En general, en la k--ésima iteración tomamos

$$\begin{aligned} t_k &= argmin\, \phi(\mathbf{x}_k + t\mathbf{s}_k) \\ &= \frac{\mathbf{s}_k^T\mathbf{r}_k}{\mathbf{s}_k^T A\mathbf{s}_k} \quad k = 1, 2, \ldots \end{aligned}$$

Una posibilidad para la elección de la k--ésima dirección de búsqueda, que pareciera bastante natural, es tomar

$$\mathbf{s}_k = \mathbf{r}_k.$$

Esta elección de la dirección de búsqueda presenta una dificultad, la cual consiste en que $\mathbf{r}_{k+1}$ resultaría ser un vector ortogonal a la k--ésima dirección de búsqueda $\mathbf{s}_k$ debido a la forma que hemos tomado el escalar t_k. El problema es que esta elección de $\mathbf{s}_k$ hace que nuestro proceso iterativo sea bastante lento.

3.2.2. Direcciones A--conjugadas. Decimos que una familia $\mathcal{F}_k$ de vectores linealmente independiente

$$\mathcal{F}_k = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\} \tag{3.14}$$

es A-conjugada si para cada par de vectores distintos $\mathbf{s}_i$ y $\mathbf{s}_j$ de la familia $\mathcal{F}_k$ se tiene que

$$\mathbf{s}_i^T A\mathbf{s}_j = 0.$$

Por el momento, supongamos que se tiene una familia $\mathcal{F}_k$ de vectores de búsqueda dada por los vectores $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$ de manera que

I: la familia conste de vectores A--conjugados, y

II: $< \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\} >= \mathfrak{K}_k(\mathbf{r}_1, A)$.

A partir de los elementos de la familia $\mathcal{F}_k$ construimos una matriz $\mathrm{T}_k \in \mathbb{R}^{n\times k}$ cuyos vectores columna sean precisamente los elementos de $\mathcal{F}_k$:

$$\mathrm{T}_k = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k] \tag{3.15}$$

Esta matriz T_k es tal que $\mathrm{T}_k^T A \mathrm{T}_k \in \mathbb{R}^{k\times k}$ es una matriz diagonal, donde todos los elementos de la diagonal son números reales estrictamente positivos.

Para mostrar esto, consideremos la entrada (i, j) de la matriz T_k

$$\left(\mathrm{T}_k^T A \mathrm{T}_k\right)_{i,j} = \mathbf{s}_i^T A \mathbf{s}_j = \begin{cases} \lambda_i > 0, & \text{si } i = j; \\ 0, & \text{si } i \neq j. \end{cases}$$

De esta forma

$$\mathrm{T}_k^T A \mathrm{T}_k = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_k \end{pmatrix} = \Lambda \in \mathbb{R}^{k\times k}.$$

Con esta elección de la matriz T_k procedemos a minimizar la función ϕ, pero ahora restringida al conjunto de búsqueda de dirección dado por

$$\mathcal{S}_k = \mathbf{x}_1 + < \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\} > .$$

Observemos que todo elemento del conjunto $\mathcal{S}_k$ tiene la forma

$$\mathbf{x}_1 + t_1\mathbf{s}_1 + \cdots + t_k\mathbf{s}_k = \mathbf{x}_1 + \mathrm{T}_k\mathbf{t}$$

donde

$$\mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_k \end{pmatrix}.$$

Consideremos la función

$$\Gamma : \mathbb{R}^k \to \mathbb{R}$$

dada por

$$\Gamma(\mathbf{t}) = \phi(\mathbf{x}_1 + \mathrm{T}_k\mathbf{t}).$$

De la definición de la función ϕ, tenemos que

$$\begin{aligned}
\Gamma(\mathbf{t}) &= \phi(\mathbf{x}_1 + \mathrm{T}_k\mathbf{t}) \\
&= (\mathbf{x}_1 + \mathrm{T}_k\mathbf{t} - \bar{\mathbf{x}})^T A\,(\mathbf{x}_1 + \mathrm{T}_k\mathbf{t} - \bar{\mathbf{x}}) \\
&= ([\mathbf{x}_1 - \bar{\mathbf{x}}] + \mathrm{T}_k\mathbf{t})^T A\,([\mathbf{x}_1 - \bar{\mathbf{x}}] + \mathrm{T}_k\mathbf{t}) \\
&= (\mathbf{x}_1 - \bar{\mathbf{x}})^T A\,(\mathbf{x}_1 - \bar{\mathbf{x}}) + (\mathbf{x}_1 - \bar{\mathbf{x}})^T A\,(\mathrm{T}_k\mathbf{t}) \\
&\quad + (\mathrm{T}_k\mathbf{t})^T A\,(\mathbf{x}_1 - \bar{\mathbf{x}}) + (\mathrm{T}_k\mathbf{t})^T A\,(\mathrm{T}_k\mathbf{t}) \\
&= \phi(\mathbf{x}_1) + 2\,(\mathrm{T}_k\mathbf{t})^T A\,(\mathbf{x}_1 - \bar{\mathbf{x}}) + (\mathrm{T}_k\mathbf{t})^T A\,(\mathrm{T}_k\mathbf{t}) \\
&= \phi(\mathbf{x}_1) + 2\mathbf{t}^T\mathrm{T}_k^T A\,(\mathbf{x}_1 - \bar{\mathbf{x}}) + \mathbf{t}^T\left(\mathrm{T}_k^T A\mathrm{T}_k\right)\mathbf{t} \\
&= \phi(\mathbf{x}_1) + 2\mathbf{t}^T\mathrm{T}_k^T A\mathbf{r}_1 + \mathbf{t}^T\Lambda\mathbf{t} \\
&= \phi(\mathbf{x}_1) + \sum_{j=1}^{k}\left(-2t_j\mathbf{s}_j^T\mathbf{r}_1 + t_j^2\lambda_j\right).
\end{aligned}$$

Los cálculso anteriores muestran que el vector $\mathbf{t}$ para el cual se minimiza la función Γ está dado por

$$\mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_k \end{pmatrix}$$

donde

$$t_j = \frac{\mathbf{s}_j^T\mathbf{r}_1}{\mathbf{s}_j^T A\mathbf{s}_j}. \tag{3.16}$$

Es de notar que hemos obtenido los mismos valores (3.14) que cuando usamos las rectas de búsqueda l_k, pues también tenemos que

$$\begin{aligned}
\mathbf{s}_k^T\mathbf{r}_k &= \mathbf{s}_k^T(\mathbf{b} - A\mathbf{x}_k) \\
&= \mathbf{s}_k^T\left(\mathbf{b} - A\left[\mathbf{x}_{k-1} + t_{k-1}\mathbf{s}_{k-1}\right]\right) \\
&= t_{k-1}\mathbf{s}_k A\mathbf{s}_{k-1} + \mathbf{s}_k^T\mathbf{r}_{k-1} \\
&= \mathbf{s}_k^T\mathbf{r}_{k-1} \\
&= \ldots \\
&= \mathbf{s}_k^T\mathbf{r}_1.
\end{aligned}$$

En vista de toda la información que hemos obtenido, podemos decir que al usar direcciones que son A--conjugadas en la determinación de direcciones de búsqueda es posible minimizar iterativamente, haciéndolo solo una dimensión a la vez.

Por otra parte,

$$\mathbf{x}_{k+1} = \mathbf{x}_1 + \mathrm{T}_k\mathbf{t}, \tag{3.17}$$

por lo que

$$\begin{aligned}\mathbf{r}_{k+1} &= \mathbf{b} - A\mathbf{x}_{k+1}\\ &= \mathbf{b} - A\left[\mathbf{x}_1 + \mathrm{T}_k\mathbf{t}\right]\\ &= \mathbf{b} - A\mathbf{x}_1 - A\left[\mathrm{T}_k\mathbf{t}\right]\\ &= \mathbf{r}_1 - A\left[\mathrm{T}_k\mathbf{t}\right],\end{aligned}$$

en consecuencia,

$$\begin{aligned}\mathrm{T}_k^T\mathbf{r}_{k+1} &= \mathrm{T}_k^T\left[\mathbf{r}_1 - A\mathrm{T}_k\mathbf{t}\right]\\ &= \mathrm{T}_k^T\mathbf{r}_1 - \mathrm{T}_k^T A\mathrm{T}_k\mathbf{t}\\ &= \mathbf{0}.\end{aligned}$$

La última igualdad es consecuencia de la elección hecha para los valores t_j. Por lo que, además

$$\mathbf{r}_{k+1}\perp < \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\} > .$$

3.2.3. Elección de las direcciones de búsqueda. A continuación estableceremos la manera de determinar las direcciones de búsqueda para determinar las aproximaciones de la solución del sistema de ecuaciones lineales.

Tomemos

$$\mathbf{s}_1 = \mathbf{r}_1.$$

Este vector genera el primer subespacio de Krylov asociado a la matriz A y al residual $\mathbf{r}_1$; es decir,

$$\mathfrak{K}_1(\mathbf{r}_1, A) =< \{\mathbf{s}_1\} > .$$

Supongamos que ya han sido determinadas k direcciones de búsqueda $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k$ de tal manera que sean A--conjugadas y tales que

$$< \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\} >= \mathfrak{K}_k(\mathbf{r}_1, A).$$

Definamos el $k+1$--ésimo residual como

$$\mathbf{r}_{k+1} = \mathbf{b} - A\mathbf{x}_{k+1}.$$

Sin perder generalidad, podemos suponer que este evctor es no nulo, pues en caso contrario se tendría ya al vector $\bar{\mathbf{x}}$. Es decir, nuetra sucesión resultaría convegente a este vector.

Para determinar la $k+1$--ésima dirección de búsqueda $\mathbf{s}_{k+1}$ consideremosla de la forma

$$\mathbf{s}_{k+1} = \mathbf{r}_{k+1} + \rho\mathbf{s}_k. \tag{3.18}$$

Es impotante recordar que deseamos que todas las direcciones sean A--conjugadas, por ello es necesario que se satisfaga la igulad

$$\mathbf{s}_k^T A\mathbf{s}_{k+1} = 0,$$

pero por otro lado, si desarrollamos la parte izquierda de la igualad anterior se tiene que

$$0 = \mathbf{s}_k^T A\left(\mathbf{r}_{k+1} + \rho \mathbf{s}_k\right) = \mathbf{s}_k^T A\mathbf{r}_{k+1} + \rho \mathbf{s}_k^T A\mathbf{s}_k.$$

En consecuencia, se ha obtenido el valor de ρ para el cual el vector $\mathbf{s}_{k+1}$ definido por la igaldad (3.18) satisface las propiedades buscadas:

$$\rho = \rho_k = -\frac{\mathbf{s}_k^T A\mathbf{r}_{k+1}}{\mathbf{s}_k^T A\mathbf{s}_k}. \tag{3.19}$$

PROPOSICIÓN 3.2.1. *La dirección de búsqueda* $\mathbf{s}_{k+1}$ *tiene las siguientes propiedades para* $j < k$*:*

(1) $\mathbf{s}_j^T A\mathbf{s}_k = (A\mathbf{s}_j)^T \mathbf{r}_{k+1}$.
(2) $\mathbf{s}_j^T A\mathbf{s}_{j+1} = 0$.
(3) $< \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{k+1}\} >= \mathfrak{K}_{k+1}(\mathbf{r}_1, A)$.

DEMOSTRACIÓN 3.2.2. (1) La demostración de este inciso es simple,

$$\begin{aligned} \mathbf{s}_j^T A\mathbf{s}_k &= \mathbf{s}_j^T A\left(\mathbf{r}_{k+1} + \rho \mathbf{s}_k\right) \\ &= \mathbf{s}_j^T A\mathbf{r}_{k+1} + \rho \mathbf{s}_j^T A\mathbf{s}_k \\ &= \left(A\mathbf{s}_j\right)^T \mathbf{r}_{k+1}. \end{aligned}$$

(2) Para demostrar esta afirmación observemos que

$$A\mathbf{s}_j \in < \{\mathbf{s}_1, \ldots, \mathbf{s}_{k-1}\} >= A\left(\mathfrak{K}_{k-1}(\mathbf{r}_1, A)\right),$$

además

$$A\left(\mathfrak{K}_{k-1}(\mathbf{r}_1, A)\right) \subset \mathfrak{K}_k(\mathbf{r}_1, A) =< \{\mathbf{s}_1, \ldots, \mathbf{s}_k\} >,$$

por otra parte el connjunto

$$< \{\mathbf{s}_1, \ldots, \mathbf{s}_k\} >$$

es ortogonal a $\mathbf{r}_{k+1}$. Por lo que la afirmación se sigue inmediatamente.

(3) Finalmente,

$$\begin{aligned} < \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{k+1}\} > &=< \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k, \mathbf{r}_{k+1}\} > \\ &=< \mathfrak{K}_k(\mathbf{r}_1, A) \bigcup \{\mathbf{r}_{k+1}\} > \\ &= \mathfrak{K}_{k+1}(\mathbf{r}_1, A. \end{aligned}$$

Los primeros dos incisos del resultado anterior nos dicen que la nueva dirección de búsqueda $\mathbf{s}_{k+1}$ es A--conjugada a todas las direcciones anteriores $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k$.

Antes de continuar, haremos algunas observaciones con respecto al método que hemos desarrollado en estas páginas.

Primero, al calcular el valor t_k,

$$\mathbf{s}_k^T\mathbf{r}_k = \left(\mathbf{r}_k + \rho_{k-1}\mathbf{s}_{k-1}\right)^T \mathbf{r}_k = \mathbf{r}_k^T\mathbf{r}_k + \rho_{k-1}\mathbf{s}_{k-1}^T\mathbf{r}_k = \|\mathbf{r}_k\|^2$$

Con toda la información obtenida en estas páginas es posible dar la descripción de un algoritmo, el cual llamaremos *Gradiente Conjugado*. Pero antes, consideremos dos observaciones que nos ayudarán adar una mejor descripción a este algoritmo

A: Es posible dar otra expresión al producto $\mathbf{s}_j^T\mathbf{r}_j$. Y, en consceuencia obtener una nueva expresión para t_j. Recordemos que el prodcto interior de $\mathbf{s}_{j-1}$ y $\mathbf{r}_j$ se anula.

$$\mathbf{s}_j^T\mathbf{r}_j = (\mathbf{r}_j + \rho_{j-1}\mathbf{s}_{j-1})^T\mathbf{r}_j = \mathbf{r}_j^T\mathbf{r}_j + \rho_{j-1}\mathbf{s}_{j-1}^T\mathbf{r}_j = \|\mathbf{r}_j\|^2. \tag{3.20}$$

De esta forma, sustituyendo en la ecuación (3.16) la nueva expresión para $\mathbf{s}_j^T\mathbf{r}_j$ obtenemos la igualdad

$$t_k = \frac{\|\mathbf{r}_k\|^2}{\mathbf{s}_k^TA\mathbf{s}_k}. \tag{3.21}$$

B: Por otra parte, es posible obtener una nueva expresión para ρ_k, como se sigue de los siguientes cálculos

$$\begin{aligned}
\|\mathbf{r}_{k+1}\|^2 &= \mathbf{r}_{k+1}^T\mathbf{r}_{k+1} \\
&= \mathbf{r}_{k+1}^T(\mathbf{b} - A\mathbf{x}_{k+1}) \\
&= \mathbf{r}_{k+1}^T(\mathbf{b} - A\mathbf{x}_k - t_kA\mathbf{s}_k) \\
&= \mathbf{r}_{k+1}^T(\mathbf{r}_k - t_kA\mathbf{s}_k) \\
&= \mathbf{r}_{k+1}^T\mathbf{r}_k - t_k\mathbf{r}_{k+1}^TA\mathbf{s}_k \\
&= -t_k\mathbf{r}_{k+1}^TA\mathbf{s}_k
\end{aligned} \tag{3.22}$$

La última igualdad se sigue del hecho que por una parte $\mathbf{r}_k$ en elemento de $\mathcal{K}(\mathbf{r}_1, A)$, $\mathcal{K}(\mathbf{r}_1, A)$ es el subespacio generado por los vectores $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k$ y a su vez este conjunto es ortogonal a $\mathbf{r}_{k+1}$. De lo anterior se desprende qe $\mathbf{r}_{k+1}^T\mathbf{r}_k = 0$. En consecuencia,

$$t_k = -\frac{\|\mathbf{r}_{k+1}\|^2}{\mathbf{r}_{k+1}^TA\mathbf{s}_k} \tag{3.23}$$

Posteriormente se sustituye esta expresión para t_k en la ecuación (3.21), por lo que obtenemos la igualdad

$$\|\mathbf{r}_{k+1}\|^2 = -t_k\mathbf{r}_{k+1}^TA\mathbf{s}_k = -\frac{\|\mathbf{r}_k\|^2}{\mathbf{s}_k^TA\mathbf{s}_k}\mathbf{r}_{k+1}^TA\mathbf{s}_k$$

Para terminar esta segunda observación, se sustituye la expresión para ρ_k en (3.19), posteriormente se despeja ρ_k de la expresión así obtenida para, finalmente, obtener que

$$\rho_k = \frac{\|\mathbf{r}_{k+1}\|^2}{\|\mathbf{r}_k\|^2}. \tag{3.24}$$

3.3. Algoritmo GC para resolver $A\mathbf{x} = \mathbf{b}$

Para terminar este capítulo damos el algoritmo correspondiente al método del gradiente conjugado.

Primera aproximación: Consideremos

- Una primera aproximación $\mathbf{x}_1$ de la solución $\bar{\mathbf{x}}$ de $A\mathbf{x} = \mathbf{b}$.
- El residual $\mathbf{r}_1 = \mathbf{b} - A\mathbf{x}_1$ correspondiente a la primera aproximación $\mathbf{x}_1$, y
- La primera dirección de búsqueda $\mathbf{s}_1 = \mathbf{r}_1$.

Iteración: Se procede a iterar hasta que se satisfaga algún criterio de paro.

- $$t_k = \frac{\|\mathbf{r}_k\|^2}{\mathbf{s}_k^T A \mathbf{s}_k}$$
  ,
- $$\mathbf{x}_{k+1} = \mathbf{x}_k + t_k \mathbf{s}_k,$$
- $$\mathbf{r}_{k+1} = \mathbf{r}_k - t_k A \mathbf{s}_k,$$
- $$\rho_k = \frac{\|\mathbf{r}_{k+1}\|^2}{\|\mathbf{r}_k\|^2},$$
- $$\mathbf{s}_{k+1} = \mathbf{r}_{k+1} + \rho_k \mathbf{s}_k,$$
- $$k \leftarrow k + 1,$$
- Fin.

Capítulo 4

Algunos Problemas

(1) (a) Sea $A \in \mathbb{R}^{m\times n}$ tal que $A = UDV^T$ es su descomposición en valores singulares. Muestra que los valores propios de la matriz simétrica A^TA son los cuadrados de de los elementos de la matriz diagonal D, quiénes son sus respectuvos vectore propios?.

(b) Sea U una matriz ortogonal, muestra que U^T es también orotogonal.

(c) Usando el hecho de que una matriz ortogonal preserva la norma (euclideana) muetsra que $\|U^TAV\| = \|A\|$ para cualquier terna A, U, V donde A es cualquier matriz y U, V son ortogonales talea que el producto matricial tenga sentido.

(2) Si H es un subespacio vectorial de $\mathbb{R}^n$ y $P : \mathbb{R}^n \to H$ es la proyección ortogonal definido por la matriz P tal que

I: $P^2 = P$,

II: $(I - P)x$ es ortogonal a Px para todo $x \in \mathbb{R}^n$.

Considere una matriz $A \in \mathbb{R}^{m\times n}$. Exprese las proyecciones ortogonales $P_1 : \mathbb{R}^n \to N(A)$ y $P_2 : \mathbb{R}^n \to R(A)$ en términos de la DVS de A. Verifica la validez de la igualdad $R(A)^\perp = N(A^T)$ usando TDVS.

(3) Construye una matriz de tamaño 2×2 la cual transforme el círculo unitario en una elipse de semieje mayor igual a 2 y que apunta en la dirección del vector $(1, 2)$ y semiejer menor igual a $1/2$. Cuántos grados de libertad se tienen, Es decir, hay que caracterizar de grado de no-unicidad de estas matrices.

(4) Considera la ecuación $Ax = b$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^n$, al multiplicar esta ecuación a la izuierad por A^T se obtienen las llamadas *ecuaciones normales* $A^TAx = A^Tb$,donde la matriz A^TA es una matriz cuadrada, por lo que el sistema está formalmente determinado. Analiza las ecuaciones normales usando el TDVS aplicado a la matriz A. En particular, contesta las siguientes preguntas: Cuando podemos asegurar la invertibilidad de A^TA, en términos de los valores singulares

de A?. Cuál es la seudoinversa de la nueva ecuación? Cuál es la conexión entre la seudoinversa de A y la de $(A^TA)^\dagger A^T$?

(5) Determine condiciones de invertibilidad para AA^T en términos de los valores singulares de A. En tal caso $A^T(AA^T)^{-1}b$ es solución de $Ax = b$? Qué puedes decir respecto de $A^T(AA^T)$? Es igual a $(A^TA)^\dagger A^Tb$ y si no lo es,en general, bajo queé condiciones sí lo es?.

(6) Demuestre las ecuaciones de Moore-Penrose:
   (a) $A\dagger AA\dagger = A\dagger$,
   (b) $AA\dagger A = A$,
   (c) $(A\dagger A)^T = A\dagger A$,
   (d) $(AA\dagger)^T = AA\dagger$.

(7) Ecuaciones normales y mal-condicionamiento: Considere la matriz $A \in \mathbb{R}^{2\times 2}$, $A = UDU^T$, donde $U = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$, $D = \begin{pmatrix} 1 & 0 \\ 0 & 10^{-k} \end{pmatrix}$, donde $\theta = \pi/3$ y $k = 10$ (por ejemplo). Usa Matlab para calcular $A^{-1}b$ y $(A^TA)^TA^Tb$, con $b = (1,1)^T$. Explica porqué no son iguales. Cambia el valor de k y analiza lo que sucede.

(8) Sea f una función real definida en el intervalo $[0,\infty)$. La *transformada de Laplace* $\mathfrak{L}(f)$ de f está definida por la integral

$$\mathfrak{L}(f) = \int_0^\infty e^{-st} f(t)dt$$

siempre y cuando la integral sea convergente. Considera el siguiente problema: Dados los valires de la transformada de Laplace en los puntos s_j, $0 < s_1 < s_2 < \cdots < s_n < \infty$ es posible estimar la función f. Para tal fin, primero aproxima la integral que define a la transformada de Laplace por ,medio de una suma finita

$$\int_0^\infty e^{-s_j t} f(t)dt \approx \sum_{k=1}^{\infty} w_k e^{-s_j t_k} f(t_k),$$

donde los w_k son los *pesos* y los t_k's son los nodos de la regla de cuadratura usada (gaussiana, regla de Simson o trapezoidal). Sean $x_k = f(t_k)$, $y_j = \mathfrak{L}(f)(s_j)$ y $a_{jk} = w_k e^{-s_j t_k}$ y escribe la aproximación numérica de la transformada de laplace por $Ax = b$, donde A es una matriz cuadrada. En ete ejemplo escoja los datos distribuidos logarítmicamente, e.g.

$$\log(s_j) = \left(-1 + \frac{j-1}{20}\right)\log 10, \;\; j = 1, 2, \ldots, 40$$

con el fin de garantizar un muestreo denso cerca del origen.

Usa tu regla de cuadratura preferida con 40 nodos t_k en el intervalo [0,5]. Por lo que $A \in \mathbb{R}^{40\times 40}$.

Con el fin de generar los datos, considera la función o señal verdadera f dada por

$$f(t) = \begin{cases} t, & \text{si } t \in [0,1); \\ \frac{3}{2} - \frac{t}{2}, & \text{si } t \in [1,3); \\ 0, & \text{si } t \geq 3. \end{cases}$$

La Transformada de Laplace puede ser calculada analíticamente. Nosotros tenemos que $\mathfrak{L}(f) = \frac{1}{2s^2}\left(2 - 3e^{-s} + e^{-3s}\right)$. Muestra los detalles de este cálculo.

Para verificar el mal--condicionamiento de este problema intente estimar los valores $x_j = f(t_j)$ resolviendo directamente la ecuación $Ax = y$ usando el comando "backlash" de Matlab, usando los datos obtenidos analíticamente no le añadiremos error alguno a estos. Calcula le descomposición de valores singulares de A y muestra sus valores singulares. Añade un pequeño error a los datos y estime el valore de x a partir de los datos con ruido usando la regularización de Tikhonov,

$$x_\delta = argmin\left(\|Ax - y\|^2 + \delta^2\|x\|^2\right).$$

Intenta distintos valores del parámetro de regularización y calcula las correspondientes discrepancias.

(9) Implementa tu propio método de Gradiente Conjugado basado en los algoritmos descritos al final de las notes de Gradiente Conjugado. Prueba el progtama con una matriz cuadrada simple de tamaño $n \times n$, simétrica, positiva definida para ver si converge ba la solución exacta en n iteraciones.

(10) Regresa a las notas sobre TSVD e implemente la matriz de convolución con kernel gaussiano de tamaño 60×60, Calcule los datos usadndo la función "boxcar" usada en las notas y añade ruido (define el nivel de ruido como quieras). Obteniendo el modelo $b = Ax_* + e$. Aplica tu algoritmo de GC al problema lineal $Ax = b$ para aproximar iterativamente le solución. Sea x_k la k--ésima iteración. Sigue con cuidado la evolución del error y del residual graficando

$$\|e_k\| = \|x_k - x_*\|, \ \|r_k\| = \|Ax_k - b\|.$$

Se debe observar lo que se conoce como *semi--convergencia*. Primero el error comienza a disminuir para posteriormente volver a crecer.

(11) Selecciona el número óptimo de iteraciones usando el principio de discrepancia, i.e. detén el número de iteraciones cuando la

norma del residual es del orden de la norma del error. Grafica las estimaciones x_k correspondientes cerca del valor óptimo de paro: para poder comparar grafica también la L--curva. Esta L--curva da un criterio semejante para el valor óptimno de paro?.