



**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**  
PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y  
DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

PROBLEMAS LINEALES DISCRETOS MAL PLANTEADOS Y SUS  
APLICACIONES EN RESTAURACIÓN DE IMÁGENES DIGITALES

TESIS  
QUE PARA OPTAR POR EL GRADO DE  
MAESTRO EN CIENCIAS

PRESENTA  
IVÁN MÉNDEZ CRUZ

DIRECTOR DE TESIS  
DR. PABLO BARRERA SÁNCHEZ  
FACULTAD DE CIENCIAS

CIUDAD DE MÉXICO  
MÉXICO, D. F. ENERO DE 2017



Universidad Nacional  
Autónoma de México

Dirección General de Bibliotecas de la UNAM

**Biblioteca Central**



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



En Ciencia e Ingeniería, hay problemas donde queremos saber los efectos de un fenómeno a partir de sus causas. Para reproducir los efectos observados, usamos un modelo que describa el fenómeno. De ese modo, podemos pensar que se lleva a cabo un proceso en el que suministramos datos para obtener una respuesta.

Otros problemas que se presentan consisten en determinar las causas del fenómeno a partir de sus efectos. Estos son problemas inversos en relación a los primeros. Esta vez conocemos tanto el proceso como su respuesta, pero desconocemos que entrada debe recibir el proceso.

Es frecuente que pequeñas perturbaciones en los datos observados o en el sistema ocasionen cambios abruptos en la solución del problema inverso. Cuando esto pasa, decimos que el problema está mal planteado. Esta situación ocurre en el procesamiento de imágenes [81], tomografía [64], visión artificial [83], espectroscopía, etc. Por ejemplo si la imagen tomada por una cámara está desenfocada y tiene ruido blanco, cuando intentamos recuperar la versión ideal, la que obtenemos puede tener más degradaciones. Por lo que es importante detectar cuando un problema está mal planteado y cómo se debe proceder.

En el presente trabajo estamos interesados en hacer un estudio de los problemas lineales discretos mal planteados que nos permita dar las herramientas necesarias para analizarlos y saber que métodos podemos usar para resolverlos. Así mismo deseamos resaltar sus aplicaciones en la restauración de imágenes digitales.

Con conocimiento previo de la descomposición en valores singulares, el autor deseaba conocer cómo y dónde se puede usar esta herramienta del Álgebra Lineal. Esta tesis cubre esa inquietud y proporciona un panorama más amplio de sus aplicaciones. El estudio que realizamos está íntimamente relacionado con esta factorización matricial.

En el primer capítulo explicamos la teoría subyacente a los problemas lineales inversos mal planteados. Introducimos los conceptos básicos e ideas, y examinamos distintos ejemplos como la deconvolución, la transformada inversa de Laplace, entre otros. Damos una formulación matemática del problema a partir de un modelo dado por una ecuación integral de Fredholm de primera clase.

En el segundo capítulo, discretizamos la ecuación integral de Fredholm. Esto nos conduce a un problema lineal de cuadrados mínimos que está muy mal condicionado. Por eso los errores de redondeo, truncamiento, así como el ruido en las observaciones, afectan drásticamente las soluciones calculadas. Estos son problemas discretos mal planteados. Para analizar su condicionamiento usamos la descomposición en valores singulares.

En lugar de resolver directamente un problema mal planteado, tratamos de obtener soluciones aproximadas de una familia de problemas bien planteados cuya discretización sea mejor condicionada. Esta es la idea de los métodos de regularización. Las tareas principales son calcular la solución aproximada y elegir el valor del parámetro que determina el problema escogido de la familia. En el tercer capítulo examinamos estos métodos y los usamos

en ciertos problemas mal planteados. Hacemos comparaciones y vemos sus aplicaciones, después vemos los criterios establecidos para elegir el parámetro.

Finalmente, en el capítulo cuatro, entramos en el procesamiento de imágenes digitales. Veremos algunas ideas y conceptos básicos de esa área. Los problemas que presentamos son de gran escala. Debemos calcular decenas de miles de valores para generar una imagen. Abordamos dos: Deblurring [55] y Super-resolución [95]. En éstos nos dan una o más imágenes difuminadas y debemos obtener otra con menos degradación. Damos los modelos de observación y su discretización. En el caso del deblurring, remarcamos el papel que desempeña la convolución. Examinamos la estructura de ambos problemas, y explicamos cómo resolverlos. Nos interesa aplicar de manera eficiente los métodos de regularización para recuperar imágenes.

En la elaboración de este proyecto nos fueron de utilidad las referencias del Dr. Humberto Madrid y la presentación de la Dra. Ángela León Macías sobre procesamiento de imágenes. Para nuestros experimentos ocupamos la rutina `dgqt` de Moré [84], las bibliotecas `REGUTOOLS` [50] de Hansen y `HNO` [55] de Hansen, Nagy y O'Leary. Para los ejemplos creamos rutinas en MATLAB que pueden consultarse en la página web

[https://sites.google.com/site/vanmctwp/tesis\\_maestria](https://sites.google.com/site/vanmctwp/tesis_maestria).

## AGRADECIMIENTOS

---

Gracias a mis padres Artemia Cruz y Bulmaro Méndez por apoyarme y estar pendientes de mis estudios en la UNAM.

Deseo agradecer la ayuda y los consejos brindados por el Dr. Pablo Barrera Sánchez. Me ha guiado durante la maestría y en la elaboración de esta tesis, además de permitirme realizar mis actividades académicas en el Laboratorio de Cómputo Científico.

Agradezco al Dr. Humberto Madrid y al Dr. Jesús Lopez Estrada por su interés, observaciones y sugerencias.

Al Mat. Fco. Javier Aurrecochea por sus consejos y correcciones de estilo, y a la Act. Berenice Abigail por las correcciones de ortografía, así como al M. en C. Guilmer González por su apoyo y sugerencias.

Estoy agradecido con los servicios que me ha facilitado la UNAM y con el apoyo económico por parte de CONACYT.

A mis compañeros del Laboratorio: Adriana Rivera, Jorge Zavaleta, Gustavo Adolfo, Javier de Jesús, Leticia Ramírez, César Carreón por su amistad y solidaridad.

A mis amigos de la UNAM.



Prefacio	III
Agradecimientos	v
<b>1. Introducción a Problemas Mal Planteados</b>	<b>1</b>
1.1. Problemas Directos e Inversos . . . . .	1
1.1.1. El problema Inverso de la Convolución . . . . .	5
1.2. Problemas Mal Planteados . . . . .	17
1.3. Ecuaciones Integrales . . . . .	20
<b>2. Problemas Discretos Mal Planteados</b>	<b>25</b>
2.1. Métodos de Discretización . . . . .	25
2.1.1. Reglas de Cuadratura . . . . .	25
2.1.2. Método de Galerkin . . . . .	26
2.1.3. Método de Colocación . . . . .	33
2.2. Problemas de Cuadrados Mínimos . . . . .	35
2.2.1. Interpretación Geométrica . . . . .	38
2.3. Modelo Lineal General de Regresión . . . . .	39
2.4. Problemas de cuadrados mínimos con rango deficiente . . . . .	42
2.4.1. De la esfera al hiperelipsoide vía SVD . . . . .	42
2.4.2. SVD en el Problema de Cuadrados Mínimos . . . . .	43
2.4.3. Propiedades de la SVD . . . . .	47
2.4.4. Aplicaciones la SVD . . . . .	48
2.5. Condicionamiento . . . . .	50
2.5.1. Problemas Mal Condicionados . . . . .	50
2.5.2. Problemas Discretos Mal Planteados . . . . .	52
2.6. Expansión en Valores Singulares . . . . .	56
2.6.1. SVE y SVD . . . . .	59
2.6.2. Condición de Picard . . . . .	60
2.7. Condición Discreta de Picard . . . . .	61
<b>3. Regularización</b>	<b>67</b>
3.1. Introducción a la Regularización mediante SVD . . . . .	68
3.1.1. SVD Truncada . . . . .	68
3.1.2. SVD Selectiva . . . . .	74
3.2. Factores Filtro . . . . .	79
3.3. Regularización de Tikhonov . . . . .	82
3.3.1. Aplicación en Imágenes Digitales . . . . .	86

3.3.2.	Aplicación en el Ajuste de Curvas . . . . .	90
3.3.3.	Aplicación en Regresión Múltiple . . . . .	96
3.3.4.	Extensiones de la Regularización de Tikhonov . . . . .	102
3.4.	El Subproblema de Región de Confianza . . . . .	105
3.4.1.	Método de Región de Confianza . . . . .	105
3.4.2.	Características del TRS . . . . .	107
3.4.3.	Métodos para el TRS . . . . .	108
3.4.4.	TRS y Regularización . . . . .	110
3.4.5.	Gradiente Conjugado en el TRS . . . . .	117
3.5.	Elección del Parámetro de Regularización . . . . .	117
3.5.1.	Principio de Discrepancia . . . . .	119
3.5.2.	Criterio de la L-Curva . . . . .	123
3.5.3.	Validación Cruzada Generalizada . . . . .	127
<b>4.</b>	<b>Problemas de Gran Escala en Restauración de Imágenes</b>	<b>133</b>
4.1.	Deblurring . . . . .	137
4.1.1.	Modelo de difuminación . . . . .	138
4.1.2.	Digitalización de Imágenes . . . . .	147
4.1.3.	Modelo Discreto para Difuminación . . . . .	149
4.1.4.	Condiciones de frontera . . . . .	152
4.1.5.	Reducción del Problema . . . . .	161
4.1.6.	Regularización en el Problema de Deblurring . . . . .	169
4.2.	Super-resolución . . . . .	177
4.2.1.	El modelo . . . . .	177
4.2.2.	Discretización del modelo . . . . .	180
4.2.3.	Método para resolver el problema de la SR . . . . .	185
4.2.4.	Regularización en SR . . . . .	191
	<b>Conclusiones</b>	<b>197</b>
	<b>Apéndice</b>	<b>199</b>
	<b>Bibliografía</b>	<b>209</b>

El estudio de los problemas mal planteados comienza en el siglo XX. En 1902, el matemático francés Jacques Hadamard introdujo el término de problema bien planteado en su artículo *Sur les problèmes aux dérivées partielles et leur signification physique* [44], [78]. Para él, un problema está bien planteado si su solución existe y está unívocamente determinada. Posteriormente, mostró que el problema de Cauchy para la ecuación de Laplace no cumple con esto. Luego, R. Courant y D. Hilbert introducen el requerimiento de que la solución dependa continuamente de los datos para que el problema esté bien planteado.



**Figura 1.1:** Jacques Salomon Hadamard (1865-1963)

La teoría de problemas inversos mal planteados se ha ido desarrollando en las décadas pasadas conforme la comunidad científica se ha percatado de su importancia en las aplicaciones. En la actualidad, éstos tienen un carácter multidisciplinario [121]. Los avances en el cómputo científico han contribuido a su estudio.

En este capítulo presentamos ejemplos que dan lugar a problemas lineales mal planteados e introducimos los conceptos básicos. Comenzamos por examinar situaciones donde la relación causa-efecto juega un papel importante.

### 1.1. Problemas Directos e Inversos

Nos interesan situaciones donde tenemos el modelo matemático de un fenómeno concreto. En estos problemas tenemos dos conjuntos  $\mathcal{V}$  y  $\mathcal{W}$ , y un operador  $T : \mathcal{V} \rightarrow \mathcal{W}$ . Las causas del fenómeno están representadas por funciones  $f \in \mathcal{V}$ , mientras que los efectos observados son funciones  $g \in \mathcal{W}$ . La relación de causalidad está dada por  $T(f) = g$ .

Consideramos dos tipos. En el primero, dadas las causas del fenómeno, queremos saber cómo generar los efectos observados.

\* **Problema Directo.** Dada  $f \in \mathcal{V}$  y el operador  $T : \mathcal{V} \rightarrow \mathcal{W}$ , obtener  $g = T(f)$ .



El otro es conocer las causas del fenómeno a partir de sus efectos observados.

\* **Problema Inverso.** Dada  $g \in \mathcal{W}$  y el operador  $T : \mathcal{V} \rightarrow \mathcal{W}$ , hallar  $f \in \mathcal{V}$  tal que  $T(f) = g$ .

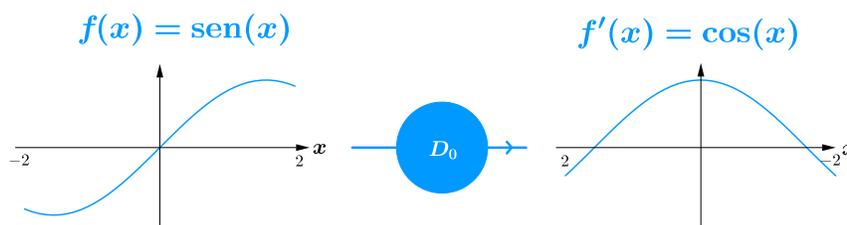


En álgebra, estadística, mecánica y en otras áreas de Matemáticas y Física aparecen problemas directos e inversos. En los ejemplos que se presentan, se plantea tanto el directo como el inverso. Ambos usan un modelo continuo y examinamos la sensibilidad de la solución respecto a pequeñas perturbaciones en los datos. Cuando se dispone solamente de algunas observaciones, trabajamos con un modelo discreto. En ese caso, se desea obtener valores aproximados de la solución.

**Ejemplo 1.1 (Diferenciación e integración).** Consideramos un ejemplo sencillo y muy sensible a perturbaciones. En la diferenciación, dada una función  $f : [-2, 2] \rightarrow \mathbb{R}$  continuamente diferenciable, aplicamos el operador

$$\mathcal{D}_0 : \begin{array}{ccc} C^1[-2, 2] & \rightarrow & C^1[-2, 2] \\ f & \mapsto & f' \end{array}$$

para hallar su derivada. En la Figura 1.2 se muestra como  $\mathcal{D}_0$  transforma la función seno en coseno.



**Figura 1.2:** El operador  $D_0$  realiza la diferenciación

❖ **Sensibilidad de su solución.** La diferenciación tiene la siguiente característica: *la derivada de una función cambia drásticamente con pequeñas perturbaciones en la función.*

Veamos esto con la función cuadrática

$$f(x) = x^2 + 1.$$

Si le agregamos una oscilación pequeña

$$r(x) = \sin(50x)/\sqrt{50},$$

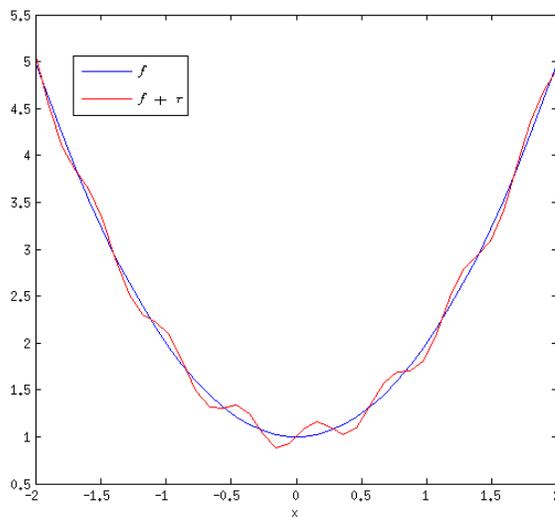
entonces  $f$  y  $f + r$  están a una distancia pequeña

$$\|(f + r) - f\|_\infty = 1/\sqrt{50}$$

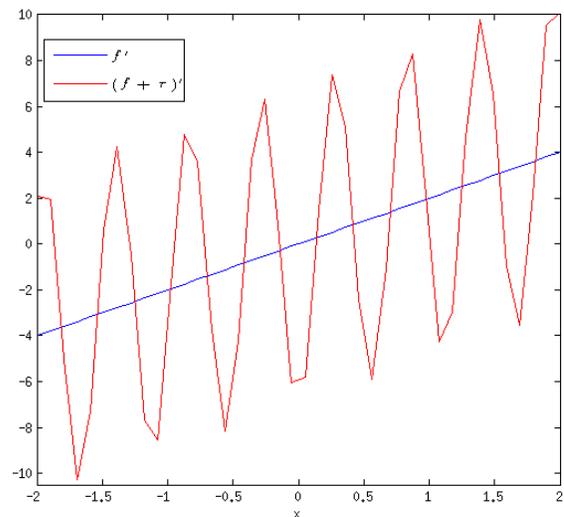
como se muestra en la Figura 1.3, mientras que las derivadas de  $f$  y  $f + r$  están a una distancia considerable

$$\|(f + r)' - f'\|_\infty = \sqrt{50}$$

En la Figura 1.4 vemos que los valores de la derivada de  $f + r$  se alejan de los respectivos valores de  $f'$ .



**Figura 1.3:**  $f(x) = x^2 + 1$  y su perturbación  $(f + r)(x) = x^2 + 1 + \sin(50x)/\sqrt{50}$



**Figura 1.4:**  $f'(x) = 2x$  y la perturbación  $(f + r)'(x) = 2x + \sqrt{50}\cos(50x)$

❖ **Problema inverso.** Ahora, dada una función  $q : [1, \infty) \rightarrow \mathbb{R}$ , queremos hallar su antiderivada, es decir, una función  $f : [1, \infty) \rightarrow \mathbb{R}$  tal que  $q = f'$ .

Consideremos el Operador

$$\mathcal{D} : C^1[1, \infty) \rightarrow C^1[1, \infty)$$

$$f \mapsto f'$$

La antiderivada de  $q$  es la función  $f$  que resuelve la ecuación

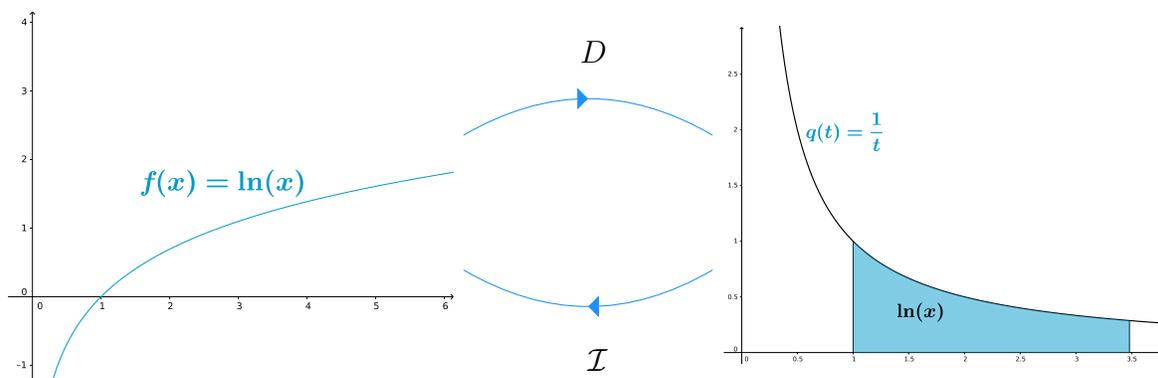
$$D(f) = q.$$

El inconveniente es que una sola función tiene una infinidad de antiderivadas, debido a que la derivada de una constante es cero. Luego, la Ecuación  $D(f) = q$  tiene infinidad de soluciones.

Supóngamos que conocemos  $f(1)$ . Entonces por el Teorema fundamental del Cálculo, tenemos que una antiderivada de  $q$  está dada por

$$f(x) = f(1) + \int_1^x q(t)dt, \quad a \leq x \leq b.$$

Esto da lugar al operador  $\mathcal{I} : C^1[1, \infty) \rightarrow C^1[1, \infty)$  que transforma la función  $q$  en su antiderivada. Este operador realiza un proceso de integración para hallar la antiderivada de  $q$ . En particular,  $\mathcal{I}$  transforma  $q(t) = 1/t$  en el logaritmo natural. Véase Figura 1.5.



**Figura 1.5:** El operador  $D$  realiza la diferenciación del logaritmo natural, mientras que la transformación  $\mathcal{I}$  realiza la integración de  $q(t) = 1/t$ .

La integración nos ofrece como solución  $f = \mathcal{I}(q)$ . En este sentido, a partir del Teorema fundamental del Cálculo podemos concluir que los procesos de diferenciación e integración son inversos uno del otro, más aún, la integración es un proceso de suavizamiento, ya que transforma funciones continuas en funciones diferenciables.

La solución del problema inverso dado por la ecuación  $T[f] = g$  puede no existir o no ser única. En general, la existencia y unicidad de su solución tiene que ver con que el operador  $T$  sea sobreyectivo o inyectivo.

**Observaciones 1.1:**

☞ En algunos casos, los elementos del conjunto  $\mathcal{W}$  no están en la imagen del conjunto  $\mathcal{V}$  bajo el operador  $T$ , es decir,  $g \notin T(\mathcal{V})$ . En ese caso, el problema inverso no tiene solución.

☞ Si hay más de una solución en el conjunto  $\mathcal{V}$ , el operador  $T$  no es inyectivo. Cuando  $T$  es inyectivo en  $\mathcal{S} \subsetneq \mathcal{V}$ , la restricción  $T|_{\mathcal{S}}$  es invertible y hay una única solución en  $\mathcal{S}$ .

**1.1.1. El problema Inverso de la Convolución**

Frecuentemente, los datos suministrados tienen errores. Por lo que un aspecto a considerar es cómo cambia la solución obtenida respecto a pequeños perturbaciones en los datos. Esto ocurre en áreas como en el procesamiento de señales.

Cuando una señal continua  $f$  presenta varios picos al estar contaminada por ruido, nos interesa tener una versión más suave de la misma. Para suavizarla, ocupamos otra señal conocida  $k$  que tenga mejores propiedades de suavidad, por ejemplo una función exponencial. Mediante una operación  $*$  realizamos un proceso de integración entre la señal original  $f$  y la auxiliar  $k$  que nos regresa una tercer señal  $g = f * k$  más suave que la original. Esta operación se conoce como convolución.

Algunos de los modelos de ingeniería electrónica y espectroscopia pueden representarse mediante la convolución. Así que en esas áreas, tenemos que tratar con el problema inverso de la convolución, llamado deconvolución. Por lo que nos parece interesante abordar este problema inverso. En el siguiente ejemplo examinamos la sensibilidad de la deconvolución de señales [123], [102].

**Ejemplo 1.2 (Deconvolución [53]).** Suavizamos la función  $f : [0, 1] \rightarrow \mathbb{R}$  dada por

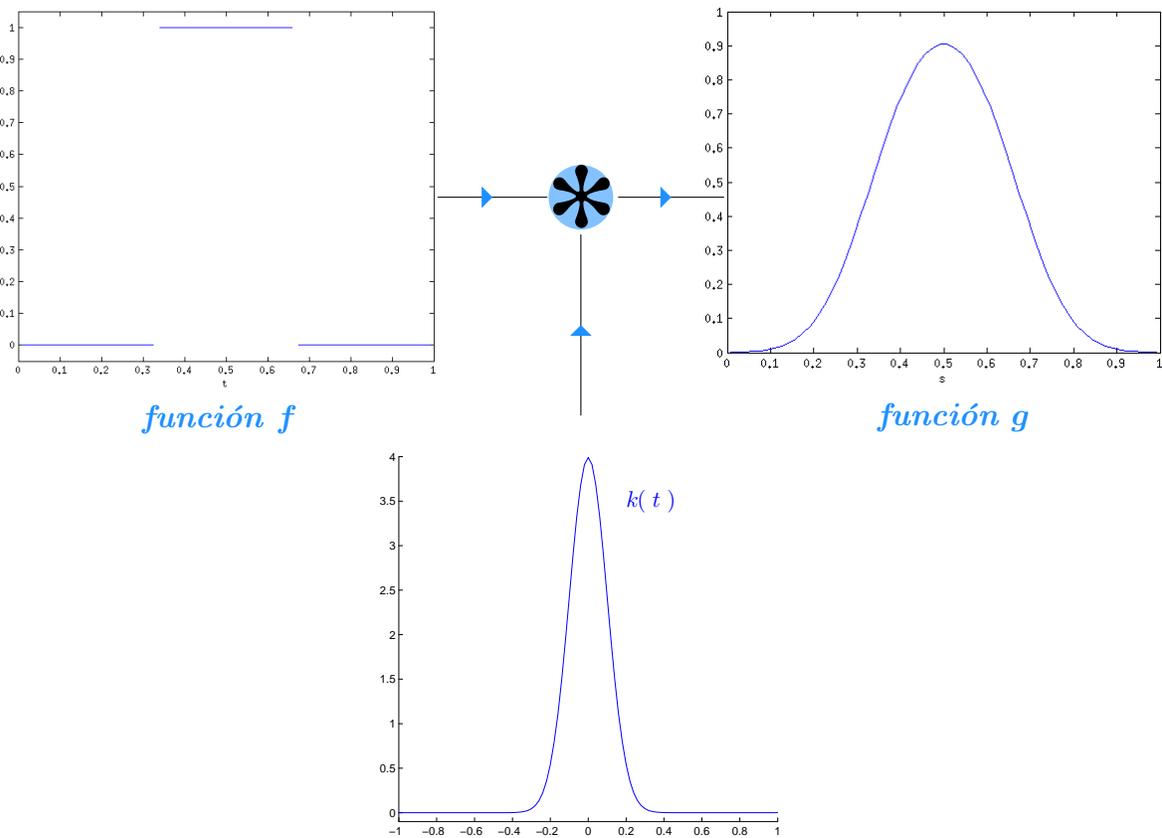
$$f(t) = \begin{cases} 1, & \text{si } \frac{1}{3} \leq t \leq \frac{2}{3}, \\ 0, & \text{en otro caso,} \end{cases}$$

Una manera de hacerlo es mediante la convolución con la función gaussiana

$$k(t) = \frac{1}{0.1\sqrt{2\pi}} \exp\left(-\frac{t^2}{0.02}\right), \quad -1 \leq t \leq 1,$$

La *convolución* de  $f$  con  $k$  está dada por

$$(f * k)(s) = \int_0^1 k(s-t)f(t)dt, \quad 0 \leq s \leq 1.$$



**Figura 1.6:** Convolución  $g = k * f$

La convolución nos produce una nueva función  $g : [0, 1] \rightarrow \mathbb{R}$  dada por  $g(s) = (f * k)(s)$ . Cada valor de  $g$  es el área debajo de la gráfica del producto de  $f$  con el reflejo desplazado de  $k$  respecto al eje vertical [53]. Por lo que el problema es:

Dadas las funciones  $f$  y  $k$ , obtener  $g = f * k$ .

Puesto que el producto de funciones integrables en  $[0, 1]$  es una función integrable sobre ese intervalo, tenemos que la convolución  $f * k$  da lugar al operador

$$T : L^1[0, 1] \rightarrow L^1[0, 1]$$

$$f \mapsto f * k$$

Así que buscamos  $g \in L^1[0, 1]$  tal que  $g = T(f)$ . En la Figura 1.6 mostramos la gráfica de la función  $g$  dada por la convolución de la función característica  $f$  con la gaussiana  $k$ .

Tomemos la partición uniforme de 21 puntos del intervalo  $[0, 1]$ . Obtenemos 20 subin-

tervalos  $I_j$  con puntos medios  $m_j$ :

$$I_j = \left[ \frac{j-1}{20}, \frac{j}{20} \right] \quad \text{y} \quad m_j = \frac{2j-1}{40}, \quad j = 1, \dots, 20$$

Además de la gaussiana  $k$ , supóngamos que solamente conocemos los valores de la función  $f$  en los puntos  $m_i$ . Queremos calcular los valores  $g_i$  de la función  $g$  dada por  $g = f * k$  en los puntos  $m_i$  para  $i = 1, \dots, 20$ . Este es el problema discreto de la convolución.

El integrando en la convolución  $f * k$  depende de las variables  $t$  y  $s$ . De éstas, solo  $t$  es variable de integración. Lo que hacemos es darle los valores  $m_i$  a la variable  $s$ . Con esto obtenemos 20 ecuaciones:

$$\begin{aligned} g_1 &= \int_0^1 k(m_1 - t)f(t)dt, \\ &\vdots \\ g_{20} &= \int_0^1 k(m_{20} - t)f(t)dt. \end{aligned}$$

Aproximamos el valor de las integrales por una suma ponderada de evaluaciones del integrando. En nuestro caso, empleamos la regla compuesta del punto medio. Por lo que los nodos en el intervalo  $[0, 1]$  son los puntos  $m_j$  y todos los pesos son iguales a  $1/20$ . Así que

$$\int_0^1 k(m_i - t)f(m_i)dt \approx \sum_{j=1}^{20} \frac{1}{20} k(m_i - m_j)f(m_j), \quad i = 1, \dots, 20.$$

De este modo, con las evaluaciones de la función  $k$  en las diferencias  $m_i - m_j$  y las de la función  $f$  en los puntos  $m_j$ , formamos el sistema de ecuaciones lineales

$$\frac{1}{20} \underbrace{\begin{bmatrix} k(m_1 - m_1) & \cdots & k(m_1 - m_{20}) \\ \vdots & & \vdots \\ k(m_{20} - m_1) & \cdots & k(m_{20} - m_{20}) \end{bmatrix}}_A \underbrace{\begin{bmatrix} f(m_1) \\ \vdots \\ f(m_{20}) \end{bmatrix}}_x = \underbrace{\begin{bmatrix} g_1 \\ \vdots \\ g_{20} \end{bmatrix}}_b$$

La matriz  $A$  de  $20 \times 20$  es simétrica porque la gaussiana  $k$  es una función par. El vector  $\mathbf{x}$  tiene los valores de  $f$  en los puntos  $m_j$ . Con la multiplicación de  $A$  por  $\mathbf{x}$ , obtenemos el vector  $\mathbf{b}$  con 20 observaciones de  $g$ . Así, obtenemos los valores de la convolución.

❖ **Problema inverso.** Ya vimos como podemos suavizar una función dada mediante la convolución con una gaussiana. Ahora, dadas las observaciones de una función suave  $g$ ,

estamos interesados en recuperar la función  $f$  que hemos suavizado. Para ello, supongamos que la función suave se obtiene a partir de la convolución de la gaussiana  $k$  con la función que buscamos. Esto da a la **deconvolución**:

Dadas de las funciones  $g$  y  $k$ , hallar la función  $f$  tal que  $f * k = g$ .

Supongamos que solamente conocemos las observaciones  $g_i$  de la función  $g$  en los puntos  $m_i$  para  $i = 1, \dots, 20$ . Queremos recuperar los valores de  $f$  en esos puntos a partir de la convolución  $f * k = g$ . Esta es la deconvolución discreta.

Sabemos que podemos obtener un vector con los valores de la convolución  $f * k$  por medio del producto matriz-vector  $\mathbf{b} = A\mathbf{x}$ . Ahora conocemos los valores  $g_i$  y resolvemos la ecuación  $A\mathbf{x} = \mathbf{b}$  para hallar el vector  $\mathbf{x}$  con  $x_j = f(m_j)$  para  $j = 1, \dots, 20$ .

En [96], Pasupathy y Damodar prueban que, salvo el factor  $1/(2\sqrt{2\pi})$ , la matriz  $A$  se factoriza como

$$A = LL^T,$$

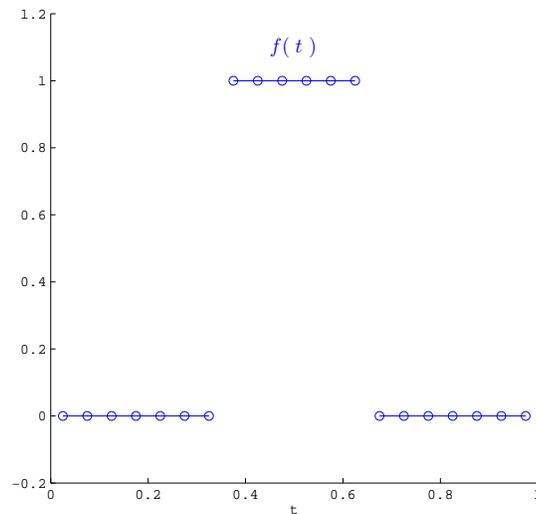
donde  $L$  es una matriz triangular inferior  $20 \times 20$  dada por

$$\begin{aligned} l_{i+1,1} &= \alpha^{i^2}, & i &= 0, \dots, 19, \\ l_{i+1,j+1} &= \frac{\alpha^{(i-j)^2} (1 - \alpha^{2i}) \dots (1 - \alpha^{2(i-j+1)})}{\sqrt{(1 - \alpha^2) \dots (1 - \alpha^{2j})}} & j &= 0, \dots, i. \\ \alpha &= \exp\left(-\frac{1}{2(20)^2(0.1)^2}\right) \end{aligned}$$

En la literatura [37], la factorización  $A = LL^T$  se llama **Factorización de Cholesky**. Puesto que  $\alpha \in (0, 1)$ , tenemos que los elementos en la diagonal principal de  $L$  son distintos de cero, y por consiguiente  $A$  es invertible. Resolvemos la Ecuación  $A\mathbf{x} = \mathbf{b}$  en dos fases. En la primera, resolvemos  $L\mathbf{y} = \mathbf{b}$  por sustitución directa, y en la segunda,  $L^T\mathbf{x} = \mathbf{y}$  por sustitución hacia atrás. Así, conseguimos la solución. En la Figura 1.7 mostramos la aproximación numérica que obtuvimos de los valores de  $f$ .

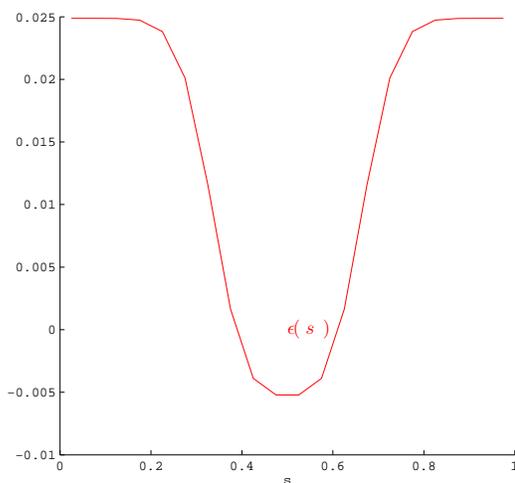
❖ **Sensibilidad de su solución.** A partir de la convolución  $f * k$ , recuperamos los valores de la función  $f$ . Veamos lo sensible que son los valores de esta función a cambios pequeños en las observaciones  $g_i$ . Para ver esto, perturbamos  $g$  con la función  $\epsilon : [0, 1] \rightarrow \mathbb{R}$  dada por

$$\epsilon(s) = 0.04 \exp\left(-0.125 \sum_{i=1}^{20} g(m_i)^2\right) \text{sinc}(0.5\pi g(s)).$$

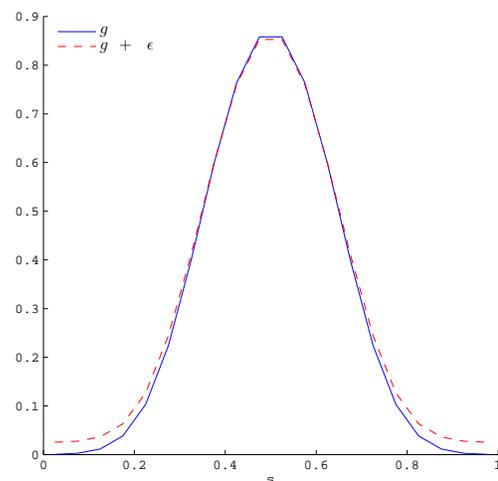


**Figura 1.7:** Valores de la función  $f$  recuperados en la convolución  $f * k = g$ .

En la Figura 1.8 mostramos la gráfica de esta función. Observamos que sus valores son del orden de  $10^{-2}$ . Por lo que la diferencia entre los respectivos valores de  $g$  y  $g + \epsilon$  es pequeña como puede verse en Figura 1.9.



**Figura 1.8:** Gráfica de función  $\epsilon$ .



**Figura 1.9:** Gráficas de funciones  $g$  y  $g + \epsilon$ .

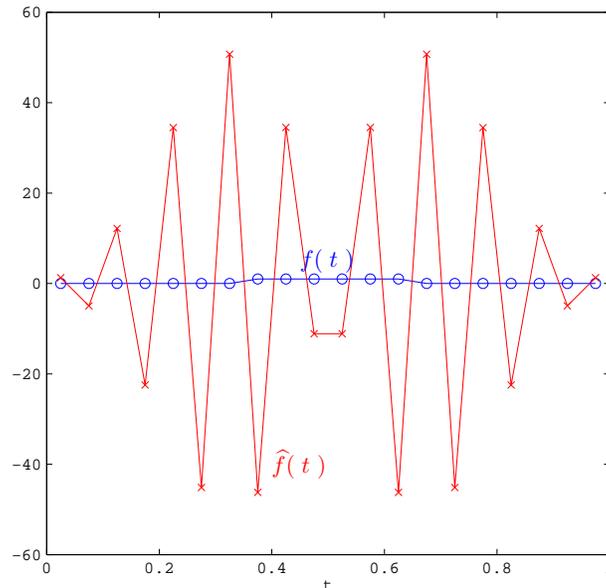
Disponemos de las observaciones de  $g + \epsilon$  en los puntos  $m_i$ . A partir de éstas, queremos calcular los valores de la función  $\hat{f} : [0, 1] \rightarrow \mathbb{R}$  tal que

$$(\hat{f} * k)(s) = g(s) + \epsilon(s), \quad 0 \leq s \leq 1.$$

Lo que hacemos es discretizar la convolución  $\hat{f} * k$  de la misma manera que hicimos sin perturbar  $g$ . Así, obtenemos el sistema de ecuaciones lineales

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

De aquí, una vez que resolvemos este sistema de ecuaciones lineales con ayuda de la Factorización de Cholesky, conseguimos los valores de  $\hat{f}$ . En la Figura 1.10 mostramos los valores de las funciones  $\hat{f}$  y  $f$  en los mismos puntos  $m_j$ . Notamos que  $\hat{f}(m_j)$  se aleja considerablemente de  $f(m_j)$ . Esto nos indica que pequeños errores en los observaciones producen cambios abruptos en los valores de la función  $f$  recuperada.



**Figura 1.10:** Valores de las funciones  $f$  y  $\hat{f}$  recuperadas en las convoluciones  $f * k = g$  y  $f * k = g + \epsilon$ , respectivamente.

En el siguiente ejemplo presentamos un problema inverso de dinámica de fluidos que obedece los principios de hidráulica y tiene aplicación en sistemas de riego tradicionales. El problema es medir el caudal del agua que pasa por un canal que lleva el agua de la cisterna a los cultivos. Los agrónomos pueden calcular el caudal mediante una estructura conocida como vertedero, que consiste en una pared delgada colocada de manera transversal al canal por donde pasa el agua. El vertedero que consideramos es un lámina delgada, donde la superficie tiene el mismo grosor y sus bordes son rectos. Véase Figura 1.11. Los vertederos más comunes tienen forma de V, trapezoidal o parabólica.

Con herramientas del cálculo y mecánica de fluidos, podemos dar un modelo que relaciona la forma del vertedero con el caudal. El principio básico está dado por la Ley de Torricelli, que relaciona la velocidad del agua con su altura en el canal mediante el principio de conservación de la energía:

$$\text{velocidad del agua} = \sqrt{2g_0 \times \text{altura del agua}},$$

donde  $g_0$  es la constante de aceleración de gravedad en caída libre.

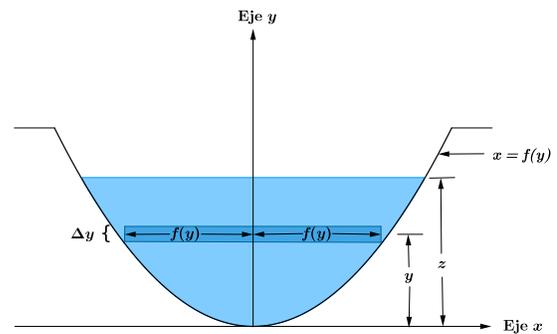
Nuestro interés está en que este modelo se representa mediante una convolución.

En el problema inverso, diseñamos el vertedero para corresponda con un rango de valores del caudal. Una vez construido, se calibra y los resultados se comparan con un estándar de la literatura [104].

**Ejemplo 1.3 (La forma del vertedero [42]).** En cultivos, el agua se almacena en un depósito. Al abrir una compuerta, dejamos pasar el agua a un canal. En medio del canal, colocamos una lámina (el vertedero) de una unidad de alto con una abertura de forma simétrica respecto a un eje vertical. Supóngamos que el agua no trae sedimentos ni se sale del canal, y que tiene una profundidad  $z > 0$ . Dada la forma de la abertura, queremos saber el volumen de agua que pasa por el área de la abertura en una unidad de tiempo, es decir, el *caudal* del agua.



**Figura 1.11:** Vertedero



**Figura 1.12:** Forma de la abertura

La forma del vertedero es una función  $f : (0, \infty) \rightarrow \mathbb{R}$  que depende de la altura  $y$ , mientras que el caudal es una función  $c : (0, \infty) \rightarrow \mathbb{R}$  que depende de la profundidad  $z$ . Por lo que debemos resolver el siguiente problema:

Determinar la función  $c$  que nos da el caudal a partir de la función  $f$  que nos da la forma de la abertura del vertedero.

Mediante la Ley de Torricelli, relacionamos el caudal del agua con la forma de la abertura. Para darnos cuenta de esta relación, vemos el volumen de agua que pasa por el vertedero en un pequeño intervalo de tiempo  $[t, t + \Delta t]$ . En ese lapso, el agua que atraviesa la abertura a una altura  $y$  forma un bloque de grosor  $\Delta y$ , ancho  $2f(y)$  y su largo es la velocidad del agua por  $\Delta t$ . En la Figura 1.12 mostramos una sección transversal de este bloque. Su volumen es

$$\Delta v = 2f(y) \times \text{velocidad del agua} \times \Delta y \Delta t.$$

De acuerdo a la Ley de Torricelli, la velocidad del agua en esa sección es

$$\sqrt{2g_0(z - y)},$$

En consecuencia,

$$\Delta v = 2\sqrt{2g_0(z - y)}f(y)\Delta y\Delta t.$$

Por lo que

$$\frac{\Delta v}{\Delta t} = 2\sqrt{2g_0(z - y)}f(y)\frac{\Delta y}{\Delta t}\Delta t.$$

El volumen y la profundidad del agua dependen del tiempo  $t$ . Así que para  $\Delta t$  suficientemente pequeño, denotado ahora por  $dt$ , tenemos que

$$\frac{dv}{dt} = 2\sqrt{2g_0(z - y(t))}f(y(t))\frac{dy}{dt}dt. \quad (1.1)$$

Supóngamos que la profundidad del agua al tiempo  $t = t_1$  es  $y = z$ . Entonces para sumar la contribución de todos los bloques de agua a distintas alturas  $y \in [0, z]$ , integramos ambos lados de la relación (1.1) sobre el intervalo  $[0, t_1]$ . Obtenemos

$$\int_0^{t_1} \frac{dv}{dt}dt = 2 \int_0^{t_1} \sqrt{2g_0(z - y(t))}f(y(t))\frac{dy}{dt}dt$$

Luego, con un cambio de variable, tenemos que

$$\int_0^{t_1} \frac{dv}{dt}dt = 2 \int_0^z \sqrt{2g_0(z - y)}f(y)dy.$$

Por otro lado, el caudal a la profundidad  $z$  está dado por todas las contribuciones de la derivada  $\frac{dv}{dt}$  desde  $t = 0$  hasta  $t = t_1$ :

$$c(z) = \int_0^{t_1} \frac{dv}{dt}dt.$$

En consecuencia, para  $g_0 = 32 \text{ ft/s}^2$ , se sigue que la función del caudal está dada por

$$c(z) = 16 \int_0^z \sqrt{(z - y)}f(y)dy.$$

### Observaciones 1.2:

 La función  $c$  que nos da el caudal es la convolución de la función de forma  $f$  con la función  $k : (0, \infty) \rightarrow \mathbb{R}$  dada por

$$k(y) = 16\sqrt{y}.$$

❖ **Problema inverso.** Hemos obtenido una expresión para el caudal del agua a partir de la forma que tiene la abertura del vertedero. Ahora, supongámonos que conocemos el caudal. Queremos diseñar el vertedero de modo que el cambio de volumen en la sección transversal de su abertura nos de los mismos valores que el caudal dado. Esto nos lleva al problema inverso:

Dada la la función  $c$  para el caudal del agua, determinar la función  $f$  que nos da la forma de la abertura del vertedero.

Supongámonos que el caudal del agua está dado por

$$c(z) = z^2.$$

Debido a que la función del caudal es la convolución  $c = k * f$ , aplicamos la deconvolución para encontrar la función de forma  $f$ . Una manera de resolver esto es mediante una transformación que nos permita manejar fácilmente la convolución.

Sean  $\mathcal{V}$  y  $\mathcal{W}$  espacios vectoriales de funciones. La **transformada de Laplace** es un operador  $\mathcal{L} : \mathcal{V} \rightarrow \mathcal{W}$  tal que a cada función  $f \in \mathcal{V}$ , le asocia la función  $\tilde{f} \in \mathcal{W}$  dada por

$$\tilde{f}(s) = \int_0^{\infty} e^{-sy} f(y) dy.$$

En particular,  $\mathcal{L}$  transforma  $k$  en

$$\tilde{k}(s) = \frac{8\sqrt{\pi}}{s^{3/2}}.$$

### Observaciones 1.3:

👉 Para que la transformada de Laplace de una función  $f$  exista es suficiente que sea continua a tramos y de orden exponencial sobre el intervalo  $(0, \infty)$ , esto es, que existan constantes  $M, \alpha > 0$  tales que para algún  $y_0 \geq 0$  se cumpla  $|f(y)| \leq Me^{\alpha y}$  para todo  $y \geq y_0$ . [107].

Una de las propiedades de  $\mathcal{L}$  es que transforma la convolución de las funciones  $k$  y  $f$  del espacio  $\mathcal{V}$  en el producto de las funciones  $\mathcal{L}[k]$  y  $\mathcal{L}[f]$  del espacio  $\mathcal{W}$ , esto es,

$$\mathcal{L}[k * f] = \mathcal{L}[k] \cdot \mathcal{L}[f]. \quad (1.2)$$

Entonces el operador  $\mathcal{L}$  transforma  $c = f * k$  en

$$\tilde{c} = \tilde{k} \tilde{f}.$$

Véase Figura 1.13. Así, con la introducción de la transformada de Laplace, tenemos:

Dadas funciones  $c, k : (0, \infty) \rightarrow \mathbb{R}$ , hallar una función continua  $f : (0, \infty) \rightarrow \mathbb{R}$  tal que

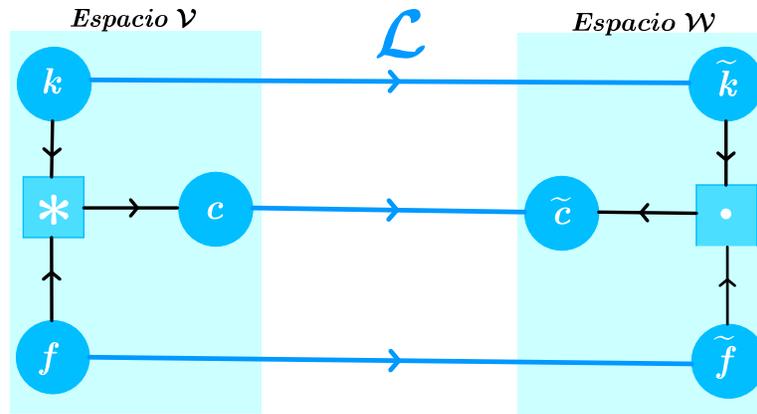
$$\mathcal{L}[f] = \frac{\mathcal{L}[c]}{\mathcal{L}[k]}. \quad (1.3)$$

La idea es expresar a  $f$  por medio de una convolución. Lo que hacemos es buscar funciones  $p$  y  $q$  tales que

$$\mathcal{L}[f] = \mathcal{L}[p] \cdot \mathcal{L}[q].$$

Pues de ser así, la identidad (1.2) implica que

$$\mathcal{L}[f] = \mathcal{L}[p * q].$$



**Figura 1.13:** Transformada de Laplace  $\tilde{c} = \tilde{k}\tilde{f}$  de la Convolución  $c = f * k$

El **Teorema de Lerch** [107] nos dice que dos funciones continuas distintas sobre  $(0, \infty)$  tienen diferentes transformadas de Laplace. En consecuencia, dos funciones continuas sobre  $(0, \infty)$  con misma transformada de Laplace son idénticas. Por lo que

$$f = p * q.$$

A partir de la función  $\tilde{k}$ , tenemos que la Ecuación (1.3) es

$$\mathcal{L}[f](s) = \frac{s^{3/2}\mathcal{L}[c](s)}{8\sqrt{\pi}}.$$

Por lo que las funciones  $p$  y  $q$  deben cumplir que

$$\mathcal{L}[p](s) \cdot \mathcal{L}[q](s) = \frac{s^{3/2}\mathcal{L}[c](s)}{8\sqrt{\pi}}.$$

Para obtener las funciones  $p$  y  $q$  a partir de esta expresión empleamos la propiedad de la transformada de Laplace para la segunda derivada:

$$L[c''](s) = s^2 L[c](s) - sc(0) - c'(0).$$

En nuestro caso  $c'(0) = c(0) = 0$ , pues el caudal está dado por la cuadrática  $c(z) = z^2$ , entonces

$$L[c''](s) = s^2 L[c](s).$$

Por consiguiente

$$\mathcal{L}[p] \cdot \mathcal{L}[q] = \frac{\mathcal{L}[c''](s)}{8\sqrt{\pi s}}.$$

Dado que

$$\mathcal{L}\left[\frac{1}{\sqrt{z}}\right] = \sqrt{\frac{\pi}{s}},$$

tomamos

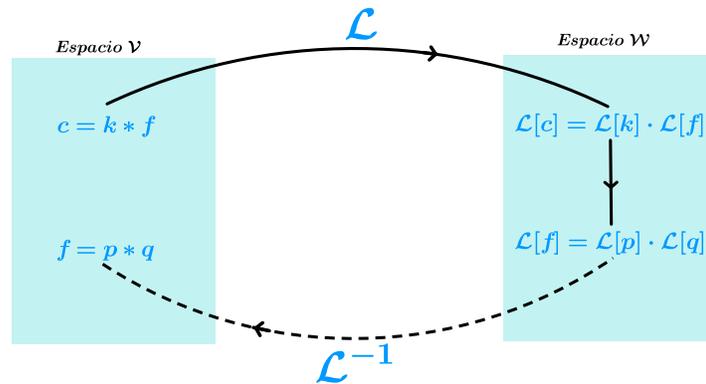
$$p(z) = c''(z) \quad \text{y} \quad q(z) = \frac{1}{8\pi\sqrt{z}}.$$

De este modo, encontramos funciones  $p$  y  $q$  tales que  $f = p * q$ , es decir,

$$f(y) = \frac{1}{8\pi} \int_0^y \frac{c''(z)}{\sqrt{y-z}} dz.$$

De aquí, podemos obtener la función  $f$  de forma a partir de la segunda derivada de la función del caudal. Así, al sustituir  $c(z) = z^2$ , tenemos que

$$f(y) = \frac{1}{4\pi} \int_0^y \frac{dz}{\sqrt{y-z}} = \frac{\sqrt{y}}{2\pi}.$$



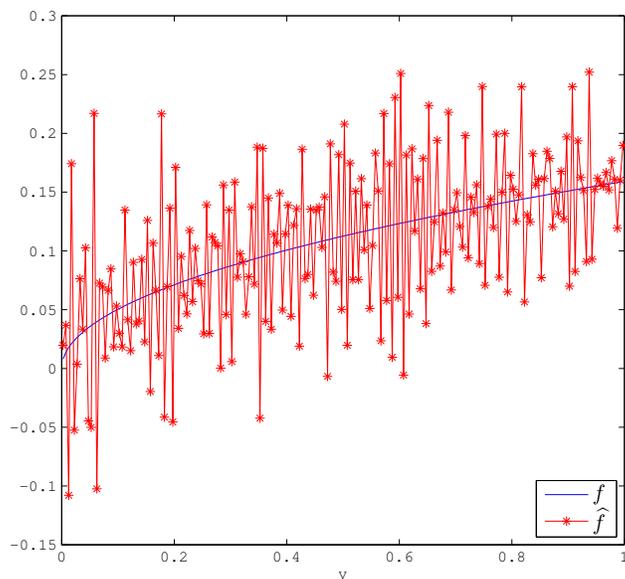
**Figura 1.14:** El operador  $\mathcal{L}$  transforma la convolución en un producto de funciones, y su inversa  $\mathcal{L}^{-1}$  transforma  $\mathcal{L}[p] \cdot \mathcal{L}[q]$  en  $p * q$ .

### Observaciones 1.4:

☞ El Teorema de Lerch nos sugiere que el operador  $\mathcal{L}$  posee una inversa  $\mathcal{L}^{-1} : \mathcal{W} \rightarrow \mathcal{V}$  que transforma  $\mathcal{L}[f]$  en  $f$  y transforma el producto de  $\mathcal{L}[p]$  con  $\mathcal{L}[q]$  en la convolución de  $p$  con  $q$  como se muestra en la Figura 1.14.

❖ **Sensibilidad de su solución.** Agregamos ruido gaussiano  $\epsilon$  de variación 0.01 al caudal  $c(z) = z^2$ . Sobre la malla uniforme de 200 puntos  $z_i$  del intervalo  $[0, 1]$  queremos calcular los valores de la función  $\hat{f} : \mathbb{R} \rightarrow \mathbb{R}$  que cumple

$$16 \int_0^z \sqrt{z-y} \hat{f}(y) dy = z^2 + \epsilon, \quad 0 \leq z \leq 1. \quad (1.4)$$



**Figura 1.15:** Valores de la función de forma  $f$  del vertedero y la función  $\hat{f}$  que resuelve la Ecuación (1.4) en los puntos medios  $m_i$ .

Aproximamos la integral con regla de cuadratura de punto medio en cada intervalo  $[0, z_i]$ . Denotamos a los puntos medios por  $m_i$ . Así, obtenemos el sistema de ecuaciones lineales

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

donde

$$b_i = z_i^2, \quad x_i = f(m_i), \quad a_{i,j} = \begin{cases} \frac{16}{200^{3/2}} \sqrt{i-j+\frac{1}{2}}, & \text{si } i \geq j, \\ 0 & \text{si } i < j, \end{cases} \quad i, j = 1, \dots, 200.$$

Los valores aproximados de la función  $\hat{f}$  en la malla uniforme de 200 puntos  $z_i$  del intervalo  $[0, 1]$  están dados por la solución de la Ecuación  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ . La matriz  $A$  es triangular inferior y todos sus elementos en la diagonal principal son distintos de cero, por lo que es invertible. Sin embargo, el ruido en el lado derecho  $\mathbf{b}$  ocasiona que la solución calculada sea una mala aproximación de los valores de  $f$  en los puntos  $m_i$  como se muestra en la Figura 1.15.

## 1.2. Problemas Mal Planteados

En relación a los problemas inversos, hemos visto que pueden tener solución única. Sin embargo, su solución puede cambiar abruptamente con pequeñas perturbaciones en los datos. A éstos se les conoce como problemas mal planteados.

Un problema está *bien planteado* en el sentido de Hadamard si cumple lo siguiente:

- \* tiene solución,
- \* la solución es única,
- \* la solución depende continuamente de los datos.

De otro modo, decimos que el problema está *mal planteado*.

### Observaciones 1.5:

 Remarcamos que si la solución no depende continuamente de los datos, los pequeños errores en los datos ocasionan que la solución presente cambios sustanciales.

Trabajamos con funciones  $f, g : (a, b) \rightarrow \mathbb{R}$  que están en espacios vectoriales  $\mathcal{V}$  y  $\mathcal{W}$ , respectivamente. El modelo que usamos para transformar  $f$  en la función  $g$  puede formularse matemáticamente como:

$$\int_a^b k(x, y)f(y)dy = g(x), \quad a < x < b, \quad (1.5)$$

donde  $k : (a, b) \times (a, b) \rightarrow \mathbb{R}$  es una función conocida.

Los problemas inversos que abordamos consisten en resolver la Ecuación (1.5), llamada *Ecuación Integral de Fredholm de Primera Clase*. El proceso o sistema físico es realizado por el operador lineal  $T : \mathcal{V} \rightarrow \mathcal{W}$  dado por

$$T[f](x) = \int_a^b k(x, y)f(y)dy.$$

La función  $k$  se llama *núcleo* del operador  $T$ .

### Observaciones 1.6:

☞ Con la ayuda de la integración, el núcleo  $k$  puede suavizar la función  $f$ . Por lo que al aplicar el operador integral  $T$ , reducimos las altas frecuencias de la función  $f$ . En cambio, al resolver la Ecuación (1.5), se amplifican las bajas frecuencias de la función  $g$ .

Para problemas inversos bien planteados en el sentido de Hadamard, podemos formular las tres condiciones en términos de la transformación  $T$ :

- \*  $T$  es sobreyectiva,
- \*  $T$  es inyectiva,
- \*  $T^{-1} : \mathcal{W} \rightarrow \mathcal{V}$  es una función continua.

En el siguiente ejemplo examinamos un operador integral muy reconocido en la Ciencia e Ingeniería:

La Transformada de Laplace. Una herramienta útil para resolver ecuaciones diferenciales ordinarias y parciales. La mayoría de los libros básicos sobre ecuaciones diferenciales tienen un capítulo dedicado a este operador.

Las transformadas integrales se remontan a los trabajos de Leonard Euler entre 1763 y 1769. De hecho, Laplace da crédito a Euler en su trabajo *Théorie analytique des probabilités* por haberlas introducido. Para finales del siglo XIX, Poincaré y Pincherle extendieron la transformada de Laplace a su forma compleja, Picard la extendió a dos variables, y Abel investigó más sobre el tema. El nombre fue propuesto por Bernstein en 1920. La primera aplicación de la versión moderna de esta transformada integral se debe a Bateman en 1910. Él la usó para transformar la ecuación de decaimiento radioactivo. Por otra parte, la aportación del cálculo operacional de Oliver Heaviside, ha permitido aplicarla en Ingeniería Eléctrica [107].

En el Ejemplo 1.3, usamos esta transformada para resolver el problema inverso de la forma del vertedero, su solución sensible al ruido en las observaciones, nos sugiere que la transformada inversa de Laplace es sensible a pequeñas perturbaciones.

**Ejemplo 1.4 (Transformada inversa de Laplace [8]).** Denotemos por  $Ce(0, \infty)$  al conjunto de funciones continuas por tramos de valores reales sobre  $(0, \infty)$  que tienen orden exponencial. El problema es

*Dada  $f \in Ce(0, \infty)$ , encontrar su transformada de Laplace  $g$ .*

El operador  $\mathcal{L} : Ce(0, \infty) \rightarrow Ce(0, \infty)$  dado por

$$\mathcal{L}[f](s) = \int_0^{\infty} e^{-sy} f(y) dy.$$

nos da la transformada de Laplace  $g = \mathcal{L}[f]$ . En particular,

$$f(t) = \frac{1}{2\sqrt{\pi t^3}} \exp(-1/4t) \implies \mathcal{L}[f](s) = \exp(-\sqrt{s}).$$

❖ **Problema inverso.** Queremos recuperar una función a partir de su transformada de Laplace:

*Dada  $g \in Ce(0, \infty)$ , hallar  $f \in Ce(0, \infty)$  tal que  $\mathcal{L}(f) = g$ .*

Si encontramos la inversa  $\mathcal{L}^{-1} : Ce(0, \infty) \rightarrow Ce(0, \infty)$  del operador  $\mathcal{L}$ , entonces la solución es  $f = \mathcal{L}^{-1}[g]$ . Damos un caso concreto: La transformada inversa de Laplace de

$$g(s) = \frac{0.2}{s^2 + 1}$$

es

$$f(t) = 0.2 \sin(t).$$

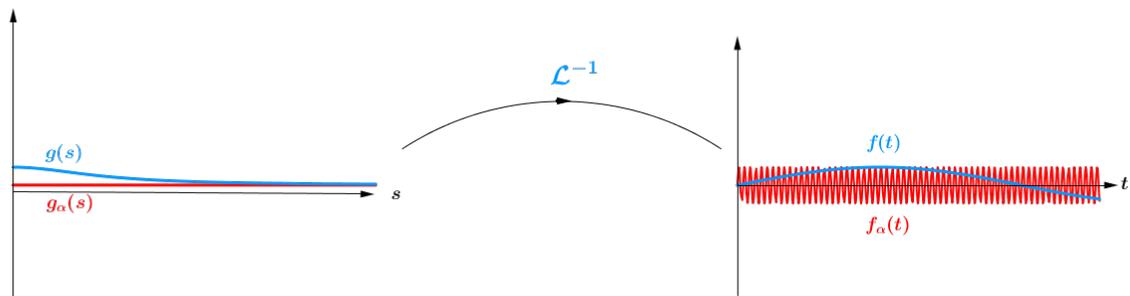
❖ **Sensibilidad de su solución.** Veamos como cambia  $\mathcal{L}^{-1}[g]$  con pequeñas perturbaciones en  $g$ . Para ello, introducimos  $\alpha = 100$  en la función  $g$  como

$$g_\alpha(s) = \frac{0.2\alpha}{s^2 + \alpha}.$$

Conforme  $s$  aumenta, la distancia  $|g(s) - g_\alpha(s)|$  disminuye. De hecho,  $g$  y  $g_\alpha$  tienen la misma asíntota horizontal. En cambio, la transformada inversa  $f_\alpha = \mathcal{L}^{-1}[g_\alpha]$  dada por

$$f_\alpha(t) = 0.2 \sin(\alpha t).$$

oscila más rápido que  $f$ . Véase Figura 1.16. Esto nos sugiere que  $\mathcal{L}^{-1}$  es sensible a las pequeñas perturbaciones que recibe la función  $g$ .



**Figura 1.16:** Sensibilidad de la transformada inversa de Laplace

Con la teoría de variable compleja, podemos obtener una expresión analítica para la inversa [107]:

$$\mathcal{L}^{-1}[g](t) = \lim_{\omega \rightarrow \infty} \frac{1}{2\pi i} \int_{\sigma - i\omega}^{\sigma + i\omega} g(s)e^{st} ds,$$

donde  $\sigma \in \mathbb{R}$  es mayor que las singularidades de  $g$ . La presencia de la exponencial en el integrando nos indica que  $\mathcal{L}^{-1}$  amplifica los errores en la función  $g$ .

Otra expresión para  $\mathcal{L}^{-1}$ , relaciona la inversa con la diferenciación [122]:

$$\mathcal{L}^{-1}[g](t) = \lim_{k \rightarrow \infty} \left[ \frac{(-1)^k}{k!} \cdot \left(\frac{k}{y}\right)^{k+1} \cdot \frac{d^k g}{ds^k} \Big|_{\frac{k}{t}} \right].$$

Vimos que la diferenciación es sensible a pequeñas perturbaciones en la función. Como aparecen derivadas de orden superior, los pequeños cambios en la función  $g$  producen cambios drásticos en  $\mathcal{L}^{-1}[g]$ .

Los ejemplos anteriores nos dan una idea de la dificultad en resolver la ecuación integral de Fredholm de primera clase. La solución de los problemas inversos que están descritos por esta ecuación son sensibles a los pequeños errores del lado derecho  $g$ . Esto nos sugiere que resolver la ecuación (1.5) para  $f$  es un problema mal planteado en el sentido de Hadamard.

### 1.3. Ecuaciones Integrales

Las ecuaciones integrales aparecen en varios lados de la Ciencia e Ingeniería. Desde problemas como la forma del vertedero en agricultura hasta en técnicas experimentales para la construcción de una maquina de rayos X en el Laboratorio Científico de los Alamos como relata Wing en [123]. Métodos de transformadas integrales como la de Laplace y Fourier son ampliamente conocidos en Ingeniería. Sin embargo, químicos, ingenieros, geólogos, entre otros, requieren de un nivel matemático más avanzado para analizar las ecuaciones integrales que aparecen en sus experimentos. En particular, las propiedades de las ecuaciones integrales de primera clase son inusuales y complicadas. Comprender estas propiedades es apropiado para buscar soluciones analíticas y numéricas de la ecuación integral.

Vemos algunos resultados de análisis funcional sobre ecuaciones integrales con el fin de saber cuando un problema determinado por una ecuación integral de Fredholm de primera clase está mal planteado.

Sea  $T : \mathcal{V} \rightarrow \mathcal{W}$  el operador dado por

$$T[f](x) = \int_a^b k(x, y)f(y)dy.$$

Queremos saber bajo qué propiedades del operador  $T$ , los problemas dados por la ecuación integral (1.5) están mal planteados en el sentido de Hadamard. Supóngamos que los

espacios  $\mathcal{V}$  y  $\mathcal{W}$  tienen la estructura de espacios vectoriales. Para concretar ideas, hacemos nuestro análisis en un espacio de funciones cuadrado integrables  $L^2(a, b)$  y lo equipamos con el producto interno

$$\langle u, v \rangle = \int_a^b u(x)v(x)dx$$

y la norma inducida

$$\|v\|_2 = \sqrt{\langle v, v \rangle}.$$

Definimos el operador  $T$  sobre el espacio normado  $\mathcal{V} = \mathcal{W} = L^2(a, b)$ .

### Observaciones 1.7:

 La transformación  $T$  es lineal.

 Si además de ser lineal, pedimos que  $T$  transforme conjuntos acotados de  $\mathcal{V}$  en conjuntos acotados de  $\mathcal{W}$ , entonces  $T$  es un operador continuo sobre  $L^2(a, b)$ .

De modo que  $T$  es continuo si podemos hallar una constante  $M > 0$  tal que

$$\|T[f]\|_2 \leq M\|f\|_2 \quad \forall f \in L^2(a, b). \quad (1.6)$$

La constante  $M$  no debe depender de la función  $f$  que demos. Los operadores que cumplen esta desigualdad se llaman **operadores acotados**. Una manera de obtener la desigualdad (1.6) es pedir que el núcleo  $k$  sea cuadrado integrable sobre  $(a, b) \times (a, b)$ . Así, la desigualdad de Cauchy-Schwarz

$$\left( \int_a^b k(x, y)f(y)dy \right)^2 \leq \left( \int_a^b |k(x, y)|^2 dy \right) \left( \int_a^b |f(y)|^2 dy \right)$$

se cumple e implica que

$$M = \int_a^b \int_a^b |k(x, y)|^2 dx dy.$$

Por lo que si el núcleo  $k$  es cuadrado integrable, entonces  $T$  es un operador acotado, y por ser lineal, se sigue que es una transformación continua sobre  $L^2(a, b)$ .

Para obtener otra propiedad del operador  $T$ , vamos a aproximarlo por una sucesión de operadores  $T_n$  sobre  $L^2(a, b)$ . Lo que hacemos es expandir el núcleo  $k$  en una base de funciones. Sea  $\{\varphi_n\}$  una sucesión de funciones ortonormales sobre  $L^2(a, b)$ . Podemos expandir cualquier función  $u \in L^2(a, b)$  como

$$u = \sum_{n=1}^{\infty} \langle u, \varphi_n \rangle \varphi_n$$

siempre y cuando

$$\|u\|_2^2 = \sum_{n=1}^{\infty} |\langle u, \varphi_n \rangle|^2.$$

Entonces la sucesión  $\{\varphi_n\}$  forma una base ortonormal de  $L^2(a, b)$ .

Así, cuando equipamos al espacio  $L^2((a, b) \times (a, b))$  con el producto interno

$$\langle u, v \rangle = \int_a^b \int_a^b u(x, y)v(x, y)dx dy,$$

vamos a tener que las funciones

$$w_{p,q}(x, y) = \varphi_p(x)\varphi_q(y)$$

forman una base ortonormal de ese espacio. Para expandir el núcleo  $k$  en la base de funciones  $w_{p,q}$ , pedimos que sea cuadrado integrable. De esa manera,

$$k(x, y) = \sum_{p,q=1}^{\infty} \langle k, w_{p,q} \rangle w_{p,q}(x, y).$$

Con esta expansión, construimos operadores sobre  $L^2(a, b)$ . Lo que hacemos es reemplazar el núcleo  $k$  del operador  $T$  por su expansión truncada

$$k_n(x, y) = \sum_{p,q=1}^n \langle k, w_{p,q} \rangle w_{p,q}(x, y).$$

De esa manera, conseguimos operadores  $T_n : L^2(a, b) \rightarrow L^2(a, b)$  dados por

$$T_n[f](y) = \int_a^b k_n(x, y)f(x)dx.$$

Estos operadores convergen a  $T$  en la norma

$$\|T\|_2 = \sup_{f \in L^2(a,b)} \frac{\|T[f]\|_2}{\|f\|_2}.$$

y transforman subespacios de  $L^2(a, b)$  en subespacios de dimensión finita generados por las funciones  $\varphi_1, \dots, \varphi_n$  [19]. Intuitivamente, podemos pensar que  $T$  está cerca de ser una transformación lineal en dimensión finita, ya que la imagen de cada  $T_n$  es de dimensión finita. En la literatura [72],  $T$  es un **operador compacto**, debido a que transforma conjuntos acotados en conjuntos con cerradura compacta. En espacios de dimensión infinita como  $L^2(a, b)$ , los operadores compactos nos dan información sobre su inversa.

**Teorema 1.1** ([21]). *Sean  $\mathcal{V}$  y  $\mathcal{W}$  espacios normados. Sea  $T : \mathcal{V} \rightarrow \mathcal{W}$  un operador continuo y compacto. Entonces  $T$  no tiene una inversa acotada  $T^{-1} : \mathcal{W} \rightarrow \mathcal{V}$  si  $\mathcal{V}$  es de dimensión infinita.*

**Observaciones 1.8:**

 Debido a que los operadores lineales y continuos entre espacios normados son acotados, el Teorema 1.1 implica que cualquier operador acotado y compacto sobre  $L^2(a, b)$  no tiene una inversa continua. Así que si nuestro problema inverso se reduce a resolver la ecuación integral (1.5) sobre  $L^2(a, b)$  con un núcleo  $k$  cuadrado integrable, entonces el problema ésta mal planteado en el sentido de Hadamard.



## PROBLEMAS DISCRETOS MAL PLANTEADOS

Los problemas inversos que vimos en el capítulo anterior dan lugar a la Ecuación integral de Fredholm de primera clase

$$\int_a^b k(x, y)f(y)dy = g(x), \quad a \leq x \leq b. \quad (2.1)$$

En ocasiones no podemos hallar una expresión analítica para la solución de esta ecuación o solamente disponemos de un número finito de observaciones. Discretizamos:

Dados  $\mathbf{b} \in \mathbb{R}^m$  y la matriz  $A \in \mathbb{R}^{m \times n}$ , hallar un vector  $\mathbf{x} \in \mathbb{R}^n$  tal que  $A\mathbf{x} = \mathbf{b}$ .

### Observaciones 2.1:

 Si  $A$  es invertible, entonces el problema discreto está bien planteado en el sentido de Hadamard.

Desde un punto de vista práctico, trabajamos con precisión finita y el lado derecho  $\mathbf{b}$  tiene errores. A pesar de que  $A$  sea invertible, ésta puede ser sensible a errores por redondeo. Lo que ocasiona cambios abruptos en la solución calculada. Así que un aspecto a considerar es la sensibilidad de la solución respecto a pequeñas perturbaciones en las observaciones. Para darnos cuenta de lo sensible que es esta solución, examinamos el sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b}$  con tres enfoques: estadístico, geométrico y numérico.

## 2.1. Métodos de Discretización

En esta sección presentamos dos métodos para discretizar la ecuación integral (2.1). Ambos requieren reglas de cuadratura para aproximar la integral definida.

### 2.1.1. Reglas de Cuadratura

Consideramos la función  $f : [a, b] \rightarrow \mathbb{R}$ . Aproximamos  $f$  por otra función tal que su integral sea fácil de calcular. En particular, si aproximamos  $f$  por polinomios de Lagrange, entonces su integral se aproxima por una suma ponderada de sus evaluaciones:

$$\int_a^b f(x)dx \approx \sum_{i=1}^p \omega_i f(x_i).$$

La suma se conoce como *regla de cuadratura*, los puntos de evaluación  $x_1, \dots, x_p$  se llaman *nodos* y los coeficientes  $\omega_1, \dots, \omega_p$  son los *pesos*. Debemos hacer una elección adecuada de nodos y pesos para que el error de aproximación sea pequeño. Algunas reglas de cuadratura más comunes son

\* *Cuadratura compuesta del punto medio.* En este caso

$$x_i = a + \left(\frac{2i-1}{2}\right) \left(\frac{b-a}{p-1}\right), \quad \omega_i = \frac{b-a}{p-1}.$$

\* *Cuadratura compuesta del trapecio.* Escogemos

$$x_i = a + (i-1) \left(\frac{b-a}{p-1}\right), \quad \omega_i = \begin{cases} \frac{b-a}{2(p-1)}, & \text{si } i \in \{1, p\} \\ \frac{b-a}{p-1}, & \text{si } i \in \{2, \dots, p-1\} \end{cases}$$

\* *Cuadratura Gaussiana.* Sean  $L_0, \dots, L_{p-1}$  polinomios definidos sobre un intervalo  $(c, d)$  que cumplen la relación

$$\int_c^d L_i(x)L_j(x)\omega(x)dx = \begin{cases} 1, & \text{si } i = j, \\ 0, & \text{en otro caso,} \end{cases} \quad (2.2)$$

donde  $\omega : (c, d) \rightarrow \mathbb{R}$  es una función de valores positivos. Los nodos  $x_i$  que elegimos son los  $p$  ceros de  $L_{p-1}$ . Interpolamos los puntos  $(x_1, f(x_1)), \dots, (x_p, f(x_p))$  con los polinomios de Lagrange  $l_1, \dots, l_p$ . Tomamos

$$\omega_i = \int_c^d l_i(x)\omega(x)dx.$$

### 2.1.2. Método de Galerkin

Un método para obtener una solución aproximada de la Ecuación (2.1) es reemplazar la función  $f$  por su proyección  $P_f$  sobre el subespacio generado por una colección de funciones linealmente independientes  $\varphi_1, \dots, \varphi_n \in L^2[a, b]$ :

$$P_f(t) = \sum_{j=1}^n x_j \varphi_j(t). \quad (2.3)$$

Después de elegir las  $\varphi_j$ 's, nuestra tarea es encontrar los valores de los coeficientes  $x_1, \dots, x_n$ . Escogemos una segunda colección de funciones linealmente independientes  $\psi_1, \dots, \psi_n$  en  $L^2[a, b]$  que sean ortogonales al residuo  $T[P_f] - g$ :

$$\int_a^b \left[ \int_a^b k(s, y) P_f(t) dt - g(s) \right] \psi_i(s) ds = 0, \quad i = 1, \dots, n.$$

Así tenemos

$$\sum_{j=1}^n x_j \underbrace{\int_a^b \int_a^b k(s, t) \psi_i(s) \varphi_j(t) ds dt}_{\mathbf{a}_{ij}} = \underbrace{\int_a^b g(s) \psi_i(s) ds}_{\mathbf{b}_i}, \quad i = 1, \dots, n.$$

Esto da lugar al sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b}$ . Una vez resuelto, formamos  $P_f$  mediante la combinación lineal (2.3) [19].

**Ejemplo 2.1 (Ecuación de Phillips [97],[51]).** Sean  $k, g : [-6, 6] \rightarrow \mathbb{R}$  las funciones dadas por

$$k(s) = \begin{cases} 1 + \cos\left(\frac{s\pi}{3}\right), & \text{si } |s| < 3, \\ 0, & \text{si } |s| \geq 3, \end{cases}$$

y

$$g(s) = (6 - |s|) \left(1 + \frac{1}{2} \cos\left(\frac{s\pi}{3}\right)\right) + \frac{9}{2\pi} \sin\left(\frac{|s|\pi}{3}\right),$$

queremos una aproximación de la función  $f : [-6, 6] \rightarrow \mathbb{R}$  que cumple con la ecuación integral de Phillips

$$\int_{-6}^6 k(s-t)f(t)dt = g(s). \quad (2.4)$$

Hacemos la partición uniforme del intervalo  $[-6, 6]$  en  $n + 1$  puntos con  $n$  múltiplo de 4:

$$s_j = -6 + (j-1) \left(\frac{12}{n}\right), \quad j = 1, \dots, n+1.$$

Aproximamos  $f$  mediante una combinación lineal  $P_f$  de  $n$  funciones sombrero:

$$\varphi_j(s) = \begin{cases} \sqrt{n/12}, & \text{si } s_j \leq s < s_{j+1}, \\ 0, & \text{en otro caso,} \end{cases} \quad j = 1, \dots, n$$

$$\varphi_n(s_{n+1}) = \sqrt{n/12}$$

Para obtener los coeficientes  $x_1, \dots, x_n$  de  $P_f$  resolvemos el sistema  $A\mathbf{x} = \mathbf{b}$ , donde

$$b_i = \sqrt{\frac{n}{12}} \int_{s_i}^{s_{i+1}} g(s)ds, \quad i, j = 1, \dots, n.$$

$$a_{i,j} = \frac{n}{12} \int_{s_j}^{s_{j+1}} \int_{s_i}^{s_{i+1}} k(s-t)dsdt,$$

Usamos la regla del trapecio para aproximar las integrales sobre los subintervalos. Así,

$$b_i \approx \frac{1}{2} \sqrt{\frac{12}{n}} (g(s_i) + g(s_{i+1})),$$

$$a_{i,j} \approx \frac{3}{n} (k(s_i - s_j) + k(s_{i+1} - s_j) + k(s_i - s_{j+1}) + k(s_{i+1} - s_{j+1})).$$

Puesto que  $k$  tiene soporte finito en  $(-3, 3)$ ,  $A$  es una matriz con ancho de banda  $n/4$ , más aún, como  $k$  es una función par, tenemos que  $A$  es simétrica. Por lo que podemos calcular la factorización  $LDL^T$ :

$$P^T A P = LDL^T,$$

donde  $L$  es una matriz triangular inferior con unos en diagonal principal,  $P$ , una matriz de permutación y  $D$ , diagonal por bloques, cada uno es escalar real o un matriz real  $2 \times 2$ . Las tres son del mismo tamaño que  $A$ . Con esta factorización, resolvemos el sistema de ecuaciones  $A\mathbf{x} = \mathbf{b}$ :

1. resolver  $L\mathbf{y} = P^T \mathbf{b}$
2. resolver  $D\mathbf{q} = \mathbf{y}$
3. resolver  $L^T \mathbf{z} = \mathbf{y}$
4.  $\mathbf{x} = P\mathbf{z}$ .

Una vez que obtenemos el vector de los coeficientes, evaluamos la expansión (2.3) en los puntos medios de los subintervalos  $[s_i, s_{i+1})$  y en los extremos  $s_1 = -6$  y  $s_{n+1} = 6$ . Comparamos esta aproximación con la solución analítica  $f = k$  de la Ecuación de Phillips. Realizamos los cálculos en una PC de 64 bits.

En las Figuras 2.1(a)-(b), mostramos las gráficas de  $P_f$  para 12 y 24 funciones sombrero, respectivamente. Observamos que  $P_f$  tiene pequeñas oscilaciones alrededor de  $t = -3$  y  $t = 3$ . Cuando pasamos de usar 12 a 24 funciones  $\varphi_j$ , las oscilaciones son de menor amplitud.

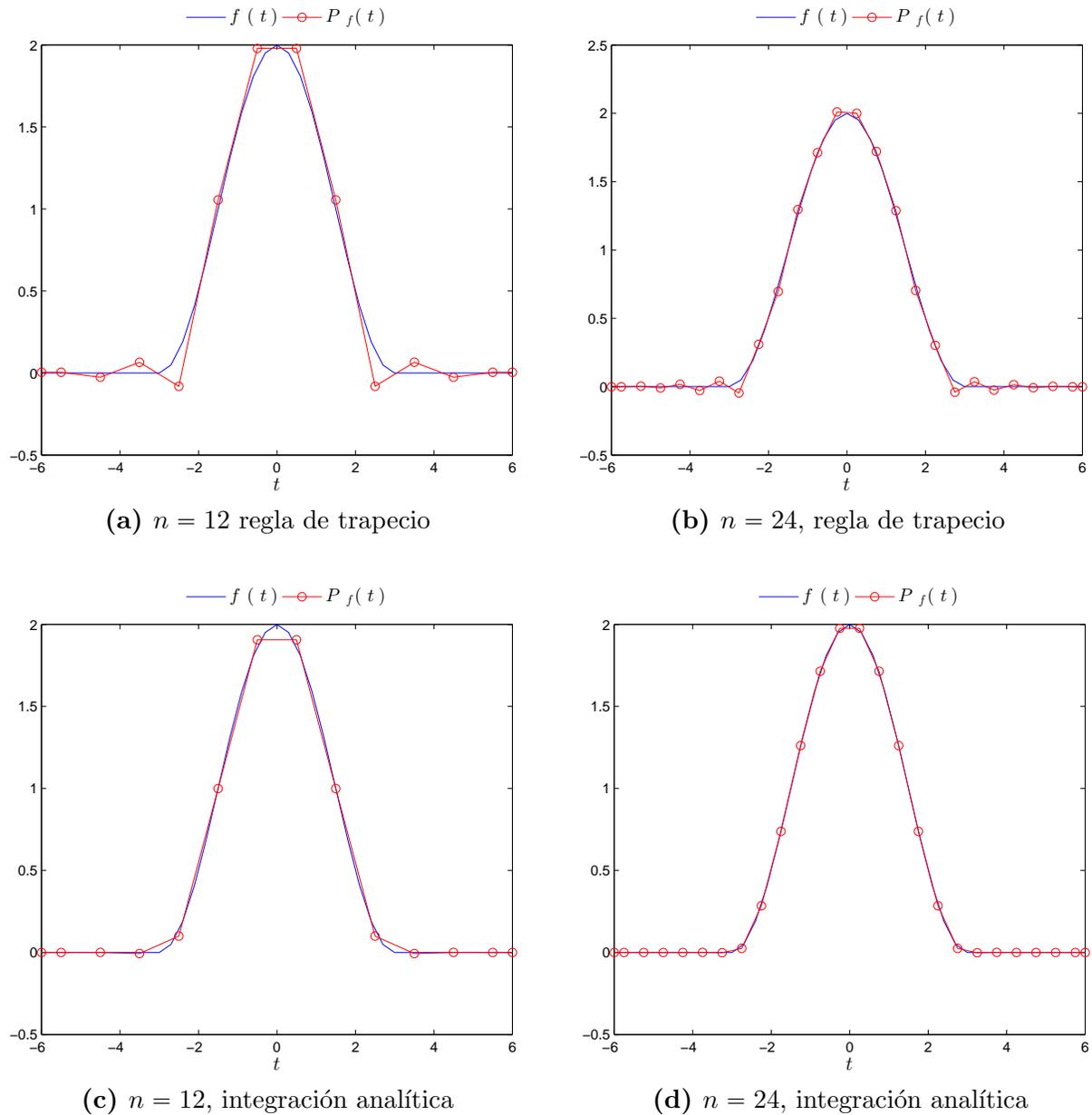
En las Figuras 2.1(c)-(d), mostramos las gráficas de  $P_f$  que calculamos con la rutina `phillips` del paquete `REGUTOOLS` de Hansen [50]. Estas rutinas usan el método de Galerkin con las mismas funciones sombrero. La diferencia es que las integrales se calculan analíticamente en vez de aproximarlas por regla de cuadratura. Por eso, vemos que la gráfica de  $P_f$  en Figuras 2.1(c)-(d) no presentan las pequeñas oscilaciones observadas en Figuras 2.1(a)-(b) y se traslapa con la solución analítica.

núm. sombreros	cuadratura	cond(A)
12	analítica	$4.50553 \times 10^2$
12	trapecio	$2.65898 \times 10^4$
24	analítica	$8.17908 \times 10^3$
24	trapecio	$4.2447 \times 10^5$

**Tabla 2.1:** Condicionamiento de matriz  $A$  obtenida en discretización por Galerkin de la ecuación de Phillips.

Cuando calculamos analíticamente las integrales, la matriz de coeficientes tiene la misma estructura: simétrica con ancho de banda  $n/4$ . Comparamos su número de condición con el

de la matriz obtenida por regla de trapecio en la Tabla 2.1. Observamos que en los cuatro casos, el problema está bien condicionado.



**Figura 2.1:** La solución  $f$  de la Ecuación de Phillips y su aproximación  $P_f$  por  $n$  funciones sombrero  $\varphi_j$ . En cada caso indicamos cómo se calculan las integrales.

En el siguiente ejemplo vemos una ecuación integral que aparece en la investigación de materiales expuestos a radiación.

Consideramos una barra colocada en un intervalo  $[a, b]$  que almacena energía cuando absorbe radiación. Supongamos que la fuente de radiación se debe a neutrones distribuidos sobre la barra que se pueden desplazar solamente a lo largo de ésta. Queremos conocer la fuerza de la fuente radioactiva a partir de las medidas de la energía depositada en un posición fija.

Las medidas están dadas por una función  $g : [a, b] \rightarrow \mathbb{R}$ . La fuerza que buscamos es descrita por una función  $f : [a, b] \rightarrow \mathbb{R}$ . Un modelo sencillo [123] para la absorción de radiación está descrito por la ecuación integral

$$\int_a^b e^{-\sigma|x-y|} f(y) dy = g(x), \quad a \leq x \leq b.$$

Discretizamos esta ecuación integral para determinar los valores de  $f$  a partir de las observaciones de  $g$ .

**Ejemplo 2.2 (Ecuación de Absorción).** A partir de la senoidal

$$g(x) = \cos(x), \quad 0 \leq x \leq 10$$

queremos dar una aproximación de la función  $f : [0, 10] \rightarrow \mathbb{R}$  que cumple la ecuación

$$\int_0^{10} e^{-0.01|x-y|} f(y) dy = g(x), \quad 0 \leq x \leq 10. \quad (2.5)$$

Hacemos la partición uniforme del intervalo  $[0, 10]$  en  $n + 1$  puntos:

$$y_j = \frac{10}{n}(j - 1), \quad j = 1, \dots, n + 1$$

Aproximamos  $f$  mediante una combinación lineal  $P_f$  de  $n$  funciones sombrero

$$\varphi_j(y) = \begin{cases} \sqrt{\frac{n}{10}}, & \text{si } y_j \leq y < y_{j+1}, \\ 0, & \text{en otro caso,} \end{cases} \quad j = 1, \dots, n.$$

Para obtener los coeficientes  $x_1, \dots, x_n$  de  $P_f$ , resolvemos el sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b}$ , donde

$$b_i = \sqrt{\frac{n}{10}} \int_{y_i}^{y_{i+1}} \cos(x) dx, \quad i = 1, \dots, n,$$

$$a_{i,j} = \frac{n}{10} \int_{y_i}^{y_{i+1}} \int_{y_j}^{y_{j+1}} e^{-0.01|x-y|} dy dx, \quad i, j = 1, \dots, n.$$

Aproximamos las integrales que nos dan los elementos de  $A$  con la regla del trapecio:

$$a_{i,j} \approx \frac{5}{2n} \left( e^{-0.01|y_i - y_j|} + e^{-0.01|y_{i+1} - y_j|} + e^{-0.01|y_i - y_{j+1}|} + e^{-0.01|y_{i+1} - y_{j+1}|} \right).$$

Como  $A$  es simétrica, sus valores propios  $\lambda_1, \dots, \lambda_n$  son reales y existe una matriz real  $U$  de tamaño  $n \times n$  tal que  $U^T U = I$  tal que

$$A = U \text{diag}(\lambda_1, \dots, \lambda_n) U^T,$$

Esta factorización de  $A$  se conoce como *Descomposición en Valores Propios (EVD)* [37]. Para  $A$  de rango completo, podemos conseguir el vector de coeficientes mediante la EVD como

$$\mathbf{x} = U \text{diag} \left( \frac{1}{\lambda_1}, \dots, \frac{1}{\lambda_n} \right) U^T \mathbf{b}.$$

Luego, formamos la combinación lineal  $P_f$ . En la Figura 2.2 comparamos  $P_f$  con la solución

$$f(y) = 50.005 \cos(y)$$

de la Ecuación Integral (2.5).

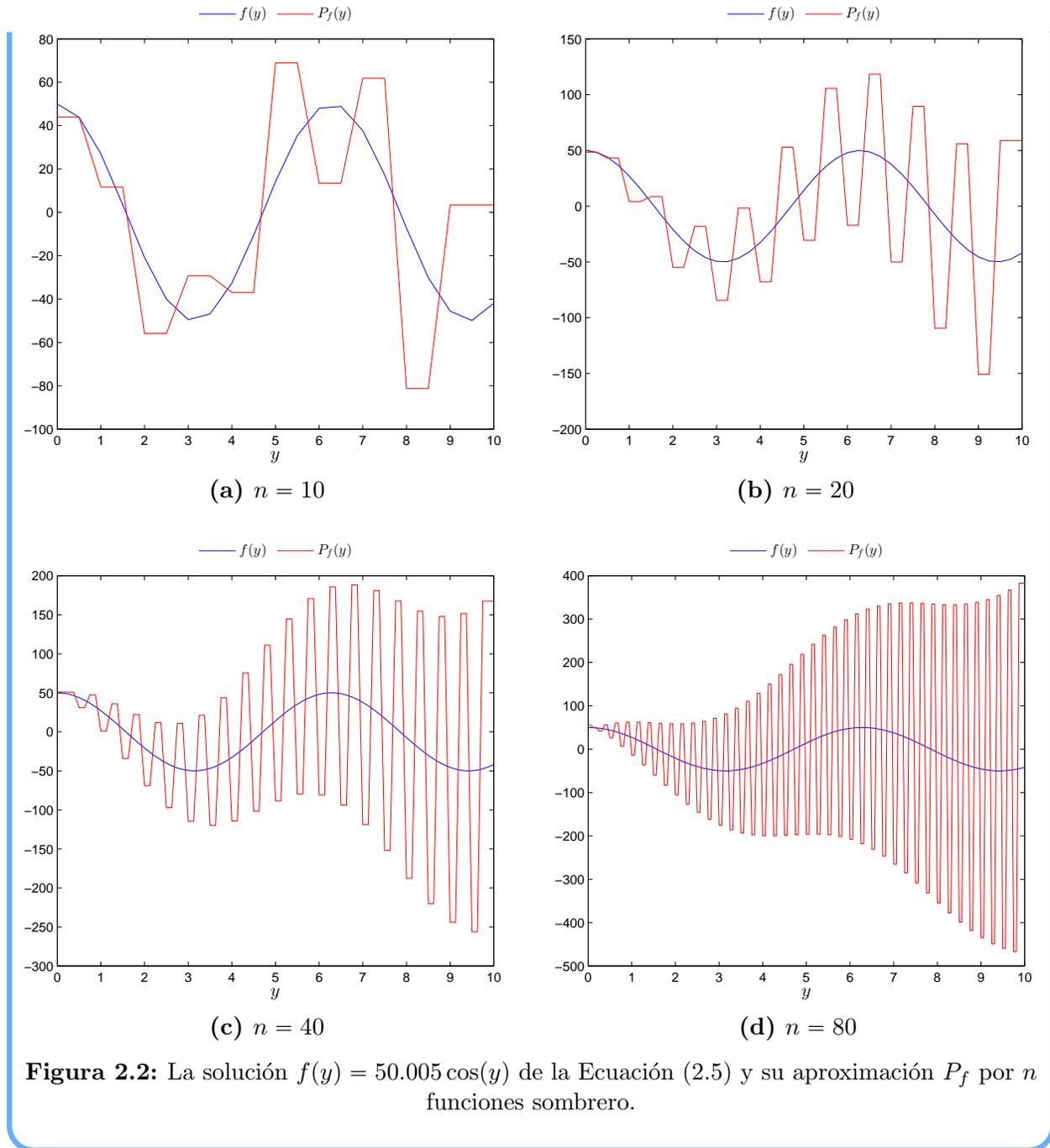
En la Figura 2.2(a) usamos 10 funciones sombrero  $\varphi_j$ . En ese caso,  $\|f - P_f\|_\infty \approx 73.912$ . En la Figura 2.2(b) empleamos 20 funciones  $\varphi_j$ . Obtuvimos que  $\|f - P_f\|_\infty \approx 108.628$ . En cambio con 40 y 80 funciones sombrero, la amplitud de la aproximación  $P_f$  se hace más grande al alejarse del origen. Como podemos ver en las Figuras 2.2(c) y 2.2(d).

En la Tabla 2.2 mostramos la aproximación de  $\|f - P_f\|_\infty$  así como el intervalo  $[\lambda_{\min}, \lambda_{\max}]$  donde están los valores absolutos de los valores propios de  $A$  conforme  $n$  se hace más grande.

$n$	$\lambda_{\min}$	$\lambda_{\max}$	$\ f - P_f\ _\infty$
10	$1.253 \times 10^{-4}$	9.6738	73.912
20	$7.741 \times 10^{-6}$	9.6750	108.628
40	$4.823 \times 10^{-7}$	9.6753	214.891
80	$3.013 \times 10^{-8}$	9.6753	427.757

**Tabla 2.2:** Valores propios más grandes y pequeños de la matriz  $A$  obtenida por discretización de ecuación 2.5 con Galerkin usando  $n$  funciones sombrero, junto con los errores de la aproximación  $P_f$  de la solución  $f$ .

Cuando aumentamos el número  $n$  de funciones sombrero de 10 a 80, el valor propio  $\lambda_{\min}$  de la matriz simétrica  $A$  pasa del orden de  $10^{-4}$  a  $10^{-8}$ , mientras que  $\lambda_{\min}$  se mantiene cerca de 9.675. Por eso los cocientes  $\lambda_{\max}/\lambda_{\min}$  aumentan.



### Observaciones 2.2:

☞ En el Ejemplo 2.2, el error por redondeo que se genera en el cálculo de los valores propios de  $A$  ocasiona que los coeficientes de la expansión (2.3) no se calculen de manera precisa. Por consiguiente, la aproximación  $P_f$  se aleja de la solución.

### 2.1.3. Método de Colocación

Dadas las observaciones  $g_1, \dots, g_n$  de la función  $g$  en los puntos  $s_1, \dots, s_n$  del intervalo  $[a, b]$ , respectivamente. deseamos calcular valores aproximados de la función  $f$  que cumple la Ecuación (2.1). Usamos el **método de colocación**.

Supongamos que los valores de  $g$  coinciden con los de la integral del producto de  $f$  con el núcleo  $k$  en cada  $s_i$ :

$$\int_a^b k(s_i, t) f(t) dy = g_i$$

Aproximamos cada integral por un método de cuadratura:

$$\int_a^b k(s_i, t) f(t) dt \approx \sum_{j=1}^p \omega_j k(s_i, t_j) f(t_j),$$

En consecuencia,

$$\sum_{j=1}^p \underbrace{\omega_j k(s_i, t_j)}_{a_{ij}} \underbrace{f(t_j)}_{x_j} = \underbrace{g_i}_{b_i}, \quad i = 1, \dots, n.$$

De aquí, obtenemos un sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b}$ .

La solución de las Ecuación  $A\mathbf{x} = \mathbf{b}$  nos da las evaluaciones de la función  $f$  en los nodos  $t_j$  de la regla de cuadratura. Llamamos **puntos de colocación** a los  $s'_i$ s.

**Ejemplo 2.3.** Retomemos la ecuación de Phillips (2.4). Esta vez aproximamos los valores de la función  $f : [-6, 6] \rightarrow \mathbb{R}$  como se describe en Phillips [97]. Tomamos a los puntos  $s_1, \dots, s_{n+1}$  de partición uniforme de  $[-6, 6]$  como puntos de colocación. Evaluamos en ambos lados de la ecuación (2.4) para obtener

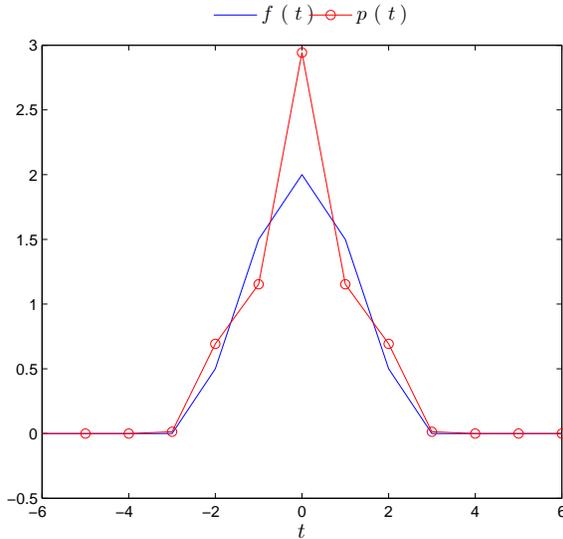
$$g_i = \int_{-6}^6 k(s_i - t) f(t) dt, \quad i = 1, \dots, n + 1.$$

Primero, aproximamos cada integral por cuadratura compuesta de Simpson con  $n + 1$  nodos que tomamos como puntos de colocación. Sea  $k_p = k(12p/n)$  para  $p = 0, \dots, n$ . Puesto que el núcleo  $k$  es función par, el método de colocación con cuadratura de Simpson nos da el sistema de ecuaciones lineales

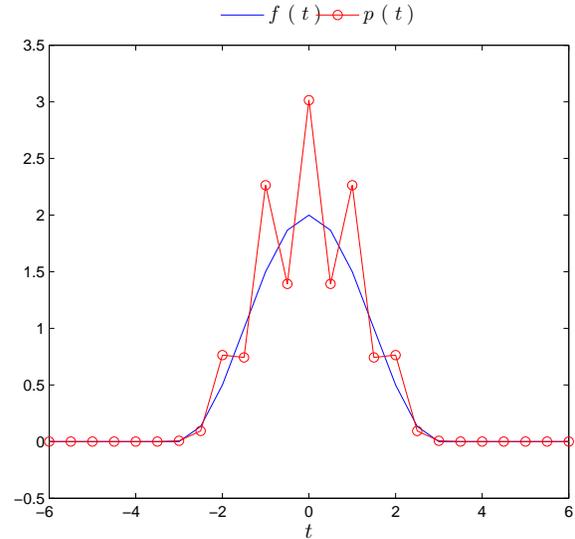
$$\frac{4}{n} \begin{bmatrix} k_0 & 4k_1 & 2k_2 & \cdots & 4k_{n-3} & 2k_{n-2} & 4k_{n-1} & k_n \\ k_1 & 4k_0 & 2k_1 & \cdots & 4k_{n-4} & 2k_{n-3} & 4k_{n-2} & k_{n-1} \\ k_2 & 4k_1 & 2k_0 & \cdots & 4k_{n-5} & 2k_{n-4} & 4k_{n-3} & k_{n-2} \\ k_3 & 4k_2 & 2k_1 & \cdots & 4k_{n-6} & 2k_{n-5} & 4k_{n-4} & k_{n-3} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ k_{n-2} & 4k_{n-3} & 2k_{n-4} & \cdots & 4k_1 & 2k_0 & 4k_1 & k_2 \\ k_{n-1} & 4k_{n-2} & 2k_{n-3} & \cdots & 4k_2 & 2k_1 & 4k_0 & k_1 \\ k_n & 4k_{n-1} & 2k_{n-2} & \cdots & 4k_3 & 2k_2 & 4k_1 & k_0 \end{bmatrix} \begin{bmatrix} f(s_1) \\ \vdots \\ f(s_{n+1}) \end{bmatrix} = \begin{bmatrix} g_1 \\ \vdots \\ g_{n+1} \end{bmatrix}.$$

$A_S \quad \mathbf{x} = \mathbf{b}$

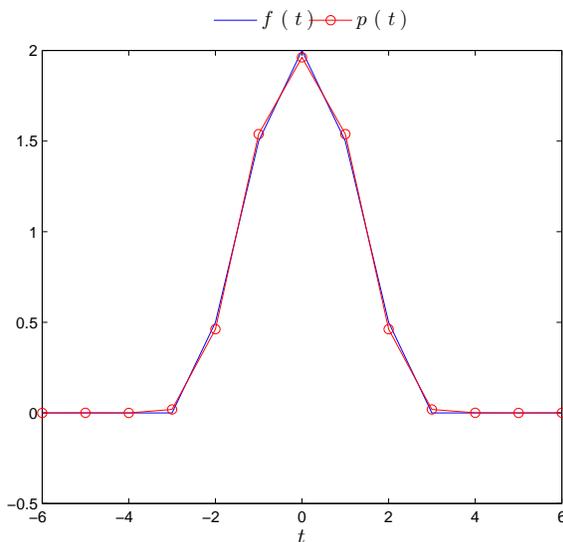
La matriz  $A_S$  es de orden  $n + 1$  y tiene ancho de banda  $n/4$  porque  $k$  tiene soporte en el intervalo  $(-3, 3)$ . Resolvemos el sistema de ecuaciones  $A_S \mathbf{x} = \mathbf{b}$  por Factorización  $LU$  con pivoteo.



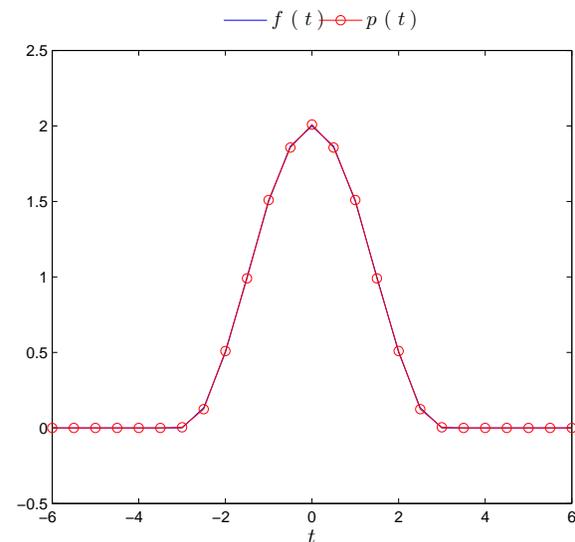
(a)  $n = 13$ , Simpson



(b)  $n = 25$ , Simpson



(c)  $n = 13$ , trapecio



(d)  $n = 25$ , Trapecio

**Figura 2.3:** Solución analítica  $f = k$  de Ecuación (2.4) y su aproximación por la poligonal  $p$  que une los puntos  $(s_1, x_1), \dots, (s_{n+1}, x_{n+1})$ , donde  $x_i$  es el valor aproximado de  $f$  en  $s_i$ , obtenidos por método de colocación. Indicamos la regla de cuadratura usada en cada caso.

Sea  $p$  la poligonal que une los puntos  $(s_i, x_i), i = 1, \dots, n + 1$ . En la Figura 2.3(a)-(b)

comparamos los valores de la solución  $f = k$  con los de la poligonal  $p$ . Para 13 puntos de colocación,  $p$  tiene soporte en  $(-3, 3)$ , pero rebasa el máximo de  $f$  en  $t = 0$ . Con 25 puntos de colocación,  $p$  tiene más oscilaciones. En ambos casos, la poligonal es simétrica. Las gráficas de  $p$  coinciden con las que obtiene Phillips en [97].

Ahora, aproximamos cada integral por la cuadratura compuesta del trapecio, donde los nodos son los  $n + 1$  puntos de colocación. Dado que  $k$  es función par, el método de colocación con cuadratura de trapecio nos da el sistema de ecuaciones lineales

$$\frac{6}{n} \begin{bmatrix} k_0 & 2k_1 & 2k_2 & \cdots & 2k_{n-2} & 2k_{n-1} & k_n \\ k_1 & 2k_0 & 2k_1 & \ddots & & 2k_{n-2} & k_{n-1} \\ k_2 & 2k_1 & 2k_0 & \ddots & \ddots & \vdots & k_{n-2} \\ \vdots & 2k_2 & \ddots & \ddots & \ddots & 2k_2 & \vdots \\ k_{n-2} & \vdots & \ddots & \ddots & 2k_0 & 2k_1 & k_2 \\ k_{n-1} & 2k_{n-2} & & \ddots & 2k_1 & 2k_0 & k_1 \\ k_n & 2k_{n-1} & 2k_{n-2} & \cdots & 2k_2 & 2k_1 & k_0 \end{bmatrix} \begin{bmatrix} f(s_1) \\ \vdots \\ f(s_{n+1}) \end{bmatrix} = \begin{bmatrix} g_1 \\ \vdots \\ g_{n+1} \end{bmatrix}.$$

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

La matriz  $A_T$  es de orden  $n + 1$  y tiene ancho de banda  $n/4$ . Para resolver la ecuación  $A_T \mathbf{x} = \mathbf{b}$  nuevamente usamos Factorización  $LU$  con pivoteo. En la Figura 2.3(c)-(d) comparamos los valores de la solución  $f$  con los de la nueva poligonal  $p$ . Observamos que las gráficas de  $f$  y  $p$  se traslapan para 13 y 25 puntos de colocación a diferencia de los casos correspondientes donde usamos cuadratura de Simpson.

### Observaciones 2.3:

👉 En el Ejemplo 2.3 obtenemos mejores resultados con la cuadratura compuesta del trapecio. Las propiedades del núcleo

$$k(s) = \begin{cases} 1 + \cos\left(\frac{s\pi}{3}\right), & \text{si } |s| < 3, \\ 0, & \text{si } |s| \geq 3, \end{cases}$$

influyen en las reglas de cuadratura que usamos. Lo que sabemos es que  $k$  es una función periódica par en  $(-3, 3)$ . En [115], Trefethen y Weideman nos comentan sobre la convergencia geométrica y exponencial de la regla del trapecio.

## 2.2. Problemas de Cuadrados Mínimos

Uno de los problemas que se presentan es hallar el valor de parámetros como velocidad, densidad o voltaje que determinan el modelo propuesto en un proceso físico. Una manera de relacionarlos con las observaciones es mediante una *regresión*. La idea es suponer que podemos inferir las causas de nuestro problema analizando las relaciones que existen entre las variables que estudiamos.

Dadas  $m$  observaciones  $b_1, \dots, b_m$  del problema, supóngamos que cada  $b_i$  depende de variables  $a_{i,1}, \dots, a_{i,n}$  y de parámetros  $x_1, \dots, x_n$ . Las variables  $a_{i,j}$  se conocen como **variables explicativas o independientes**. A partir de los valores de éstas queremos hacer predicciones.

Decimos que la regresión es lineal si cada  $b_i$  depende linealmente de los parámetros. En particular, consideramos una regresión lineal donde cada  $b_i$  es combinación lineal de las variables explicativas:

$$b_i = \sum_{j=1}^n x_j a_{ij}, \quad i = 1, \dots, m.$$

Esto da lugar a un sistema de ecuaciones lineales

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

La matriz  $A_{m \times n}$  se conoce como **matriz de regresión o de diseño**.

La estimación que hacemos de los parámetros de una regresión lineal nos conduce al

**Problema lineal de cuadrados mínimos.** Dada la matriz  $A \in \mathbb{R}^{m \times n}$  con  $m \geq n$  y el vector  $\mathbf{b} \in \mathbb{R}^m$ , hallar un vector  $\mathbf{x}^\dagger \in \mathbb{R}^n$  tal que

$$\|A\mathbf{x}^\dagger - \mathbf{b}\|_2^2 = \min_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2^2 \quad (2.6)$$

Con el método de Galerkin vimos que la discretización de la ecuación de absorción (2.5) nos da una expansión que se aleja de la solución analítica conforme usamos más funciones sombrero. Esta vez la discretizamos con el método de colocación. Aproximamos los valores de la función en los puntos de colocación.

Si tomamos menos puntos de colocación que el número de observaciones, no esperamos hallar una solución exacta del sistema de ecuaciones lineales que formamos; en cambio, buscamos una solución cercana tratando de reducir las discrepancias.

**Ejemplo 2.4.** Retomemos el Ejemplo 2.2. Generamos 40 observaciones  $g_i$  de la función  $g$  en los puntos

$$x_i = \frac{i-1}{4}, \quad i = 1, \dots, 40.$$

Queremos aproximar los valores de la función  $f : [0, 10] \rightarrow \mathbb{R}$  que cumple la ecuación (2.5) sobre la partición

$$y_j = \frac{j-1}{2}, \quad j = 1, \dots, 21$$

del intervalo  $[0, 10]$ .

Tomemos  $x_1, \dots, x_{40}$  como los puntos de colocación. A partir de la ecuación (2.5) tene-

mos que

$$g_i = \int_0^{10} e^{-0.01|x_i-y|} f(y) dy, \quad i = 1, \dots, 40.$$

Aproximamos cada integral definida mediante la cuadratura compuesta del trapecio con los 21 puntos  $y_j$  como nodos:

$$\int_0^{10} e^{-0.01|x_i-y|} f(y) dy \approx \frac{1}{4} \left[ e^{-0.01|x_i-y_1|} f(y_1) + e^{-0.01|x_i-y_{21}|} f(y_{21}) + 2 \sum_{j=2}^{20} e^{-0.01|x_i-y_j|} f(y_j) \right].$$

De esta manera formamos el sistema de ecuaciones lineales

$$\frac{1}{4} \begin{bmatrix} e^{-0.01|x_1-y_1|} & 2e^{-0.01|x_1-y_2|} & \dots & 2e^{-0.01|x_1-y_{20}|} & e^{-0.01|x_1-y_{21}|} \\ \vdots & \vdots & & \vdots & \vdots \\ e^{-0.01|x_{40}-y_1|} & 2e^{-0.01|x_{40}-y_2|} & \dots & 2e^{-0.01|x_{40}-y_{20}|} & e^{-0.01|x_{40}-y_{21}|} \end{bmatrix} \begin{bmatrix} f(y_1) \\ \vdots \\ f(y_{21}) \end{bmatrix} = \begin{bmatrix} g_1 \\ \vdots \\ g_{40} \end{bmatrix}.$$

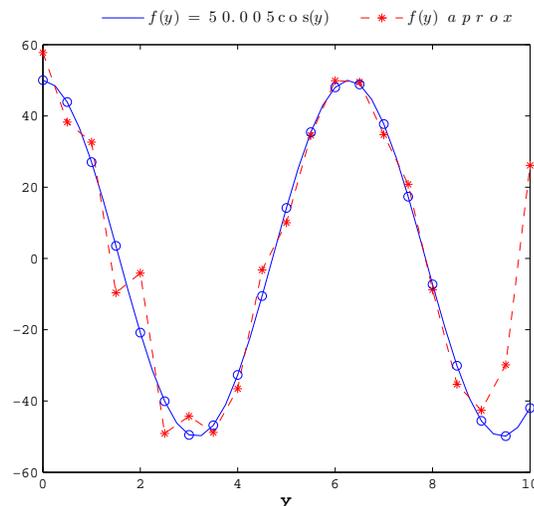
$A \qquad \qquad \qquad \mathbf{z} \qquad \qquad \qquad = \qquad \qquad \qquad \mathbf{b}$

La matriz  $A$  es de tamaño  $40 \times 21$ . Por lo que el sistema de ecuaciones está sobredeterminado. Una manera de obtener una solución aproximada de la ecuación  $Az = \mathbf{b}$  en  $\text{Col}(A)$  es con *el método de cuadrados mínimos*:

$$\min_{\mathbf{y} \in \mathbb{R}^{21}} \|Az - \mathbf{b}\|_2^2.$$

Los puntos críticos están dados por la solución de las *ecuaciones normales*

$$A^T Az = A^T \mathbf{b}$$



**Figura 2.4:** Solución  $f(y) = 50.005 \cos(y)$  de Ecuación (2.5) y valores aproximados  $f(y_i)$  por método de colocación.

Dado que  $A$  tiene rango completo por columnas, la matriz simétrica  $A^T A$  es positiva definida. Así que podemos usar la Factorización de Cholesky para resolver las ecuaciones normales, y con ello obtener los valores aproximados  $z_i = f(y_i)$ .

En la Figura 2.4 se muestran los valores  $f(y_i)$  obtenidos con el método de colocación y los comparamos con los de la solución analítica (2.2). En algunos puntos observamos que la separación entre los valores correspondientes es mayor que 10. Esto se debe a que los valores aproximados por el método de colocación se ven influenciados por la regla de cuadratura que usemos y los errores por redondeo.

### 2.2.1. Interpretación Geométrica

Examinamos la Ecuación  $A\mathbf{x} = \mathbf{b}$  para ver la geometría del método de cuadrados mínimos. Debido a que  $\mathbf{b}$  no necesariamente está en el espacio columna  $\text{Col}(A)$ , buscamos el vector  $\mathbf{p} \in \text{Col}(A)$  más cercano a  $\mathbf{b}$  en norma euclidiana:

$$\|\mathbf{p} - \mathbf{b}\|_2 = \min_{\mathbf{c} \in \text{Col}(A)} \|\mathbf{c} - \mathbf{b}\|_2.$$

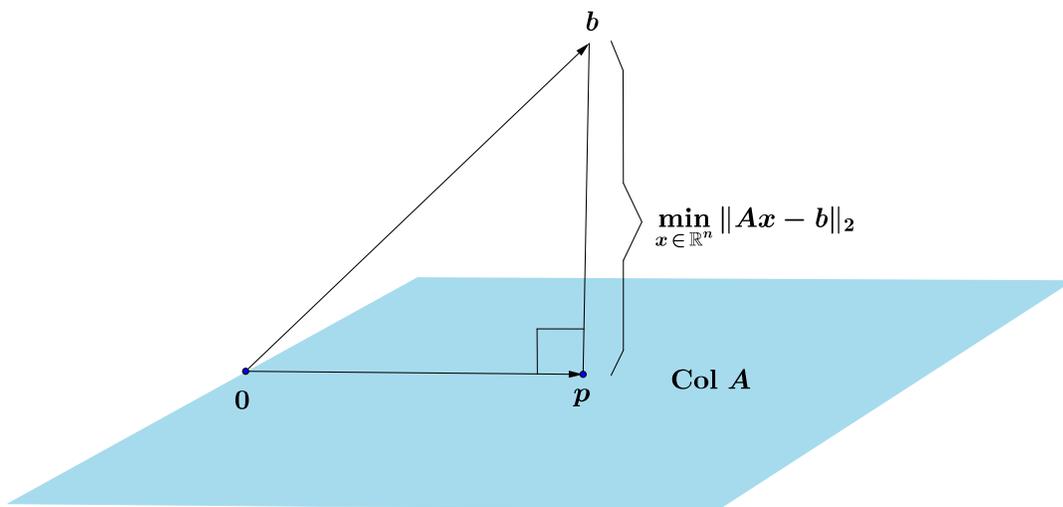
En este sentido, podemos decir que  $\|\mathbf{p} - \mathbf{b}\|_2$  es la distancia de  $\mathbf{b}$  a  $\text{Col}(A)$ . Geométricamente, el vector  $\mathbf{p}$  es una proyección de  $\mathbf{b}$  sobre  $\text{Col}(A)$  tal que la diferencia  $\mathbf{p} - \mathbf{b}$  sea perpendicular a las columnas de  $A$ . Véase Figura 2.5.

Puesto que  $\mathbf{p} \in \text{Col}(A)$ , existe un vector  $\mathbf{x}^\dagger \in \mathbb{R}^n$  tal que

$$A\mathbf{x}^\dagger = \mathbf{p}.$$

Para hallar  $\mathbf{x}^\dagger$ , resolvemos las ecuaciones normales

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



**Figura 2.5:** Proyección ortogonal de  $\mathbf{b}$  sobre  $\text{Col}(A)$

Supongamos que  $A$  tiene rango completo por columnas. Entonces  $A^T A$  es invertible. Sea

$$A^\dagger = (A^T A)^{-1} A^T.$$

Las ecuaciones normales tienen como única solución a

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

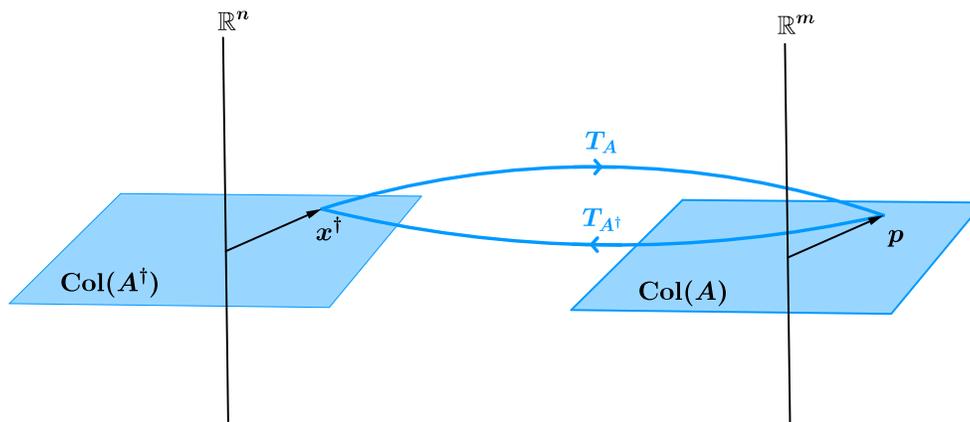
Este vector se conoce como *solución de cuadrados mínimos* de la Ecuación  $A\mathbf{x} = \mathbf{b}$ . Luego, el mapeo lineal

$$\begin{aligned} T_A : \text{Col}(A^\dagger) &\mapsto \text{Col}(A) \\ \mathbf{x}^\dagger &\mapsto A\mathbf{x}^\dagger \end{aligned}$$

transforma la solución  $\mathbf{x}^\dagger$  en la proyección  $\mathbf{p}$ , mientras que su inversa

$$\begin{aligned} T_{A^\dagger} : \text{Col}(A) &\mapsto \text{Col}(A^\dagger) \\ \mathbf{y} &\mapsto A^\dagger \mathbf{y}. \end{aligned}$$

transforma  $\mathbf{p}$  en  $\mathbf{x}^\dagger$  como se muestra en la Figura 2.6.



**Figura 2.6:** Las transformaciones  $T_A$  y  $T_{A^\dagger}$ .

## 2.3. Modelo Lineal General de Regresión

En esta sección usamos conceptos básicos de estadística para buscar un estimador de la solución de una regresión lineal con errores aleatorios. La clave es comprender el aspecto estadístico del método de cuadrados mínimos.

En 1809, Gauss justificó porqué el método de cuadrados mínimos puede verse como un procedimiento estadístico. Él supuso que los errores estaban correlacionados y distribuidos normalmente con esperanza cero y misma varianza. Más tarde, dió una fundamentación

teórica del método de cuadrados mínimos en sus memorias *Theoria Combinationis* de 1821 y 1823. Markov dio claridad a algunas de las hipótesis que usaba Gauss [10].

Supongamos que las observaciones ideales  $b_i^{\text{exacto}}$  están dadas por la regresión lineal

$$b_i^{\text{exacto}} = \sum_{j=1}^n x_j a_{ij} \quad i = 1, \dots, m,$$

Entonces el vector de observaciones libres de ruido es

$$\mathbf{b}^{\text{exacto}} = A\mathbf{x}.$$

La solución de cuadrados mínimos de la Ecuación (2.3) es

$$\mathbf{x}^\dagger = A^\dagger \mathbf{b}_{\text{exacto}},$$

Usualmente, las observaciones  $b_1, \dots, b_m$  contienen errores aleatorios  $\epsilon_1, \dots, \epsilon_m$ . A diferencia de los modelos deterministas, los modelos estadísticos toman en cuenta la presencia de estos errores mediante la siguiente relación:

$$\text{Observaciones} = \text{Modelo Matemático} + \text{Error Aleatorio.}$$

Así que

$$b_i = b_i^{\text{exacto}} + \epsilon_i, \quad i = 1, \dots, m.$$

Como cada  $\epsilon_i$  es una variable aleatoria, se sigue que  $b_i$  también lo es. Luego, con el vector aleatorio

$$\boldsymbol{\epsilon} = (\epsilon_1 \cdots \epsilon_m)^T,$$

obtenemos el vector de observaciones aleatorias

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

A partir de estas observaciones queremos hallar un estimador para el vector  $\mathbf{x}$ .

Usamos el método de cuadrados mínimos para estimar  $\mathbf{x}$ . Lo que hacemos es minimizar el tamaño de los errores aleatorios:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\boldsymbol{\epsilon}\|_2^2,$$

es decir,

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

Los puntos críticos de la función objetivo están dados por la solución

$$\mathbf{x}_{LS} = A^\dagger \mathbf{b}$$

del sistema de ecuaciones normales

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

$\mathbf{x}_{LS}$  es un estimador para  $\mathbf{x}$  llamado *estimador de cuadrados mínimos*.

**Observaciones 2.4:**

👉 El ruido que introduce el vector aleatorio  $\boldsymbol{\epsilon}$  en observaciones se propaga como

$$\boldsymbol{\epsilon}^\dagger = A^\dagger \boldsymbol{\epsilon}.$$

Ahora, veamos que características nos conviene que tenga un estimador de  $\boldsymbol{x}$ .

Decimos que  $\hat{\boldsymbol{x}}$  es un *estimador insesgado* de  $\boldsymbol{x}$  si su valor esperado coincide con el valor del parámetro, es decir,  $E(\hat{\boldsymbol{x}}) = \boldsymbol{x}$ . De este modo, los valores estimados están centrados alrededor de la solución.

El *mejor estimador lineal insesgado* de  $\boldsymbol{x}$  es aquel estimador lineal que tiene la varianza más pequeña de entre todos los estimadores insesgados de  $\boldsymbol{x}$ .

Buscamos condiciones suficientes para que el método de cuadrados mínimos nos proporcione un estimador con las características anteriores. Para ello examinamos los valores alrededor de los cuales se distribuyen los errores  $\epsilon_i$ , así como la dispersión de estos errores. En términos de la esperanza  $E(\epsilon_i)$  y la varianza  $\text{var}(\epsilon_i)$ , pedimos que los errores aleatorios cumplan con lo siguiente:

1.  $E(\epsilon_i) = 0$  para  $i = 1, \dots, n$

Los errores no están sesgados.

2.  $\text{var}(\epsilon_i) = \sigma^2$  para  $i = 1, \dots, n$

La varianza del error no varía a lo largo de las observaciones. Esta condición se llama *homocedasticidad*

3.  $E(\epsilon_i \epsilon_j) = 0$  para  $i, j = 1, \dots, n$  con  $i \neq j$ .

Los errores no están correlacionados.

Estas son las *condiciones de Gauss-Markov*, que en términos del vector aleatorio  $\boldsymbol{\epsilon}$  se expresan como

- \* valor esperado  $E(\boldsymbol{\epsilon}) = \mathbf{0}$ ,
- \* matriz de covarianza  $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 I$ .

El modelo  $\boldsymbol{b} = A\boldsymbol{x} + \boldsymbol{\epsilon}$  que satisface estas condiciones se conoce como *modelo lineal general de regresión*. Con este,  $\boldsymbol{x}_{LS}$  es el mejor estimador lineal insesgado de  $\boldsymbol{x}$ .

**Teorema 2.1** (Gauss-Markov [108]). Sea  $A \in \mathbb{R}^{m \times n}$  de rango completo por columnas, sea  $\boldsymbol{\epsilon}$  vector aleatorio de  $\mathbb{R}^m$  con  $E(\boldsymbol{\epsilon}) = \mathbf{0}$  y  $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 I$  y sea  $\boldsymbol{x}_{LS}$  el estimador de cuadrados mínimos del vector  $\boldsymbol{x}$  tal que  $\boldsymbol{b} = A\boldsymbol{x} + \boldsymbol{\epsilon}$ . Entonces entre la clase de estimadores lineales insesgados de  $\boldsymbol{a}^T \boldsymbol{x}$ ,  $\boldsymbol{a}^T \boldsymbol{x}_{LS}$  es el único estimador con varianza mínima para cualquier  $\boldsymbol{a} \in \mathbb{R}^n$ .

El Teorema 2.1 es la razón por la que el método de cuadrados mínimos se usa para resolver una regresión lineal en un modelo de Gauss-Markov.

### Observaciones 2.5:

 No es necesario que el ruido  $\epsilon$  tenga distribución normal para usar el Teorema 2.1.

## 2.4. Problemas de cuadrados mínimos con rango deficiente

Hemos resuelto la ecuación  $A\mathbf{x} = \mathbf{b}$  cuando la matriz  $A$  de tamaño  $m \times n$  con  $m \geq n$  tiene rango completo. Si  $A$  tiene rango deficiente,  $A^T A$  no es invertible. Así que evitamos resolver las ecuaciones normales y procedemos de una manera distinta. Lo que hacemos es cambiar la base de  $\text{Col}(A)$  por otra que tiene vectores ortonormales. Una manera es usar la *Descomposición en Valores Singulares (SVD)*. Sea  $r$  el rango de  $A$ . Una SVD de  $A$  es una factorización matricial

$$A = \underbrace{\begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_r & \mathbf{u}_{r+1} & \cdots & \mathbf{u}_m \end{bmatrix}}_U \underbrace{\left[ \begin{array}{cc|c} \sigma_1 & 0 & \mathbf{0} \\ & \ddots & \\ 0 & \sigma_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} \end{array} \right]}_{\Sigma}_{m \times n} \underbrace{\begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_r & \mathbf{v}_{r+1} & \cdots & \mathbf{v}_n \end{bmatrix}^T}_{V^T},$$

donde las columnas de  $U$  y  $V$  son bases ortonormales de  $\mathbb{R}^m$  y  $\mathbb{R}^n$ , respectivamente, y  $\sigma_1 \geq \cdots \geq \sigma_r > 0$  son los *valores singulares* de  $A$ . Consulte el Apéndice para una prueba.

### 2.4.1. De la esfera al hiperelipsoide vía SVD

Empleemos la SVD de  $A$  para ver como esta matriz deforma la esfera unitaria

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}.$$

Sea  $\mathbf{x} \in \mathcal{S}$ . Consideremos la SVD

$$A = U\Sigma V^T.$$

El vector con las coordenadas de  $\mathbf{x}$  en la base de vectores singulares  $\mathbf{v}_1, \dots, \mathbf{v}_n$  es  $V^T \mathbf{x}$ . Dado que la matriz  $V$  es ortogonal, tenemos que  $V^T \mathbf{x} \in \mathcal{S}$ . Así que la matriz  $V^T$  solamente gira la esfera unitaria.

El vector  $\Sigma V^T \mathbf{x}$  tiene coordenadas

$$z_i = \begin{cases} \sigma_i \mathbf{v}_i^T \mathbf{x}, & i \in \{1, \dots, r\}, \\ 0 & i \in \{r+1, \dots, m\} \end{cases}$$

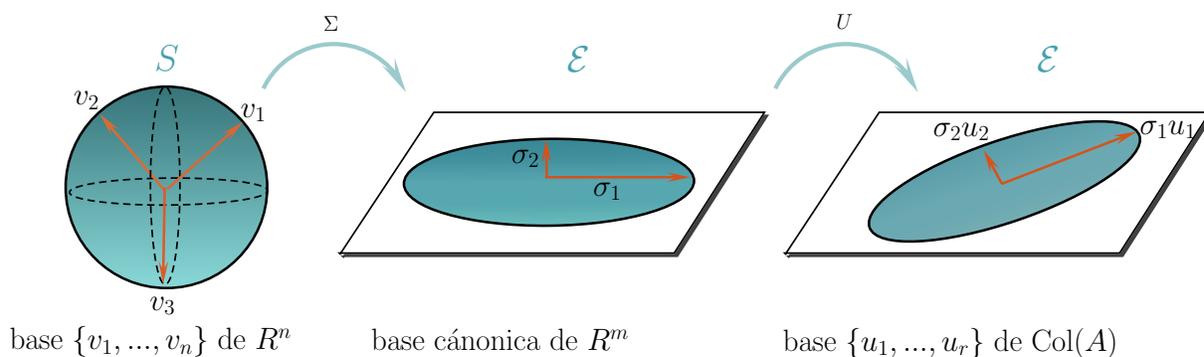
en la base canónica  $\{\mathbf{e}_1, \dots, \mathbf{e}_m\}$  de  $\mathbb{R}^m$ . Éstas cumplen que

$$\frac{z_1^2}{\sigma_1^2} + \dots + \frac{z_r^2}{\sigma_r^2} \leq 1.$$

Luego,  $\Sigma V^T \mathbf{x}$  está en el hiperelipsoide  $\mathcal{E}$  con semiejes  $\mathbf{e}_1, \dots, \mathbf{e}_r$  de longitud  $\sigma_1, \dots, \sigma_r$ . Por consiguiente, la matriz  $\Sigma$  transforma  $\mathcal{S}$  en  $\mathcal{E}$ .

La matriz  $U$ , transforma el vector  $\Sigma V^T \mathbf{x}$  en  $A\mathbf{x}$ . Por lo que  $A\mathbf{x}$  tiene las mismas coordenadas  $z_i$  de  $\Sigma V^T \mathbf{x}$ , pero en la base ortonormal  $\{u_1, \dots, u_n\}$ . Por lo que  $U$  solamente gira el hiperelipsoide  $\mathcal{E}$ .

En consecuencia, la matriz  $A$  transforma la esfera unitaria en un hiperelipsoide sobre  $\mathbb{R}^r$  con semiejes  $u_1, \dots, u_r$  de longitudes  $\sigma_1, \dots, \sigma_r$ .



**Figura 2.7:** Deformación de la esfera unitaria bajo  $A$

### Observaciones 2.6:

Si  $A$  tiene rango deficiente, entonces el hiperelipsoide  $\mathcal{E}$  se encuentra en un subespacio de menor dimensión que  $\mathbb{R}^n$  e incluimos el interior de  $\mathcal{E}$ .

Las respectivas longitudes de los semiejes mayor y menor son

$$\sigma_1 = \max_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x}\|_2 \quad \text{y} \quad \sigma_r = \min_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x}\|_2,$$

### 2.4.2. SVD en el Problema de Cuadrados Mínimos

Sea  $A = U\Sigma V^T$  una SVD de la matriz  $A$ . Entonces

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2^2 = \min_{\mathbf{x} \in \mathbb{R}^n} \|U\Sigma V^T \mathbf{x} - \mathbf{b}\|_2^2$$

Debido a que  $U$  es una matriz ortogonal, tenemos que

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|U\Sigma V^T \mathbf{x} - \mathbf{b}\|_2^2 = \min_{\mathbf{x} \in \mathbb{R}^n} \|\Sigma V^T \mathbf{x} - U^T \mathbf{b}\|_2^2.$$

Sean

$$\mathbf{c} = U^T \mathbf{b} \quad \text{e} \quad \mathbf{y} = V^T \mathbf{x}.$$

Dado que  $V$  es ortogonal,  $V^T$  es invertible, entonces

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\Sigma V^T \mathbf{x} - U^T \mathbf{b}\|_2^2 = \min_{\mathbf{y} \in \mathbb{R}^n} \|\Sigma \mathbf{y} - \mathbf{c}\|_2^2.$$

Los elementos de la matriz  $\Sigma$  son iguales a cero fuera de su diagonal principal, y sobre esta diagonal solamente los valores singulares  $\sigma_1, \dots, \sigma_r$  son distintos de cero. Así que

$$\min_{\mathbf{y} \in \mathbb{R}^n} \|\Sigma \mathbf{y} - \mathbf{c}\|_2^2 = \min_{\mathbf{y} \in \mathbb{R}^n} \left[ \sum_{i=1}^r |\sigma_i y_i - c_i|^2 + \sum_{i=1}^r |c_i|^2 \right].$$

Luego, todo vector  $\mathbf{y}$  con componentes

$$y_i = \begin{cases} \frac{c_i}{\sigma_i}, & i = 1, \dots, r, \\ \text{arbitrario}, & i = r + 1, \dots, n, \end{cases} \quad (2.7)$$

minimiza el tamaño del residuo  $\Sigma \mathbf{y} - \mathbf{c}$ . En consecuencia, el conjunto de vectores  $V \mathbf{y}$ , donde  $y_i$  está dado por (2.7) nos da una familia de soluciones de cuadrados mínimos. Dentro de esa familia, seleccionamos el vector

$$\mathbf{x}^\dagger = V \mathbf{y}$$

de menor norma euclidiana. El vector  $\mathbf{x}^\dagger$  se llama *solución de cuadrados mínimos de norma mínima*. Puesto que  $V$  es una matriz ortogonal, tenemos que

$$\|\mathbf{x}^\dagger\|_2 = \|\mathbf{y}\|_2.$$

Como queremos minimizar  $\|\mathbf{x}^\dagger\|_2$  y las componentes  $\mathbf{y}$  están dadas por (2.7), entonces basta tomar

$$y_{i+r} = \dots = y_n = 0.$$

Así que podemos expandir  $\mathbf{x}^\dagger$  en los primeros  $r$  vectores singulares de izquierda  $\mathbf{v}_1, \dots, \mathbf{v}_r$  como

$$\mathbf{x}^\dagger = \sum_{i=1}^r \frac{c_i}{\sigma_i} \mathbf{v}_i$$

En resumen, la solución de cuadrados mínimos de norma mínima  $\mathbf{x}^\dagger$  para la ecuación  $A \mathbf{x} = \mathbf{b}$  se obtiene de la siguiente manera:

1. Calcular una SVD  $A = U \Sigma V^T$ ,
2. Formar el producto  $\mathbf{c} = U^T \mathbf{b}$ ,
3.  $\mathbf{x}^\dagger = (c_1/\sigma_1) \mathbf{v}_1 + \dots + (c_r/\sigma_r) \mathbf{v}_r$ .

**Ejemplo 2.5 (Reconstrucción de haz de luz [109]).** El haz de luz que emite una fuente pasa por una abertura delgada de longitud infinita y ancho uno. La luz se difracta al pasar por la abertura. La intensidad de luz incidente  $f$  sobre la abertura está en función del ángulo de  $\theta \in [-\pi/2, \pi/2]$ , mientras que la intensidad de luz difractada  $g$  depende de un ángulo  $\phi \in [-\pi/2, \pi/2]$ . Con el paquete REGUTOOLS [51] generamos la Tabla 2.3 con 20 observaciones de la intensidad  $g$  en los ángulos

$$\phi_i = \left(i - \frac{1}{2}\right) \frac{\pi}{20} - \frac{\pi}{2}, \quad i = 1, \dots, 20.$$

Tenemos el siguiente problema:

Dados los valores de la función  $g : [-\pi/2, \pi/2] \rightarrow \mathbb{R}$  para la intensidad de luz difractada, encontrar los valores de la función  $f : [-\pi/2, \pi/2] \rightarrow \mathbb{R}$  para la intensidad de luz incidente.

$i$	1	2	3	4	5	6	7	8	9	10
$\phi_i$	-1.492	-1.335	-1.178	-1.021	-0.863	-0.706	-0.549	-0.392	-0.235	-0.078
$g(\phi_i)$	0.549	0.847	1.265	1.813	2.451	3.067	3.501	3.637	3.497	3.244
$i$	11	12	13	14	15	16	17	18	19	20
$\phi_i$	0.078	0.235	0.392	0.549	0.706	0.863	1.021	1.178	1.335	1.492
$g(\phi_i)$	3.038	2.902	2.730	2.422	1.983	1.499	1.061	0.718	0.473	0.307

**Tabla 2.3:** Valores de función  $g$  en ángulos de difracción  $\phi_i$

El modelo matemático que relaciona las funciones  $f$  y  $g$  está dado por la ecuación integral

$$\int_{-\pi/2}^{\pi/2} (\cos \phi + \cos \theta)^2 \left( \frac{\sin(\pi(\sin \phi + \sin \theta))}{\pi(\sin \phi + \sin \theta)} \right)^2 f(\theta) d\theta = g(\phi). \quad (2.8)$$

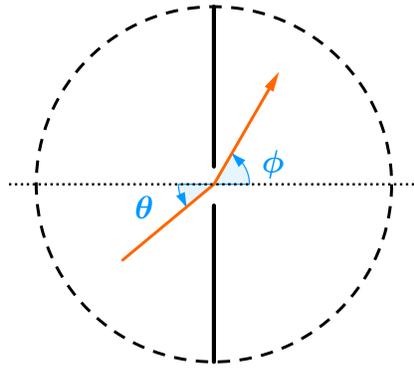
El núcleo representa la difracción de la luz con longitud de onda uno en la abertura de acuerdo con la formulación de la difracción de Kirchhoff [109], §8.3 [11].

Para conocer los valores de  $f$ , discretizamos la ecuación (2.8) con el método de colocación en los 20 ángulos  $\theta_i = \phi_i$ . La integral se aproxima con una cuadratura compuesta del punto medio. Los nodos son los  $\phi_i$ s. Así, obtenemos un sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b}$ , donde

$$a_{i,j} = \frac{\pi}{20} (\cos \phi_i + \cos \phi_j)^2 \left( \frac{\sin(\pi(\sin \phi_i + \sin \phi_j))}{\pi(\sin \phi_i + \sin \phi_j)} \right)^2,$$

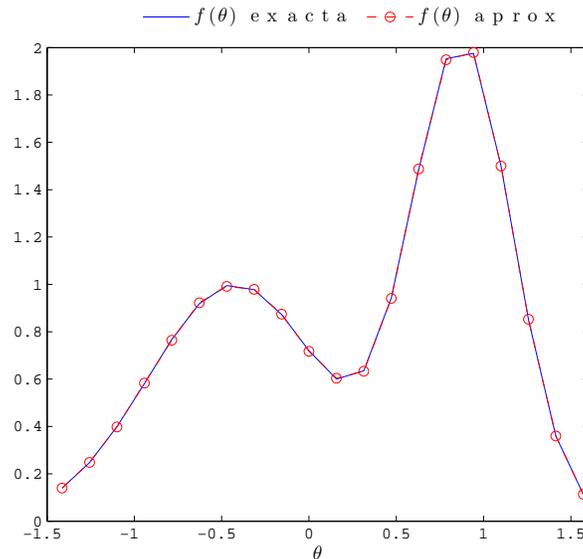
$$b_i = g(\phi_i), \quad i, j = 1, \dots, 20.$$

$$x_i = f(\phi_i),$$



**Figura 2.8:** Ángulo de incidencia  $\theta$  y ángulo de difracción  $\phi$  del haz de luz que pasa por la abertura.

La matriz  $A$  de tamaño  $20 \times 20$  tiene rango numérico 18. Por lo que buscamos una solución de cuadrados mínimos. Por lo que  $A$  es de rango deficiente. Así que no podemos obtener una Factorización  $QR$  de  $A$  con matriz triangular superior invertible, ni podemos usar la Factorización de Cholesky de  $A^T A$ . Lo que hacemos es usar la SVD para calcular la solución de cuadrados mínimos de norma mínima  $\mathbf{x}^\dagger$ .



**Figura 2.9:** Función  $f$  y sus valores aproximados por  $\mathbf{x}^\dagger$

En la figura 2.9 comparamos los valores de la solución

$$f(\theta) = 2 \exp(-6(\theta - 0.8)^2) + \exp(-2(\theta + 0.5)^2)$$

de la ecuación (2.8) con los valores de las componentes del vector  $\mathbf{x}^\dagger$ . Ahora, la discrepancia entre los respectivos valores no es mayor que 0.1. Por eso los valores de  $f$  y los de  $\mathbf{x}^\dagger$  se traslapan.

### 2.4.3. Propiedades de la SVD

Una de las propiedades de la SVD nos permite saber si una matriz dada está cerca de otra con rango deficiente. Para precisar esto, medimos la distancia entre matrices con la norma espectral

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2.$$

#### Observaciones 2.7:

☞ La norma espectral es la longitud del semieje mayor del hiperelipsoide que resulta de deformar la esfera unitaria con la matriz  $A$ . Esta longitud es precisamente el mayor de los valores singulares de  $A$ , a saber,  $\sigma_1$ , entonces

$$\|A\|_2 = \sigma_1.$$

Buscamos una matriz que minimize  $\|A - B\|_2$  para cada matriz  $B$  de rango  $k \leq r$ . A partir de la SVD de  $A$ , definimos las matrices

$$A_k = U \left[ \begin{array}{cc|c} \sigma_1 & 0 & \mathbf{0} \\ & \ddots & \\ 0 & & \sigma_k \\ \hline & \mathbf{0} & \mathbf{0} \end{array} \right]_{m \times n} V^T, \quad k = 1, \dots, r.$$

#### Observaciones 2.8:

☞ En términos de los vectores singulares de izquierda y derecha,  $\mathbf{v}_i$  y  $\mathbf{u}_i$ , respectivamente, podemos formar  $A_k$  como la suma de  $k$  matrices de rango uno:

$$A_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

De este modo, solamente necesitamos los primeros  $k$  valores singulares de  $A$  y sus respectivos vectores singulares para formar  $A_k$ .

Las matrices  $A_k$  nos dan la distancia más pequeña entre  $A$  y las matrices de rango  $k$ .

**Teorema 2.2** ([37]). *Sea  $A \in \mathbb{R}^{m \times n}$  una matriz de rango  $r$ . Entonces*

$$\min_{\substack{B \in \mathbb{R}^{m \times n} \\ \text{rank}(B) \leq k}} \|B - A\|_2 = \|A - A_k\|_2 = \sigma_{k+1} \quad k = 1, \dots, r.$$

Este Teorema nos dice que a pesar de que una matriz sea invertible en teoría, si ésta se aproxima por una matriz de rango deficiente, entonces algunas de sus columnas pueden ser tratadas como linealmente dependientes en los cálculos numéricos. A su vez, podemos detectar esta situación mediante los valores singulares de la matriz.

Si usamos la norma de Frobenius

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|^2},$$

entonces

$$\|A\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2}$$

porque  $U$  y  $V$  son matrices ortogonales. Con esta norma, cada  $A_k$  es aproximación de menor rango de  $A$ .

**Teorema 2.3** ([37]). *Sea  $A \in \mathbb{R}^{m \times n}$  una matriz de rango  $r$ . Entonces*

$$\min_{\substack{B \in \mathbb{R}^{m \times n} \\ \text{rank}(B) \leq k}} \|B - A\|_F = \|A - A_k\|_F \quad k = 1, \dots, r.$$

#### 2.4.4. Aplicaciones la SVD

Además de usar la SVD en el problema de cuadrados mínimos, esta factorización matricial tiene aplicaciones en distintas áreas de la Ciencia:

- \* Análisis del peso de recién nacidos [87],
- \* Extracción de electrocardiogramas fetales [14],
- \* Reconocimiento de patrones y de rostros [32],
- \* Recuperación de información en bases de datos [14],
- \* Compresión de imágenes digitales [24].

**Ejemplo 2.6 (Compresión de Imágenes Digitales [2]).** Vamos a usar la SVD para almacenar la imagen digital mostrada en la Figura 2.10 con una menor cantidad de píxeles. Esta es una imagen de la Biblioteca Central de la UNAM de  $1006 \times 1418$  píxeles en JPG.



**Figura 2.10:** Imagen de la Biblioteca central de la UNAM

Una imagen digital  $m \times n$  en blanco y negro tiene  $m$  columnas cada una con  $n$  píxeles, es decir, su matriz es de tamaño  $n \times n$ . La componente  $(i, j)$  es un número entre 0 y 1. El cero corresponde al color negro, mientras que 1 indica el color blanco.

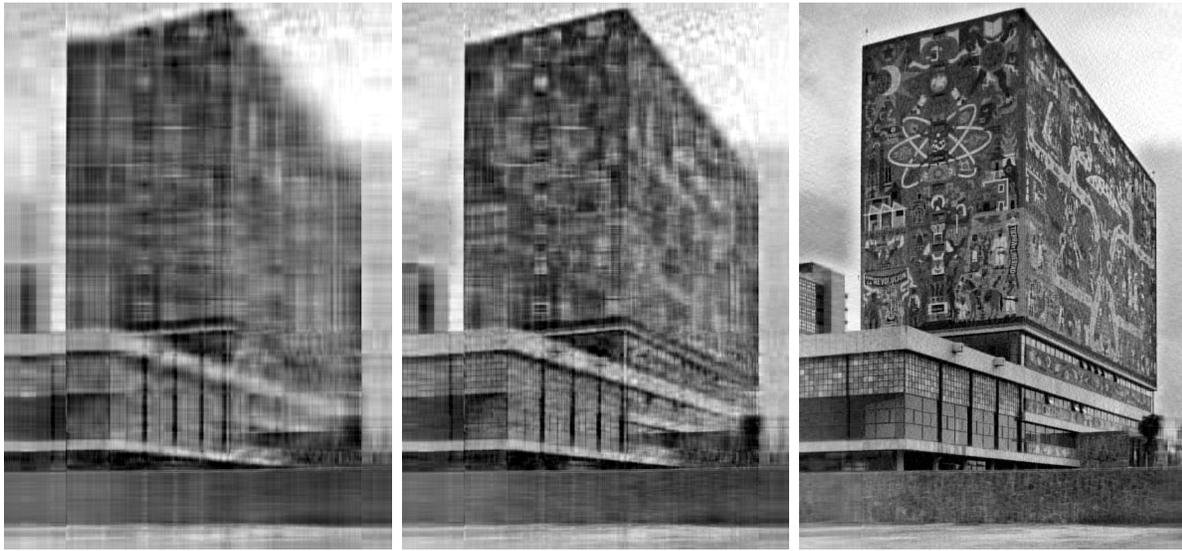
Los niveles de truncamiento que usamos son  $k = 10, 20, 100$ . En cada caso generamos la matriz  $A_k$ . En la Figura 2.11, mostramos las imágenes obtenidas. Conforme aumentamos  $k$ , la SVD nos dan una imagen que donde se aprecian más detalles de la escena. Observamos que podemos generar una aproximación correcta con  $k = 100$ , en lugar de usar los 1006 valores singulares. En cada caso, medimos el error relativo

$$\frac{\|A - A_k\|_F}{\|A\|_F} = \sqrt{\frac{\sum_{i=k+1}^r \sigma_i^2}{\sum_{i=1}^r \sigma_i^2}}$$

y el cociente del número de píxeles de la imagen comprimida entre el el número de píxeles de la imagen original  $(m + n)k/(mn)$  para darnos una idea de la reducción que ofrece  $A_k$  respecto a  $A$ . Consulté [40] para medidas de compresión de imágenes.

$k$	$\ A - A_k\ _F / \ A\ _F$	$(m + n)k / (mn)$
10	0.2006	.016993
20	0.1597	.033985
100	0.0666	.16993

**Tabla 2.4:** Errores relativos y cocientes  $(m + n)k/(mn)$  para distintos valores de  $k$ .

(a)  $k = 10$ (b)  $k = 20$ (c)  $k = 100$ 

**Figura 2.11:** Compresión de imagen de la Biblioteca Central con SVD mediante la matriz  $A_k$  para tres niveles de truncamiento  $k$ .

## 2.5. Condicionamiento

### 2.5.1. Problemas Mal Condicionados

La discretización de la ecuación integral (2.1) nos da un sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b}$ . En la práctica, tanto el vector  $\mathbf{b}$  como la matriz  $A$  tienen perturbaciones. Considere el sistema perturbado  $(A + E)\mathbf{x} = \mathbf{b} + \mathbf{e}$ . Queremos medir como los errores relativos de los datos  $\|\mathbf{e}\|/\|\mathbf{b}\|$  y  $\|E\|/\|A\|$  amplifican el error relativo de la solución  $\|\mathbf{x} - \underline{\mathbf{x}}\|/\|\mathbf{x}\|$ .

**Teorema 2.4** ([23]). *Suponga que  $A$  es invertible,  $\mathbf{b} \neq 0$  y  $\|A^{-1}\| \|E\| < 1$ . Si*

$$A\mathbf{x} = \mathbf{b} \quad \text{y} \quad (A + E)\underline{\mathbf{x}} = \mathbf{b} + \mathbf{e}.$$

entonces

$$\frac{\|\mathbf{x} - \underline{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(A)}{1 - \kappa(A) \frac{\|E\|}{\|A\|}} \left( \frac{\|E\|}{\|A\|} + \frac{\|\mathbf{e}\|}{\|\mathbf{b}\|} \right).$$

donde  $\kappa(A) = \|A\| \|A^{-1}\|$ .

**Observaciones 2.9:**

☞  $\kappa(A) \geq 1$ .

☞ Cuando  $A$  es invertible, tenemos que

$$\min\{\|E\| : A + E \text{ es singular}\} = \frac{1}{\|A^{-1}\|}.$$

Así, la hipótesis  $\|A^{-1}\| \|E\| < 1$  del Teorema 2.4 asegura que  $A + E$  es invertible [24].

$\kappa(A)$  se conoce como el **número de condición** de la matriz  $A$ . Si el valor de  $\kappa(A)$  es grande, por ejemplo de  $10^{16}$ ,  $10^{50}$ ,  $10^{100}$  o mayor, se dice que  $A$  está **mal condicionada**. De otro modo, está **bien condicionada**. Decidir que tan grande debe ser  $\kappa(A)$  para que  $A$  esté mal condicionada depende tanto de la precisión usada para la matriz y el lado derecho como de la precisión deseada en la solución calculada. En particular, si los errores relativos tanto en la matriz como en el lado derecho sean del orden de  $10^{-d}$ , entonces para que el error relativo en la solución tenga una precisión de al menos  $10^{-t}$ , necesitamos que  $\kappa(A) \leq 0.5 \times 10^{d-t}$ .

En aritmética exacta, las soluciones  $\underline{\mathbf{x}}$  y  $\underline{\underline{\mathbf{x}}}$  de las ecuaciones

$$A\mathbf{x} = \mathbf{b} \quad \text{y} \quad (A + E)\mathbf{x} = \mathbf{b} + \mathbf{e}.$$

satisfacen  $\|\underline{\mathbf{x}} - \underline{\underline{\mathbf{x}}}\| \rightarrow 0$  cuando  $\|\mathbf{e}\| \rightarrow 0$  debido a la continuidad de  $A^{-1}$ . En cambio, si  $A$  está mal condicionada, en aritmética finita puede pasar que  $\|\underline{\mathbf{x}} - \underline{\underline{\mathbf{x}}}\| \gg 0$  cuando  $\|\mathbf{e}\| \approx \epsilon_{mach}$ .

Puesto que todas normas son equivalentes en un espacio vectorial de dimensión finita, tenemos que los números de condición de matrices invertibles en  $\mathbb{R}^{n \times n}$  son equivalentes en el siguiente sentido:

$$\begin{aligned} 1/n \cdot \kappa_2(A) &\leq \kappa_1(A) \leq n\kappa_2(A), \\ 1/n \cdot \kappa_\infty(A) &\leq \kappa_2(A) \leq n\kappa_\infty(A), \\ 1/n^2 \cdot \kappa_1(A) &\leq \kappa_\infty(A) \leq n^2\kappa_1(A). \end{aligned}$$

En particular, cuando usamos la norma espectral, podemos expresar  $\kappa_2(A)$  en términos de los valores singulares  $\sigma_1 \geq \dots \geq \sigma_n$  de  $A$  como

$$\kappa_2(A) = \sigma_1/\sigma_n,$$

debido a que

$$\|A\|_2 = \max_{\|\mathbf{x}\|_2=1} \|A\mathbf{x} - \mathbf{b}\|_2 = \sigma_1, \quad \text{y} \quad \frac{1}{\|A^{-1}\|_2} = \min_{\|\mathbf{x}\|_2=1} \|A\mathbf{x} - \mathbf{b}\|_2 = \sigma_n.$$

### Observaciones 2.10:

☞ Vimos que la matriz  $A$  deforma la esfera unitaria  $\mathcal{S}$  en un hiperelipsoide  $\mathcal{E}$  con semiejes de longitudes  $\sigma_1, \dots, \sigma_n$ . Por lo que  $\kappa_2(A)$  es el cociente entre las longitudes de los semiejes mayor y menor de  $\mathcal{E}$ .

☞ En general, para una matriz  $A$  de rango  $r$ ,

$$\kappa_2(A) = \sigma_1/\sigma_r.$$

En ese caso,  $A$  está mal condicionada si por ejemplo  $\sigma_1/\sigma_r \geq 1/\epsilon_{mach}$ .

Cuando  $\mathbf{b} \notin \text{Col}(A)$ , examinamos la sensibilidad de la solución de cuadrados mínimos de la ecuación  $A\mathbf{x} = \mathbf{b}$  respecto a pequeñas perturbaciones en  $A$  y  $\mathbf{b}$ .

**Teorema 2.5** ([59]). Sea  $A \in \mathbb{R}^{m \times n}$  de rango completo con  $m \geq n$ . Sean

$$\underline{\mathbf{x}} = \arg \min_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2 \quad \text{y} \quad \underline{\underline{\mathbf{x}}} = \arg \min_{\mathbf{x} \in \mathbb{R}^n} \|(A + E)\mathbf{x} - (\mathbf{b} + \mathbf{e})\|_2.$$

Sea  $\mathbf{r} = \mathbf{b} - A\underline{\mathbf{x}}$ . Si

$$\|E\|_2 \leq \alpha \|A\|_2, \quad \|\mathbf{e}\|_2 \leq \alpha \|\mathbf{b}\|_2 \quad \text{y} \quad \kappa_2(A)\alpha < 1,$$

entonces

$$\frac{\|\underline{\mathbf{x}} - \underline{\underline{\mathbf{x}}}\|_2}{\|\underline{\mathbf{x}}\|_2} \leq \frac{\kappa_2(A)\alpha}{1 - \kappa_2(A)\alpha} \left( 2 + (\kappa_2(A) + 1) \frac{\|\mathbf{r}\|_2}{\|A\|_2 \|\underline{\mathbf{x}}\|_2} \right).$$

El Teorema 2.5 nos dice que  $\|\mathbf{r}\|_2$  influye en el tamaño de la región donde se encuentra la solución de cuadrados mínimos cuando perturbamos  $A$  y  $\mathbf{b}$ .

### 2.5.2. Problemas Discretos Mal Planteados

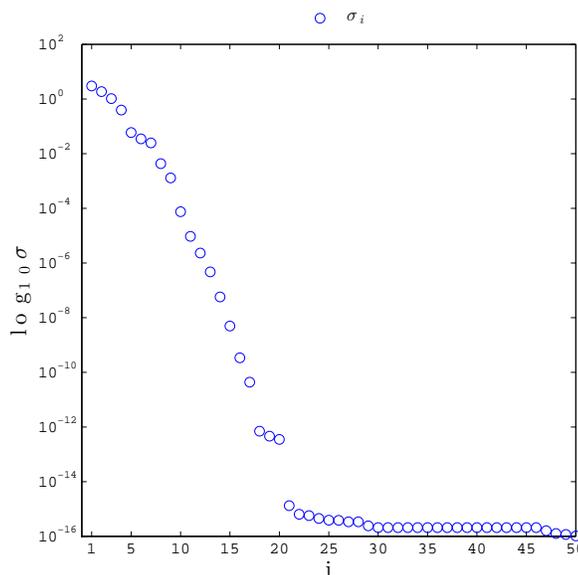
Relacionamos los  $r$  valores singulares positivos  $\sigma_i$  de  $A$  con el condicionamiento. En particular, nos interesa dos problemas mal condicionados:

\* **Problema de rango deficiente.** La sucesión  $\{\sigma_i\}_{i=1}^n$  decaí a cero y tiene al menos un salto, esto es,  $\sigma_k \gg \sigma_{k+1}$  para algún  $k \in \{1, \dots, r-1\}$ .

\* **Problema discreto mal planteado.** La sucesión  $\{\sigma_i\}_{i=1}^n$  decaí a cero sin saltos.

En ambos, el mal condicionamiento se refleja en la manera en que decaen los valores singulares.

**Ejemplo 2.7.** Retomemos la reconstrucción del haz de luz del Ejemplo 2.5. Esta vez generamos 50 observaciones de la función  $g$  en el intervalo  $[-\pi/, \pi/2]$ . Con el mismo método de discretización, formamos un sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b}$ . Esta vez  $A$  es de tamaño  $50 \times 50$ . Ordenamos sus valores singulares  $\sigma_1 \geq \dots \geq \sigma_{50}$ .



**Figura 2.12:** Gráfica en escala logarítmica de los valores singulares  $\sigma_i$  de la matriz  $A$  del Ejemplo 2.5 contra su subíndice  $i$ . Usamos 50 observaciones.

En la Figura 2.12 mostramos la gráfica de los valores singulares de  $A$ , marcados por  $(\circ)$ , contra su subíndice  $i$ . El eje vertical  $\sigma$  está en escala logarítmica base 10. Notamos dos saltos. El primero entre  $\sigma_{17} = 4.331 \times 10^{-11}$  y  $\sigma_{18} = 6.966 \times 10^{-13}$  y el otro ocurre entre  $\sigma_{20} = 3.51 \times 10^{-12}$  y  $\sigma_{21} = 1.317 \times 10^{-15}$ . Los siguientes  $\sigma_i$  decaen rápidamente hasta el orden de  $10^{-16}$ . Esto nos dice que el problema discreto del Ejemplo 2.5 es de rango numéricamente deficiente.

Como los valores singulares pueden tender de diferentes maneras a cero, tenemos distintos grados de mal condicionamiento. Distinguimos tres tipos de problemas [1]:

\* *Ligeramente mal planteado:*

$$\sigma_j = \mathcal{O}(j^{-\alpha}), \quad \alpha \leq 1.$$

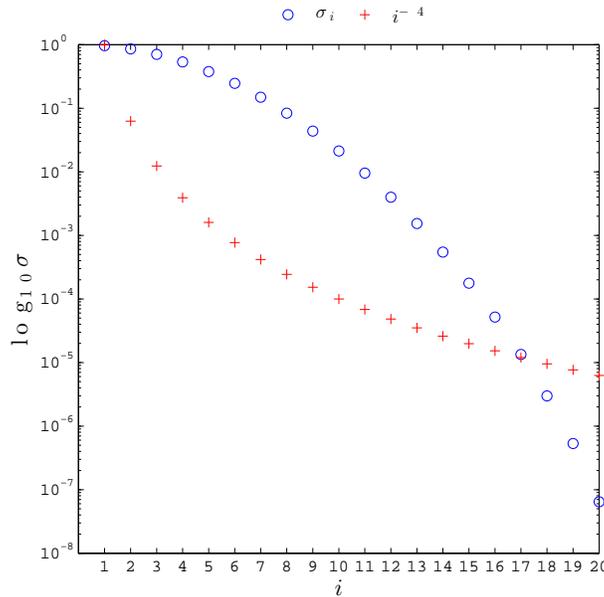
\* *Moderadamente mal planteado.* Los recíprocos de los valores singulares tienen orden polinomial:

$$\sigma_j = \mathcal{O}(j^{-\alpha}), \quad \alpha > 1.$$

\* *Severamente mal planteado* Los valores singulares tienden a cero exponencialmente:

$$\sigma_j = \mathcal{O}(e^{-j^\alpha}), \quad \alpha > 0.$$

**Ejemplo 2.8.** Cuando discretizamos el problema de deconvolución del Ejemplo 1.2, por el método de colocación de 20 puntos, obtenemos un sistema de ecuaciones  $A\mathbf{x} = \mathbf{b}$ , donde  $A$  es la matriz invertible  $20 \times 20$ .



**Figura 2.13:** Gráfica en escala logarítmica de los valores singulares  $\sigma_i$  de la matriz  $A_{20 \times 20}$  del Ejemplo 1.2 contra su subíndice  $i$  junto con la de la sucesión  $\{i^{-4}\}_{i=1}^{20}$ .

En la Figura 2.13 mostramos la gráfica de los valores singulares  $\sigma_1 \geq \dots \geq \sigma_{20}$  de  $A$ , marcados por  $(\circ)$ , contra su subíndice  $i$  en una escala logarítmica base 10 sobre el eje vertical. La sucesión  $\{\sigma_i\}$  decae gradualmente sin saltos desde  $\sigma_1 = 9.614 \times 10^{-1}$  hasta  $\sigma_{20} = 6.464 \times 10^{-8}$ , pero no decae tan rápido como  $i^{-4}$ . En consecuencia, el problema discreto de deconvolución del Ejemplo 1.2 está moderadamente mal planteado.

A continuación vemos un ejemplo donde la discretización de una ecuación integral por el método de colocación nos da un sistema de ecuaciones lineales mal condicionado que está severamente mal planteado. Para darnos cuenta, comparamos el decaimiento de los valores singulares de la matriz de coeficientes con el de la exponencial del subíndice. Esto explica porque los valores aproximados se alejan de los de la solución analítica.

**Ejemplo 2.9.** Queremos aproximar los valores de la función  $f : [0, 1] \rightarrow \mathbb{R}$  que cumple la ecuación integral

$$\int_0^1 \sqrt{s^2 + t^2} f(t) dt = \frac{1}{3} \left( (1 + s^2)^{3/2} - s^3 \right), \quad 0 \leq s \leq 1. \quad (2.9)$$

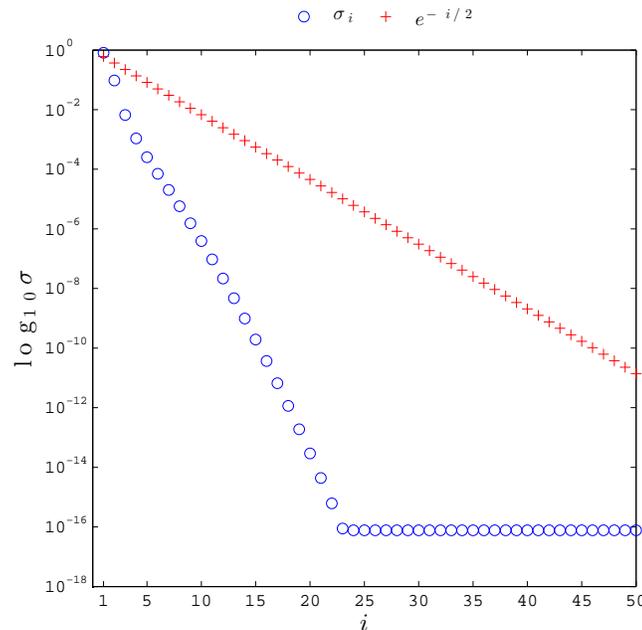
Discretizamos en los puntos de colocación

$$s_i = \frac{1}{50} \left( i - \frac{1}{2} \right), \quad i = 1, \dots, 50.$$

y aproximamos las integrales con la cuadratura compuesta del punto medio de 50 puntos. Así, obtenemos el sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b}$ , donde

$$\begin{aligned} x_i &= f(s_j), \\ a_{i,j} &= \frac{1}{50} \sqrt{s_i^2 + s_j^2}, \quad i, j = 1, \dots, 50. \\ b_i &= \frac{1}{3} \left( (1 + s_i^2)^{3/2} - s_i^3 \right), \end{aligned}$$

En este caso, la matriz  $A$  está mal condicionada ya que  $\kappa_2(A) \approx 2.805 \times 10^{18}$ .

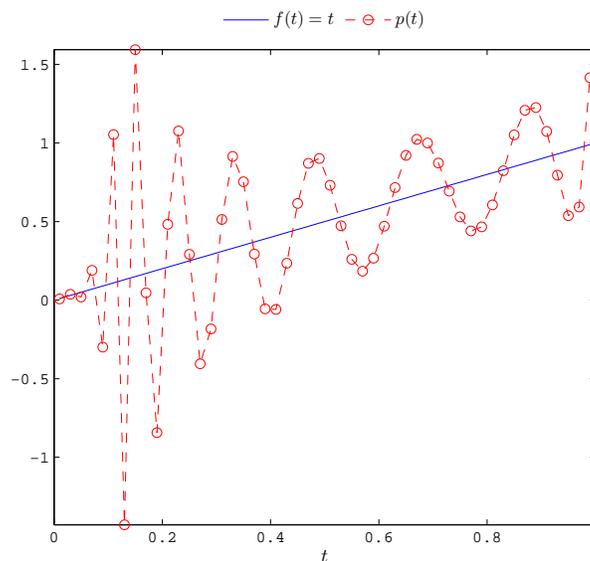


**Figura 2.14:** Gráfica en escala logarítmica de los valores singulares  $\sigma_i$  de la matriz  $A_{50 \times 50}$  de la discretización de la ecuación (2.9) contra el subíndice  $i$  junto con la sucesión  $\{e^{-i/2}\}_{i=1}^{50}$ .

En la Figura 2.14 graficamos los valores singulares  $\sigma_i$  de  $A$  contra su subíndice  $i$  en escala logarítmica base 10 sobre el eje vertical, junto con la sucesión  $\{e^{-i/2}\}$ . Como usamos una escala logarítmica, observamos que los valores  $e^{-i/2}$  aparecen en una recta. Los valores singulares decaen rápidamente. De hecho,

$$\sigma_i < e^{-i/2}, \quad i = 2, \dots, 50.$$

Por eso el problema discreto ésta severamente mal planteado.



**Figura 2.15:** Gráficas de la solución  $f(t) = t$  de la Ecuación (2.9) y de la poligonal  $p$  que une los puntos  $(t_1, x_1), \dots, (t_{50}, x_{50})$ .

En la Figura 2.15 mostramos los valores de la solución

$$f(t) = t, \quad 0 \leq t \leq 1$$

de la ecuación (2.9) junto con los valores  $x_i$  de la solución de cuadrados mínimos de norma mínima de la ecuación  $A\mathbf{x} = \mathbf{b}$ . La poligonal  $p$  que une los puntos  $(t_1, x_1), \dots, (t_{50}, x_{50})$  oscila en comparación de la recta que nos da la solución. Esto se debe al mal condicionamiento de  $A$ .

## 2.6. Expansión en Valores Singulares

Los problemas mal planteados en el sentido de Hadamard que hemos abordado pueden formularse de la siguiente manera:

Dada una función  $g \in L^2[a, b]$  y el operador  $T : L^2[a, b] \rightarrow L^2[a, b]$  definido por

$$T[f](y) = \int_a^b k(x, y) f(y) dy,$$

hallar una función  $f \in L^2[a, b]$  tal que  $T[f] = g$

Vamos a examinar la relación entre estos y los problemas discretos mal planteados.

La solución  $f$  de la ecuación  $T[f] = g$  debe ser cuadrado integrable. Damos una condición necesaria para que  $\|f\|_2 < \infty$ . La idea es expandir  $f$  y  $g$  en bases de funciones, pues si hay una relación entre los coeficientes de la función  $f$  con los de la función  $g$ , entonces asociamos la existencia de la función  $f$  con la información sobre  $g$ . Para ello, generalizamos la SVD al espacio vectorial de dimensión infinita  $L^2[a, b]$ .

Consideramos el operador  $T^* : L^2[a, b] \rightarrow L^2[a, b]$  dado por

$$T^*[g](y) = \int_a^b k(x, y)g(x)dx.$$

La composición de  $T^*$  con  $T$  nos da el operador  $S = T^* \circ T$  sobre  $L^2[a, b]$  dado por

$$S[f](z) = \int_a^b \left[ \int_a^b k(x, z)k(x, y)dy \right] f(x)dx.$$

Supongamos que  $k$  es cuadrado integrable. Entonces  $T, T^*$  y  $S$  son operadores compactos.

**Observaciones 2.11:** [72]

 Para operadores compactos, los valores propios se definen de la misma manera que en dimensión finita, esto es,  $\lambda \in \mathbb{C}$  es un valor propio del operador  $S$  si existe  $f \in L^2[a, b]$  tal que  $S[f] = \lambda f$ . La función  $f$  se llama *función propia*.

 Los valores propios de un operador compacto forman una sucesión en  $\mathbb{C}$  que se acumula en cero.

 Con el producto interno sobre  $L^2[a, b]$ ,  $S$  satisface la identidad

$$\langle S[f], g \rangle = \langle f, S[g] \rangle \quad \forall f, g \in L^2[a, b].$$

En la literatura, los operadores lineales y acotados sobre  $L^2[a, b]$  que cumplen esta identidad se llaman *auto-adjuntos*.

 Los valores propios de un operador auto-adjunto son reales no negativos y las funciones propias asociadas a distintos valores propios son ortogonales

Sea  $\{\lambda_n\}$  la sucesión de valores propios positivos de  $S$  en orden decreciente. El *Teorema Espectral* para operadores compactos auto-adjuntos [72] nos dice que  $S$  tiene una sucesión ortonormal  $\{v_n\}$  de funciones propias asociadas a la sucesión  $\{\lambda_n\}$  tal que

$$S[f] = \sum_{n=1}^{\infty} \langle f, v_n \rangle v_n \quad \forall f \in L^2[a, b].$$

Sean

$$\mu_n = \sqrt{\lambda_n} \quad \text{y} \quad u_n = \frac{1}{\mu_n} T[v_n] \quad \forall n \in \mathbb{N}.$$

Entonces

$$T[v_n] = \mu_n u_n \quad \forall n \in \mathbb{N}. \quad (2.10)$$

**Observaciones 2.12:**

☞ La sucesión de funciones ortonormales  $\{v_n\}$  forma una base ortonormal para  $\text{Ker}(T)^\perp$ .

☞  $\{u_n\}$  es una sucesión de funciones ortonormales de  $L^2[a, b]$ .

Sea  $f \in L^2[a, b]$ . Denotemos por  $P_f$  a la proyección ortogonal de  $f$  sobre  $\text{Ker}(T)$ . Como

$$f - P_f \in \text{Ker}(T)^\perp,$$

entonces

$$f = P_f + \sum_{n=1}^{\infty} \langle f, v_n \rangle v_n. \quad (2.11)$$

Puesto que  $P_f \in \text{Ker}(T)$ , se sigue que

$$\langle P_f, v_n \rangle = 0 \quad \forall n \in \mathbb{N}.$$

**Observaciones 2.13:**

☞ Dado que el núcleo del operador integral  $T$  es cuadrado integrable, tenemos que

$$T \left[ \sum_{n=1}^{\infty} \langle f, v_n \rangle v_n \right] = \sum_{n=1}^{\infty} \langle f, v_n \rangle T[v_n].$$

La observación 2.13 implica que la imagen de la expansión (2.11) bajo  $T$  es

$$T[f] = \sum_{n=1}^{\infty} \langle f, v_n \rangle T[v_n]. \quad (2.12)$$

En consecuencia, mediante la relación (2.10) podemos rescribir la expansión (2.12) como

$$T[f] = \sum_{n=1}^{\infty} \mu_n \langle f, v_n \rangle u_n \quad \forall f \in L^2[a, b]. \quad (2.13)$$

Esta representación de  $T$  se llama **expansión en valores singulares (SVE)**. Los números positivos  $\mu_n$  se conocen como **valores singulares**, las funciones ortonormales  $v_n$  y  $u_n$  son **funciones singulares** de izquierda y derecha, respectivamente [41]. Esta expansión generaliza la SVD.

**Observaciones 2.14:**

☞ Dado que el núcleo  $k$  de  $T$  es cuadrado integrable, entonces mediante la SVE de  $T$  podemos expandir  $k$  como [54] [110]:

$$k(x, y) = \sum_{n=1}^{\infty} \mu_n u_n(x) v_n(y). \quad (2.14)$$

### 2.6.1. SVE y SVD

Veamos la relación entre la SVE y la SVD. Lo que hacemos es discretizar la ecuación integral (2.1) con el método de Galerkin. Recordemos que escogemos dos conjuntos  $\{\varphi_i\}_{i=1}^n$  y  $\{\psi_i\}_{i=1}^n$  de funciones ortonormales en  $L^2[a, b]$ . Con estas funciones, obtenemos una matriz  $A_{n \times n}$  y el lado derecho  $\mathbf{b}$  dados por

$$\begin{aligned} a_{i,j} &= \int_a^b \int_a^b k(x, y) \varphi_j(y) \psi_i(x) dx dy, \\ b_i &= \int_a^b g(x) \psi_i(x) dx, \end{aligned} \quad i, j = 1, \dots, n.$$

Supongamos que  $A$  es invertible. Relacionamos sus valores singulares con los del operador  $T$ . Para ello comparamos el tamaño del núcleo  $k$  con el de la matriz  $A$  en términos de sus valores singulares.

Debido a que  $u_n$  y  $v_n$  son ortonormales, a partir de la expansión (2.14) de  $k$ , tenemos

$$\|k\|_2^2 = \sum_{i=1}^{\infty} \mu_i^2.$$

Por otra parte, sean  $\sigma_1^{(n)}, \dots, \sigma_n^{(n)}$  los valores singulares de  $A$ . Entonces

$$\|A\|_F^2 = \sum_{i=1}^n \left[ \sigma_i^{(n)} \right]^2.$$

Con estas expresiones, podemos verificar las siguientes desigualdades [54]:

$$\begin{aligned} 0 \leq \mu_i - \sigma_i^{(n)} &\leq \|k\|_2^2 - \|A\|_F^2, \\ \sigma_i^{(n)} \leq \sigma_i^{(n+1)} &\leq \mu_i, \end{aligned} \quad i = 1, \dots, n.$$

En consecuencia, los valores singulares de  $A$  se aproximan a los de  $T$  conforme el orden de la matriz aumenta y su discrepancia está acotada por la diferencia de los tamaños del núcleo y la matriz.

Ahora, consideramos la ecuación lineal  $T[f] = g$  y su discretización  $A\mathbf{x} = \mathbf{b}$ . Además de comparar los valores singulares de  $A$  y  $T$ , relacionamos los coeficientes  $\langle u_j, g \rangle$  de  $g$  con los de  $\mathbf{b}$  en base de vectores singulares, a saber,  $\mathbf{u}_j^T \mathbf{b}$ .

Puesto que las funciones  $\psi_1, \dots, \psi_n$  son ortonormales, la proyección de  $g$  sobre el subespacio generado por éstas es

$$g^{(n)} = \sum_{i=1}^n b_i \psi_i.$$

A partir de las componentes  $u_{i,j}$  y  $v_{i,j}$  de las matrices  $U$  y  $V$ , definimos las funciones

$$u_j^{(n)} = \sum_{i=1}^n u_{i,j} \psi_i \quad \text{y} \quad v_j^{(n)} = \sum_{i=1}^n v_{i,j} \varphi_i, \quad j = 1, \dots, n.$$

Éstas satisfacen la identidad

$$\langle u_j^{(n)}, g^{(n)} \rangle = \mathbf{u}_j^T \mathbf{b}, \quad j = 1, \dots, n.$$

Por lo que

$$\left| \langle u_j, g \rangle - \langle u_j^{(n)}, g^{(n)} \rangle \right| \rightarrow 0 \quad \text{cuando } n \rightarrow \infty,$$

implica que los coeficientes  $\mathbf{u}_j^T \mathbf{b}$  de  $\mathbf{b}$  aproximan a los coeficientes  $\langle u_j, g \rangle$  de  $g$ .

De esta manera, la SVD de  $A$  en dimensión finita puede pensarse como una aproximación de la SVE de  $T$  en dimensión infinita [48].

### 2.6.2. Condición de Picard

Vamos a usar la SVE para dar una condición necesaria para que la ecuación  $T[f] = g$  tenga solución. Supongamos que  $g$  está en la imagen de  $T$ . Entonces con la SVE (2.13) tenemos

$$g = \sum_{n=1}^{\infty} \mu_n \langle f, v_n \rangle u_n.$$

Dado que las funciones singulares  $u_n$  son ortonormales, se sigue que  $\langle g, u_n \rangle = \mu_n \langle f, v_n \rangle$  para cada  $n \in \mathbb{N}$ . Luego, la expansión (2.11) de  $f$  se reescribe como

$$f = P_f + \sum_{n=1}^{\infty} \frac{\langle g, u_n \rangle}{\mu_n} u_n, \quad (2.15)$$

más aún, como la sucesión  $\{v_n\}$  es base ortonormal de  $\text{Ker}(T)^\perp$  y  $P_f \in \text{Ker}(T)$ , tenemos

$$\|f\|_2^2 = \|P_f\|_2^2 + \sum_{n=1}^{\infty} \left( \frac{\langle g, u_n \rangle}{\mu_n} \right)^2.$$

Esta identidad nos da una condición necesaria para la solución  $f$ .

**Condición de Picard.** Sea  $\{\mu_n\}$  la sucesión de valores singulares del operador compacto  $T : L^2[a, b] \rightarrow L^2[a, b]$  dado por

$$T[f](x) = \int_a^b k(x, y) f(y) dy,$$

y sea  $\{u_n\}$  su sucesión de funciones singulares de izquierda. Entonces la ecuación  $T[f] = g$  tiene solución dada por (2.15) si  $g \in \text{Im}(T)$  y

$$\sum_{n=1}^{\infty} \left( \frac{\langle g, u_n \rangle}{\mu_n} \right)^2 < \infty.$$

Esta condición nos dice que si los coeficientes de  $g$  en la base ortonormal  $\{u_n\}$  decaen a cero más rápido que los valores singulares de  $T$ , entonces la ecuación integral (2.1) tiene solución [21].

## 2.7. Condición Discreta de Picard

Queremos saber si un problema mal planteado dado por la ecuación (2.1) satisface la condición de Picard cuando solamente tenemos un número finito de observaciones. Para ello, damos una versión discreta de la condición de Picard discretos mal planteados. De modo que si ésta se cumple, el problema continuo la satisface.

El operador  $T$  tiene un número infinito de valores singulares. Por lo que necesitamos que la serie

$$\sum_{i=1}^{\infty} \left( \frac{\langle g, u_i \rangle}{\mu_i} \right)^2$$

converja. En cambio, en el problema discreto la matriz  $A$  con  $r$  valores singulares positivos, el tamaño de la solución en términos de la SVD es

$$\sum_{i=1}^r \left( \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \right)^2.$$

Esta suma corresponde con la serie anterior hasta el  $r$ -ésimo término cuando discretizamos por Galerkin.

En vez de dar una condición sobre el tamaño de la solución de la ecuación  $A\mathbf{x} = \mathbf{b}$ , vemos si los coeficientes  $\mathbf{u}_i^T \mathbf{b}$  decaen más rápido que valores singulares  $\sigma_i$ , pues esto corresponde a que los cocientes  $\langle g, u_n \rangle$  decaigan más rápido que los valores singulares  $\mu_i$ . Al respecto, Hansen [54] propone lo siguiente:

Sea  $\tau > 0$  un nivel de perturbación dado. Decimos que el problema dado por la ecuación  $A\mathbf{x} = \mathbf{b}$  cumple con la **Condición Discreta de Picard (DPC)** si para todos los valores singulares  $\sigma_i$  de  $A$  mayores que  $\tau$ , los coeficientes correspondientes  $|\mathbf{u}_i^T \mathbf{b}|$  de  $\mathbf{b}$ , en promedio, decaen a cero más rápido que  $\sigma_i$ .

Podemos inspeccionar visualmente si la DPC se cumple con ayuda de la **gráfica de Picard**, que consiste en una gráfica en escala logarítmica base 10 sobre el eje vertical que muestra los valores singulares  $\sigma_i$  de  $A$  en orden decreciente contra su subíndice  $i$ , junto con los valores absolutos de los coeficientes  $\mathbf{u}_i^T \mathbf{b}$  de  $\mathbf{b}$  y los valores absolutos de los cocientes  $(\mathbf{u}_i^T \mathbf{b})/\sigma_i$ .

### Observaciones 2.15:

 Los cocientes  $(\mathbf{u}_i^T \mathbf{b})/\sigma_i$  son los coeficientes de la solución de cuadrados mínimos de norma mínima  $\mathbf{x}^\dagger$  de la ecuación  $A\mathbf{x} = \mathbf{b}$  en la base de vectores singulares de derecha.

Para corroborar numéricamente la DPC, promediamos los coeficientes  $\mathbf{u}_i^T \mathbf{b}$  con la media geométrica móvil

$$\rho_i = \left( \prod_{j=i-q}^{i+q} |\mathbf{u}_j^T \mathbf{b}| \right)^{\frac{1}{2q+1}}, \quad i = q + 1, \dots, n - q$$

para  $q = \{0, 1, 2, 3\}$ . La DPC se satisface si el cociente  $\rho_i/\sigma_i$  decae monótonamente para aquellos valores singulares  $\sigma_i$  por arriba del umbral  $\tau$ .

Primero damos el ejemplo de un problema discreto mal planteado que NO cumple la DPC. Usamos el ejemplo de deconvolución del capítulo 1.

**Ejemplo 2.10.** El problema discreto de deconvolución en el Ejemplo 1.2 consiste en resolver el sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b}$ , donde el vector de observaciones  $\mathbf{b}$  está libre de ruido y  $A$  es positiva definida dada por

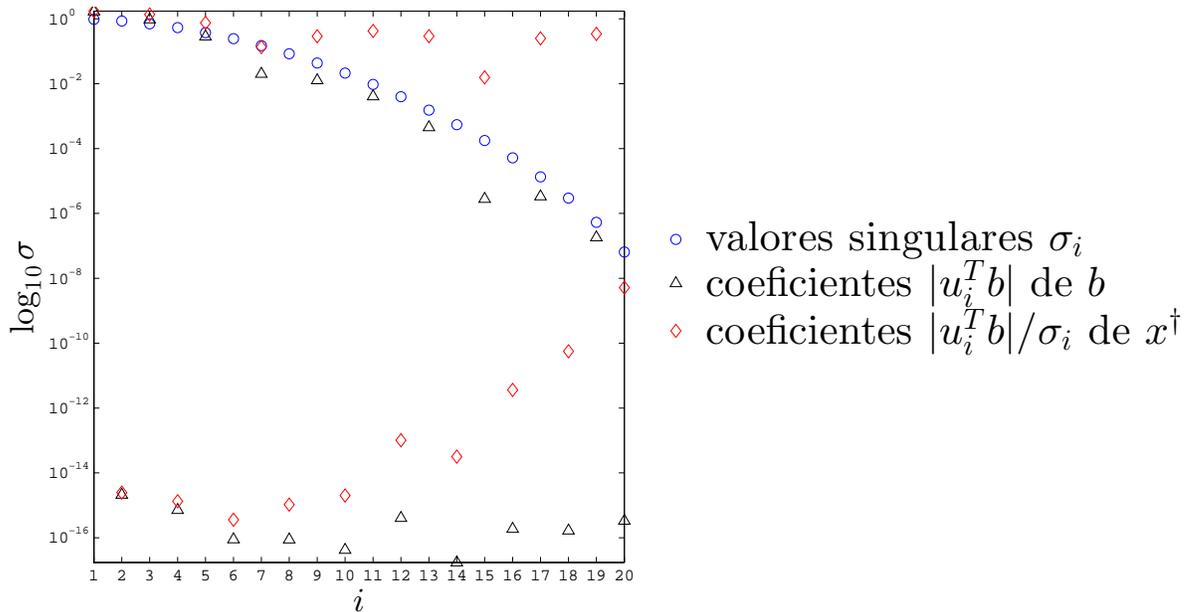
$$A = \begin{bmatrix} k(m_1 - m_1) & \cdots & k(m_1 - m_{20}) \\ \vdots & & \vdots \\ k(m_{20} - m_1) & \cdots & k(m_{20} - m_{20}) \end{bmatrix},$$

$$k(t) = \frac{1}{0.1\sqrt{2\pi}} \exp\left(-\frac{t^2}{0.02}\right), \quad m_j = \frac{2j-1}{40}.$$

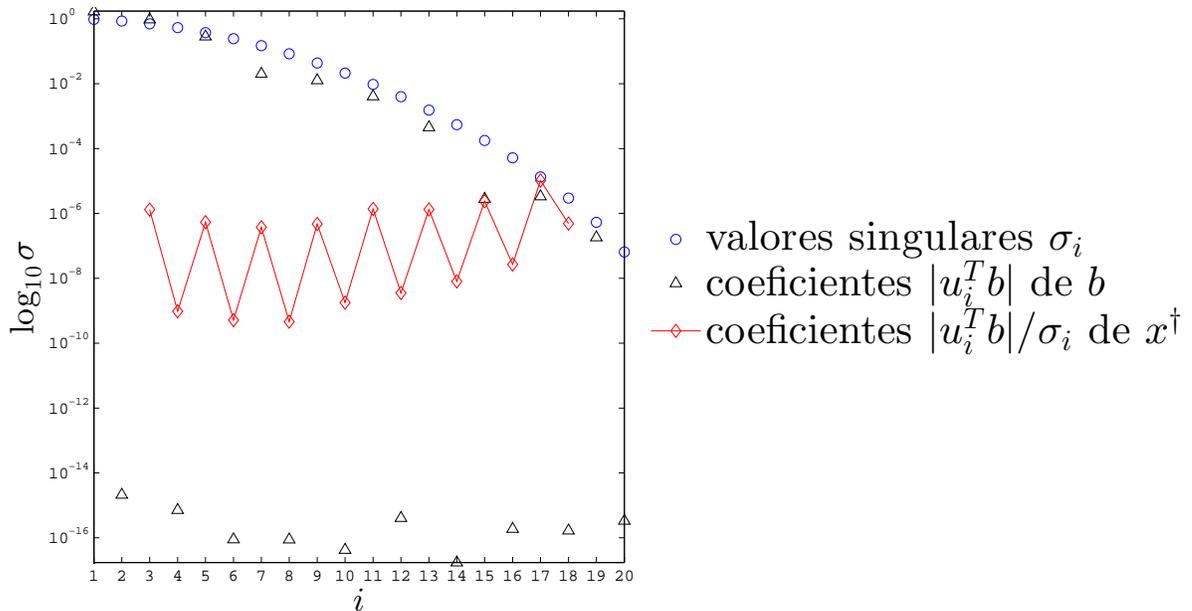
En el Ejemplo 2.8 vimos que este problema está moderadamente mal planteado. Ahora, queremos saber si cumple con la DPC.

En la Figura 2.16 mostramos la gráfica de Picard del problema discreto de deconvolución. Los 20 valores singulares  $\sigma_i$  de  $A$ , marcados por  $(\circ)$  y ordenados de manera decreciente, decaen gradualmente de 0.961 hasta  $6.464 \times 10^{-8}$ . Por otra parte, los valores absolutos de los coeficientes  $\mathbf{u}_i^T \mathbf{b}$  de  $\mathbf{b}$  en la base de vectores singulares de derecha, marcados por  $(\triangle)$ , se mantienen entre  $1.829 \times 10^{-7}$  y 1.6739 para subíndice impar, mientras que los de índice par se mantienen entre  $1.734 \times 10^{-17}$  y  $2.118 \times 10^{-15}$ . Luego, los valores absolutos de los coeficientes  $\mathbf{u}_i^T \mathbf{b}/\sigma_i$  de la solución  $\mathbf{x}^\dagger$  de la ecuación  $A\mathbf{x} = \mathbf{b}$  en la base de vectores singulares de derecha, marcados por  $(\diamond)$ , se mantienen entre 0.016 y 1.74 en los subíndices impares, y aumentan de  $3.605 \times 10^{-16}$  hasta  $5.152 \times 10^{-9}$  en subíndices pares.

Como los coeficientes  $\mathbf{u}_i^T \mathbf{b}/\sigma_i$  de  $\mathbf{x}^\dagger$  oscilan entre índice par e impar, entonces la poligonal que une su media geométrica móvil  $\rho_i$  con  $q = 2$  se comporta en zigzag como se muestra en la Gráfica de Picard 2.17. Esto nos sugiere que el problema de deconvolución del Ejemplo 1.2 no cumple con la Condición de Picard.



**Figura 2.16:** Gráfica de Picard para el problema discreto de deconvolución  $Ax = b$  del Ejemplo 1.2. Mostramos en escala logarítmica a los valores singulares de  $A$  en orden decreciente contra su subíndice  $i$ , así como los coeficientes del lado derecho  $b$  y de la solución  $x^\dagger$  en la base de vectores singulares de derecha de  $A$ .



**Figura 2.17:** Gráfica de Picard para el problema discreto de deconvolución  $Ax = b$  del Ejemplo 1.2 con la poligonal, marcada por (◇), que une las medias geométricas móviles de los coeficientes  $|u_i^T b|/\sigma_i$  de la solución  $x^\dagger$ .

Ahora, damos un ejemplo que cumple la DPC. Retomamos el Ejemplo 1.3. El problema directo es obtener la función del caudal mediante la convolución de  $k(y) = 16\sqrt{y}$  con la función que nos da la forma del vertedero. El problema inverso es una deconvolución.

Cuando discretizamos el problema inverso, obtenemos un problema discreto mal planteado. Nos interesa ver que ese problema cumple la DPC con observaciones libres de ruido. Vemos el comportamiento de los valores absolutos de los coeficientes de la solución de norma mínima del problema inverso mal condicionado en la base de vectores singulares de izquierda.

Así, si el problema discreto cumple la DPC, podemos pensar que el problema continuo de la forma del vertedero satisface la condición de Picard.

**Ejemplo 2.11.** Cuando examinamos la sensibilidad de la solución del problema discreto de la forma del vertedero en el Ejemplo 1.3, formamos una ecuación  $A\mathbf{x} = \mathbf{b}$ , donde  $A$  es una matriz triangular inferior dada por

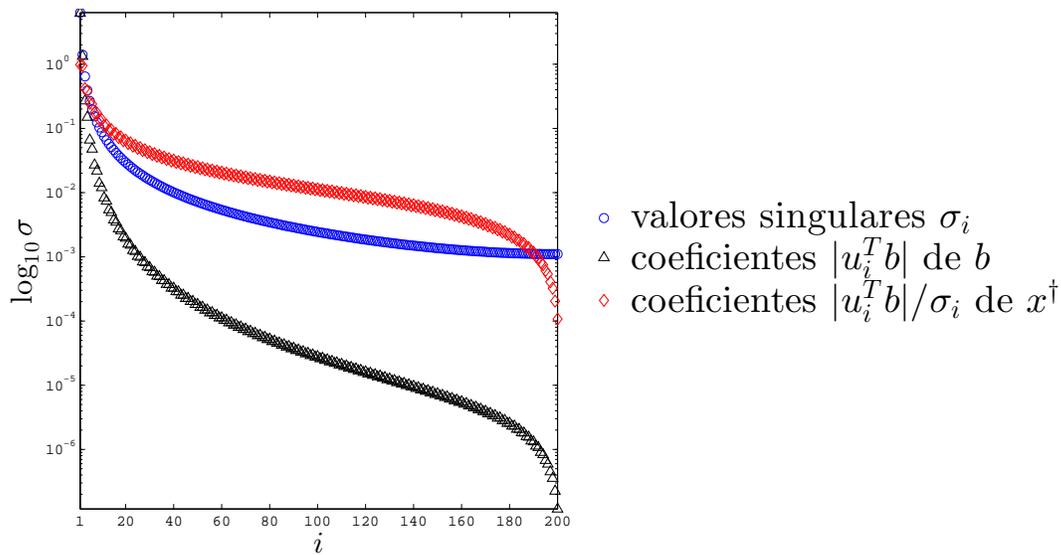
$$a_{i,j} = \begin{cases} \frac{16}{200^{3/2}} \sqrt{i - j + \frac{1}{2}}, & \text{si } i \geq j, \\ 0 & \text{en otro caso,} \end{cases} \quad i, j = 1, \dots, 200.$$

Queremos ver si este problema discreto satisface la DPC.

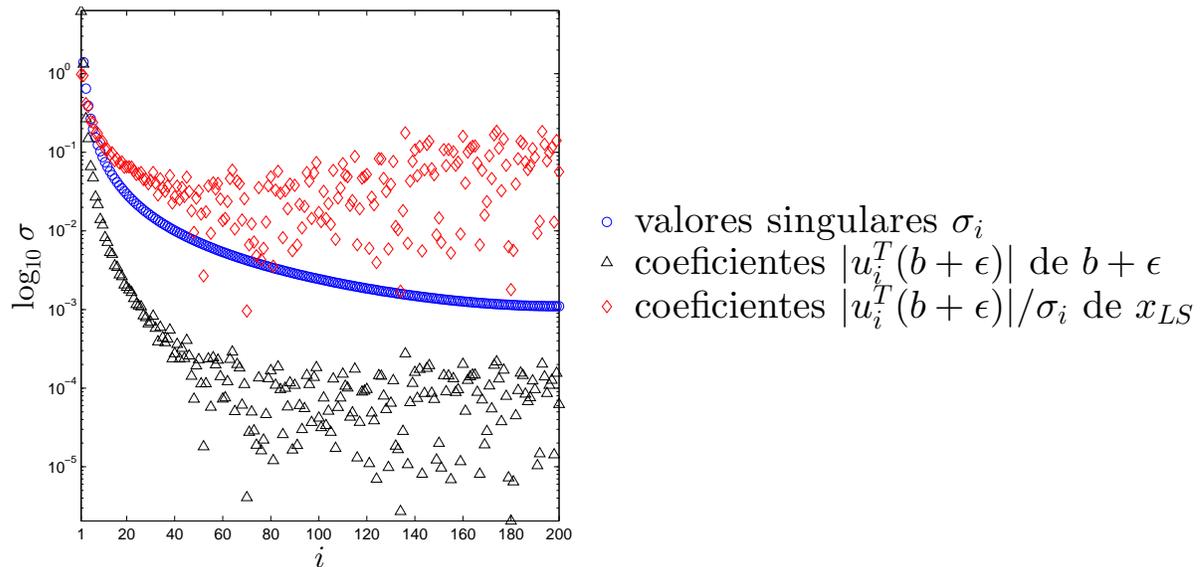
Primero, consideramos las observaciones libres de ruido  $b_i = z_i^2$  en los 200 puntos de la malla uniforme del intervalo  $[0, 1]$ . En la Figura 2.18 mostramos la gráfica de Picard correspondiente. Los 200 valores singulares  $\sigma_i$  de  $A$ , marcados por  $(\circ)$  y ordenados de manera decreciente, decaen gradualmente de 6.344 a  $1.101 \times 10^{-3}$ . Por lo que el problema discreto está ligeramente mal planteado. Observamos en la Gráfica de Picard 2.18 que los valores absolutos de los coeficientes  $\mathbf{u}_i^T \mathbf{b}$  del lado derecho  $\mathbf{b}$  en base de vectores singulares de derecha, marcados por  $(\triangle)$ , decaen más rápido que sus respectivos valores singulares. Por eso los valores absolutos de los coeficientes  $\mathbf{u}_i^T \mathbf{b} / \sigma_i$  de la solución  $\mathbf{x}^\dagger$  de la ecuación  $A\mathbf{x} = \mathbf{b}$ , marcados por  $(\diamond)$ , se muestran como una sucesión decreciente. Así que el problema discreto mal planteado de la forma del vertedero cumple con la DPC.

Ahora, supongamos que el vector  $\mathbf{b}$  tiene ruido aditivo  $\boldsymbol{\epsilon}$  idénticamente distribuido bajo una gaussiana de media cero y de varianza 0.01. Esta vez en la Gráfica de Picard tenemos que los valores absolutos de los coeficientes  $\mathbf{u}_i^T (\mathbf{b} + \boldsymbol{\epsilon})$  de  $\mathbf{b} + \boldsymbol{\epsilon}$  decrecen de  $i = 1$  hasta  $i = 29$ , y después se dispersan como se ve en la Figura 2.19. Sea  $\mathbf{x}_{LS}$  el estimador de cuadrados mínimos para  $\mathbf{x}^\dagger$ . Entonces los valores absolutos de los coeficientes  $\mathbf{u}_i^T (\mathbf{b} + \boldsymbol{\epsilon}) / \sigma_i$  de  $\mathbf{x}_{LS}$  en la base de vectores singulares de derecha decrecen inicialmente hasta dispersarse a partir de  $i = 29$ .

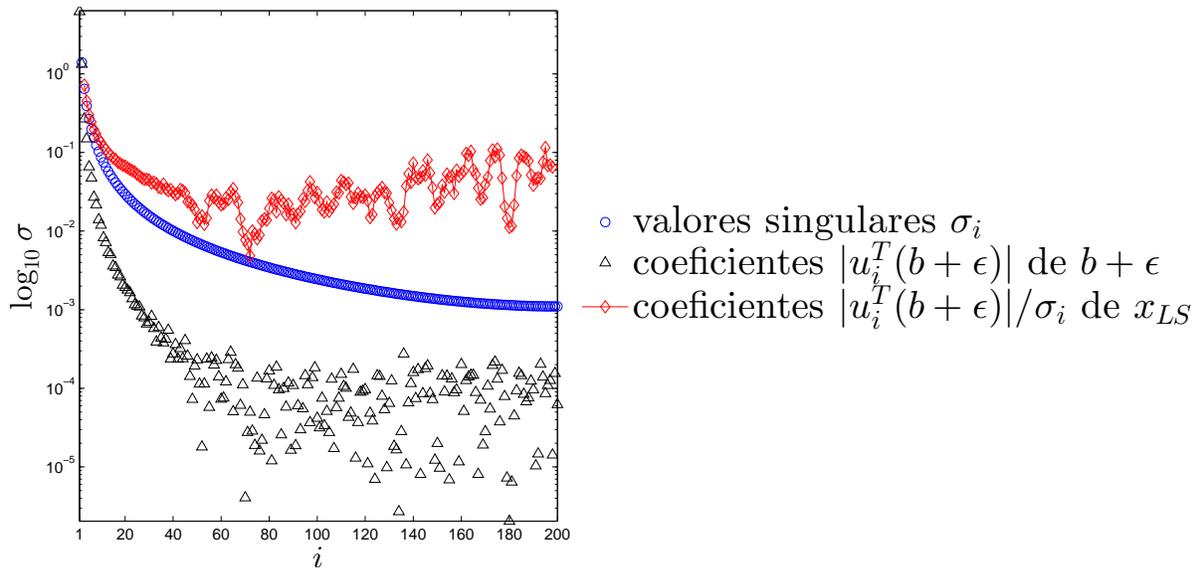
En la Figura 2.20 mostramos otra Gráfica de Picard del problema discreto con observaciones con ruido. En esta ocasión, usamos la media geométrica móvil  $\rho_i$  con  $q = 2$  para concentrar la dispersión de los coeficientes  $|\mathbf{u}_i^T (\mathbf{b} + \boldsymbol{\epsilon})| / \sigma_i$  de  $\mathbf{x}_{LS}$  en una poligonal, marcada con  $(\diamond)$ , que decrece desde el subíndice  $i = 3$  hasta  $i = 31$  y posteriormente oscila.



**Figura 2.18:** Gráfica de Picard para el problema discreto  $Ax = b$  de la forma del vertedero del Ejemplo 1.3. Mostramos en escala logarítmica a los valores singulares de  $A$  en orden decreciente contra su subíndice  $i$ , así como los valores absolutos de los coeficientes del lado derecho  $b$  y de la solución  $x^\dagger$  en la base de vectores singulares de derecha de  $A$ .



**Figura 2.19:** Gráfica de Picard para el problema discreto  $Ax = b + \epsilon$  de la forma del vertedero del Ejemplo 1.3. Esta vez mostramos en escala logarítmica a los valores absolutos de coeficientes de  $b + \epsilon$  y del estimador de cuadrados mínimos  $x_{LS}$  de  $x^\dagger$  en la base de vectores singulares de derecha de  $A$ .



**Figura 2.20:** Gráfica de Picard para el problema discreto  $A\mathbf{x} = \mathbf{b} + \epsilon$  de la forma del vertedero del Ejemplo 1.3. Mostramos en escala logarítmica a la poligonal, marcada por (◇), con las medias geométricas móviles de los coeficientes  $|u_i^T(\mathbf{b} + \epsilon)|/\sigma_i$  del estimador  $\mathbf{x}_{LS}$  de  $\mathbf{x}^\dagger$  en la base de vectores singulares de derecha de  $A$ .

Así, el problema discreto mal planteado cumple parcialmente la DPC para los 29 valores singulares más grandes cuando las observaciones tienen ruido.

Estamos interesados en resolver problemas lineales mal planteados, su discretización da lugar a un sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b}$ , donde la matriz  $A$  está mal condicionada. En consecuencia, los errores de truncamiento en la discretización y los errores por redondeo que acumulamos en los cálculos numéricos ocasionan que la solución calculada tenga altas oscilaciones

Queremos conseguir una buena aproximación de la solución del problema discreto mal planteado. Sin embargo, los ejemplos del Capítulo 1 muestran las dificultades que se presentan en la resolución de estos problemas. En vez de tratar directamente con el problema, lo que hacemos es modificarlo adecuadamente.

Tikhonov fue el primero en sugerir un enfoque nuevo para resolver problemas mal planteados. Antes, estos se resolvían usando solamente información adicional para aislar una clase de soluciones admisibles. Esta formulación modifica al problema mal planteado y lo reemplaza por otro bien planteado. Partiendo del hecho de que además de los datos aproximados, conocemos la precisión del error en los datos, en 1963 [113] [114], él establece y justifica el método más popular para resolver problemas mal planteados, llamado regularización [61].

De manera independiente, el método de regularización para resolver problemas mal planteados fue introducido en América por Phillips [97] en 1962. Él introduce un término de suavizamiento en la solución que resulta de la inversión del sistema de ecuaciones obtenido por regla de cuadratura. Un año más tarde, Twomey [117] extendió el trabajo de Phillips e introduce otro término de suavizamiento.

La idea de la *regularización* es reemplazar el problema mal planteado por otro bien planteado de modo que la solución del nuevo problema sea una buena aproximación de la solución buscada.

Escogemos una colección de problemas bien planteados, elegimos uno de esta colección, y calculamos su solución. Los problemas de la colección dependen de un parámetro llamado *parámetro de regularización*. Por cada valor de éste, tenemos un problema distinto, su solución se llama *solución regularizadora*.

Una vez que elegimos el valor del parámetro de regularización, la solución regularizadora se toma como aproximación de la solución del problema mal planteado. En resumen, en la regularización realizamos las siguientes tareas:

- \* Calcular soluciones regularizadoras en función del parámetro de regularización.
- \* Elegir el parámetro de regularización adecuado.

**Observaciones 3.1:**

- ☞ La familia de problemas puede ser finita, o bien depender de un parámetro continuo, o de uno discreto.
- ☞ En la práctica, un problema bien planteado puede estar mal condicionado. Así que en la regularización, lo reemplazamos por otro que además de estar bien planteado, no está tan mal condicionado.
- ☞ Además de los errores de truncamiento y por redondeo, los problemas tienen observaciones con ruido. En ese caso buscamos un estimador de la solución del problema mal planteado.

**3.1. Introducción a la Regularización mediante SVD**

Para regularizar problemas discretos mal planteados usamos la SVD como nuestro punto de partida. Sean  $\sigma_1 \geq \dots \geq \sigma_r$  los valores singulares positivos de la matriz  $A$  de tamaño  $m \times n$  y rango  $r$ . Sean  $\mathbf{u}_1, \dots, \mathbf{u}_r$  y  $\mathbf{v}_1, \dots, \mathbf{v}_r$  sus vectores singulares de izquierda y de derecha, respectivamente. Como los valores singulares pequeños de  $A$  son responsables del mal condicionamiento, lo primero que intentamos es removerlos.

**3.1.1. SVD Truncada**

Mediante la SVD, tenemos que  $A$  es la suma de matrices de rango uno

$$A = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

Al remover los valores singulares más pequeños, truncamos esta suma. Dado  $k \in \{1, \dots, r\}$ , quitamos los  $r - k$  valores singulares más pequeños. Obtenemos la matriz

$$A_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

**Observaciones 3.2:**

- ☞ Recordamos que los Teoremas 2.2 y 2.3 nos dicen que  $A_k$  es la matriz de rango  $k$  más cercana a  $A$  con la norma espectral y la norma Frobenius para  $k = 1, \dots, r$ .

Si reemplazamos  $A$  por  $A_k$  en el Problema de Cuadrados Mínimos, entonces tenemos la siguiente colección de problemas:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|A_k \mathbf{x} - \mathbf{b}\|_2, \quad k = 1, \dots, r. \quad (3.1)$$

En cada caso, la solución de norma mínima que obtenemos es

$$\mathbf{x}_k = \sum_{i=1}^k \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i, \quad k = 1, \dots, r$$

Este método se conoce como **SVD truncada**. La matriz  $A_k$  se llama matriz TSVD de  $A$  con nivel de truncamiento  $k$  [51],[56].

### Observaciones 3.3:

☞ La deducción de  $\mathbf{x}_k$  es análoga a la de la solución de cuadrados mínimos de norma mínima  $\mathbf{x}_{\text{LS}}$ . De hecho, cuando  $k = r$ , tenemos que  $A_k = A$  y  $\mathbf{x}_k = \mathbf{x}_{\text{LS}}$ .

☞ Los Problemas (3.1) están mejor condicionados que el problema de cuadrado mínimos (2.6), pues con los valores singulares de  $A$  en orden decreciente tenemos que

$$\kappa_2(A_k) = \frac{\sigma_1}{\sigma_k} \leq \frac{\sigma_1}{\sigma_r} = \kappa_2(A)$$

La idea de la regularización por SVD truncada es quedarse con los coeficientes  $\mathbf{u}_i^T \mathbf{b}$  de  $\mathbf{b}$  asociados a los valores singulares más grandes. En este caso, el parámetro de regularización es el nivel de truncamiento  $k$  y la solución regularizada es  $\mathbf{x}_k$ . En [47], Hansen justifica el uso de la SVD truncada como método de regularización para matrices ligeramente mal condicionadas.

**Ejemplo 3.1.** En el problema de la reconstrucción del haz de luz (Ejemplo 2.5) debemos resolver la ecuación integral

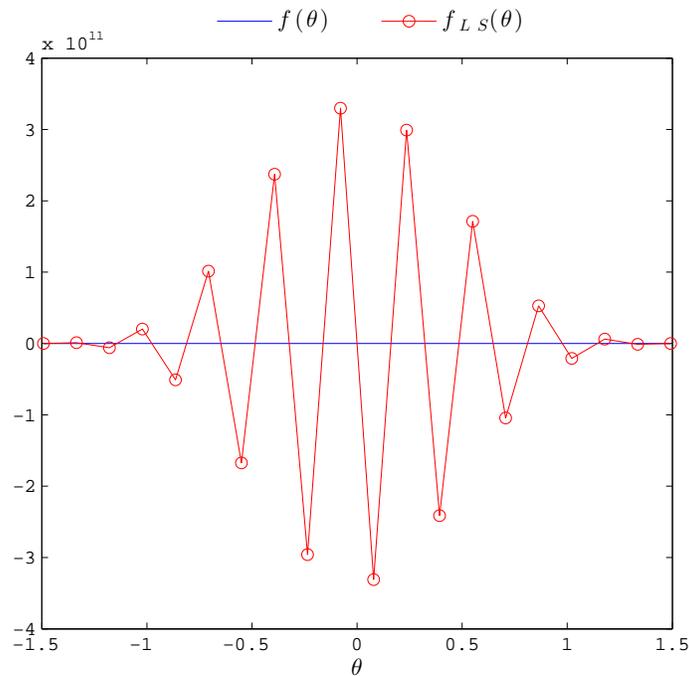
$$\int_{-\pi/2}^{\pi/2} (\cos \phi + \cos \theta)^2 \left( \frac{\sin(\pi(\sin \phi + \sin \theta))}{\pi(\sin \phi + \sin \theta)} \right)^2 f(\theta) d\theta = g(\phi), \quad |\phi| \leq \frac{\pi}{2}. \quad (3.2)$$

Usamos el método de colocación en 20 ángulos  $\phi_i \in [-\pi/2, \pi/2]$  con cuadratura compuesta del punto medio para obtener valores aproximados de la función  $f : [-\pi/2, \pi/2] \rightarrow \mathbb{R}$ . De esa manera, el problema discreto es resolver la ecuación  $A\mathbf{x} = \mathbf{b}$ , donde

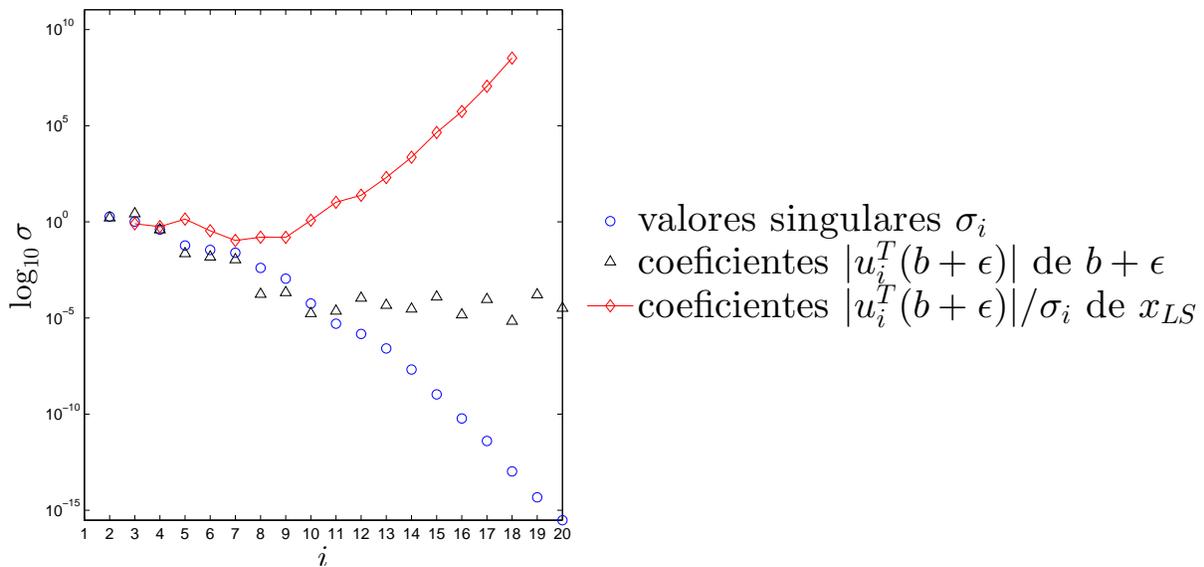
$$a_{i,j} = \frac{\pi}{20} (\cos \phi_i + \cos \phi_j)^2 \left( \frac{\sin(\pi(\sin \phi_i + \sin \phi_j))}{\pi(\sin \phi_i + \sin \phi_j)} \right)^2, \quad b_i = g(\phi_i), \quad i, j = 1, \dots, 20$$

Anteriormente ya calculamos la solución  $\mathbf{x}^\dagger$  de cuadrados mínimos con norma mínima de la ecuación  $A\mathbf{x} = \mathbf{b}$ , con las 20 observaciones  $b_i$  de la Tabla 2.3. Denotamos a su  $i$ -ésima componente por  $\mathbf{x}^\dagger(i)$ . Para aproximar los valores de  $f$ , usamos la poligonal que en  $\phi_i$  tiene el valor  $\mathbf{x}^\dagger(i)$ . En la Figura 2.9 vimos la gráfica de esta aproximación.

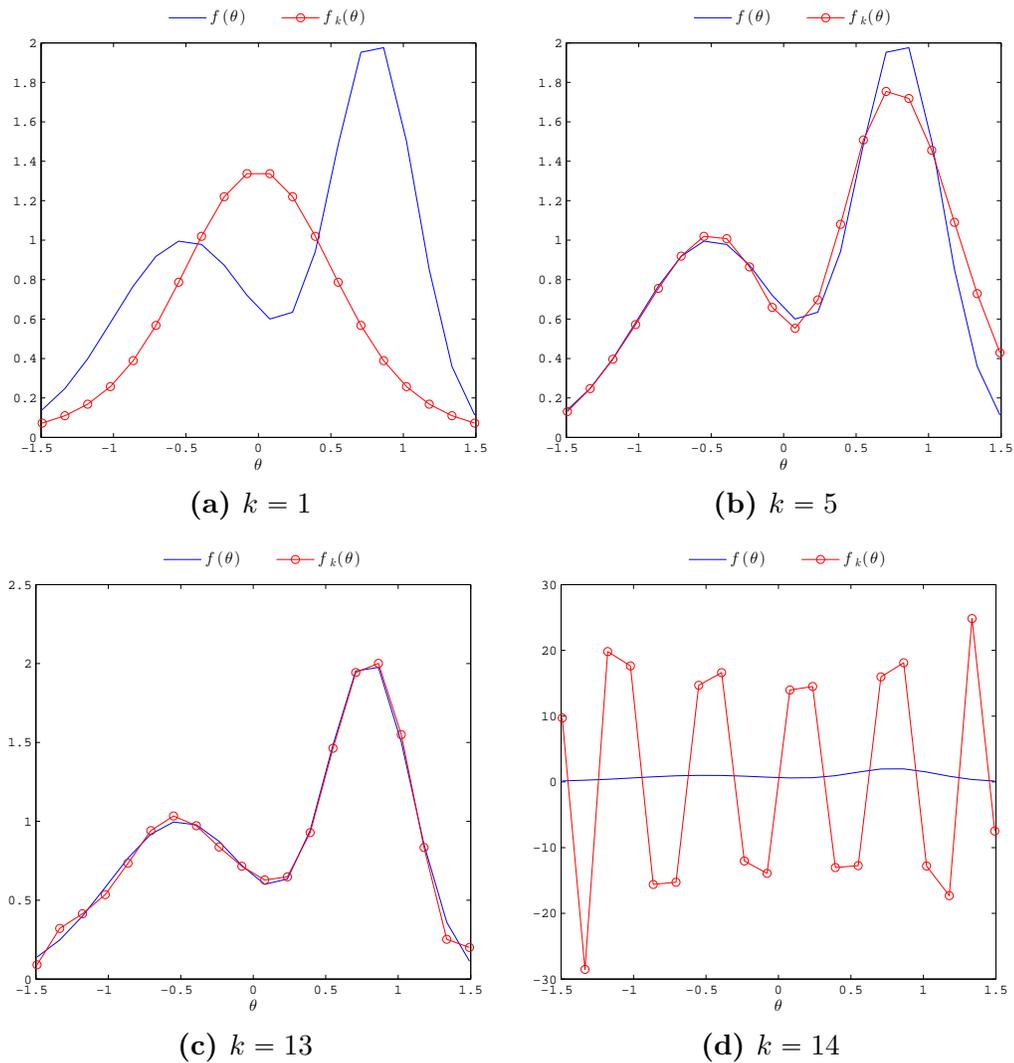
Ahora, al lado derecho  $\mathbf{b}$  le agregamos ruido  $\boldsymbol{\epsilon}$  que está idénticamente distribuido bajo una gaussiana de media cero y desviación estandar 0.01. Queremos recuperar los valores de la función  $f$  a partir del sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ .



**Figura 3.1:** Poligonales  $f$  y  $f_{LS}$  que en el ángulo  $\theta = \phi_i$  tienen los valores  $\mathbf{x}^\dagger(i)$  y  $\mathbf{x}_{LS}(i)$ , respectivamente.



**Figura 3.2:** Gráfica de Picard para el problema discreto  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$  de reconstrucción 1D del haz del Ejemplo 2.5. Mostramos en escala logarítmica a los valores singulares de  $A$ , los coeficientes de  $\mathbf{b} + \boldsymbol{\epsilon}$  en base de vectores singulares de derecha de  $A$  y la poligonal con las medias geométricas móviles de los coeficientes  $|u_i^T(\mathbf{b} + \boldsymbol{\epsilon})|/\sigma_i$  del estimador  $\mathbf{x}_{LS}$  de la solución de norma mínima  $\mathbf{x}^\dagger$  de  $A\mathbf{x} = \mathbf{b}$ .



**Figura 3.3:** Para la reconstrucción 1D del haz de luz en el Ejemplo 2.5 aproximamos  $f$  por la poligonal  $f_k$  que en los puntos  $t_i$  tiene los valores de la componentes de la solución regularizada  $\mathbf{x}_k$  obtenida por SVD truncada.

Sea  $\mathbf{x}_{LS}$  el estimador de cuadrados mínimos de  $\mathbf{x}^\dagger$  que obtenemos a partir de la ecuación  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ . En el ángulo  $\phi_i$ , la  $i$ -ésima componente del estimador  $\mathbf{x}_{LS}$ , denotada por  $\mathbf{x}_{LS}(i)$ , aproxima a  $\mathbf{x}^\dagger(i)$ . Mediante una poligonal  $f_{LS}$  unimos los puntos  $(\phi_i, \mathbf{x}_{LS}(i))$ . En la Figura 3.1 comparamos los valores de las componentes de  $\mathbf{x}_{LS}$  y  $\mathbf{x}^\dagger$  con las gráficas de  $f_{LS}$  y  $f$ , respectivamente. Los valores de  $f_{LS}$  sobrepasan notablemente a los de  $f$ , pues las discrepancias entre  $\mathbf{x}_{LS}$  y  $\mathbf{x}^\dagger$  son considerables. Esto se debe al mal condicionamiento del problema discreto. De hecho,  $\kappa_2(A) \approx 9.7534 \times 10^{15}$ .

Como la sucesión de los valores singulares  $\sigma_i$  de  $A$  decrece más rápido que  $e^{-i/2}$ , tenemos

un problema severamente al planteado. En la Figura 3.2 mostramos la Gráfica de Picard para este problema discreto mal planteado. Vemos que los valores singulares  $\sigma_i$  de  $A$  decrecen más rápido que los valores absolutos de los coeficientes  $\mathbf{u}_i^T(\mathbf{b} + \boldsymbol{\epsilon})$  de  $\mathbf{b} + \boldsymbol{\epsilon}$  en la base de vectores singulares de derecha de  $A$ . Por eso observamos que la curva con la media geométrica móvil de los coeficientes  $|\mathbf{u}_i^T(\mathbf{b} + \boldsymbol{\epsilon})|/\sigma_i$  crece. Luego, la DPC no se cumple. Aun así, tratamos de regularizar el problema discreto mal planteado.

Usamos la SVD truncada para regularizar el problema. Damos soluciones regularizadas  $\mathbf{x}_k$  para cuatro valores del nivel de truncamiento  $k$ . Mediante una poligonal  $f_k$  unimos los 20 puntos  $(\phi_i, \mathbf{x}_k(i))$ . En la Figura 3.3(a) nos quedamos solamente con el valor singular más grande ( $k = 1$ ). La aproximación  $f_k$  que obtenemos en este caso es una campana con máximo en  $\theta = 0$  que no se ajusta a  $f$ . En la Figura 3.3(b), tomamos  $k = 5$ . Esta vez la poligonal  $f_k$  se ajusta mejor a  $f$ , aunque tiene discrepancias pequeñas para  $0 \leq \theta \leq -\pi/2$ . Para  $k = 13$ , se reducen las diferencias entre los valores de  $f_k$  y  $f$  sin tener más oscilaciones como pueve verse en la Figura 3.3(c). Cuando  $k = 14$ , se generan oscilaciones de alta amplitud en la poligonal  $f_k$  que sobrepasan a los valores de  $f$ , pues las componentes ruidosas tienen mayor peso. Véase Figura 3.3(d).

En la práctica, las observaciones se ven afectadas por errores que repercuten en la solución del problema. En las señales con ruido gaussiano o en las mediciones tomadas por un dispositivo, el error es aleatorio.

En ocasiones, en un problema discreto mal planteado dado por la ecuación  $A\mathbf{x} = \mathbf{b}$ , se sabe que el vector de observaciones es

$$\mathbf{b} = \mathbf{b}_{\text{exacto}} + \boldsymbol{\epsilon},$$

donde  $\mathbf{b}_{\text{exacto}}$  es desconocido y  $\boldsymbol{\epsilon}$  es un vector aleatorio no observable. En términos de los valores y vectores singulares, la solución regularizadora  $\mathbf{x}_k$  puede expresarse como

$$\mathbf{x}_k = \sum_{i=1}^k \frac{\mathbf{u}_i^T \mathbf{b}_{\text{exacto}}}{\sigma_i} \mathbf{v}_i + \sum_{i=1}^k \frac{\mathbf{u}_i^T \boldsymbol{\epsilon}}{\sigma_i} \mathbf{v}_i, \quad k = 1, \dots, r.$$

El primer término es la solución de cuadrados mínimos del problema libre de ruido

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|A_k \mathbf{x} - \mathbf{b}_{\text{exacto}}\|_2,$$

mientras que el segundo término es la propagación del error en la solución del problema perturbado

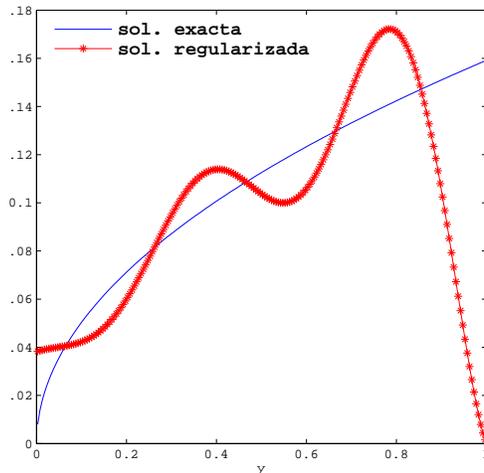
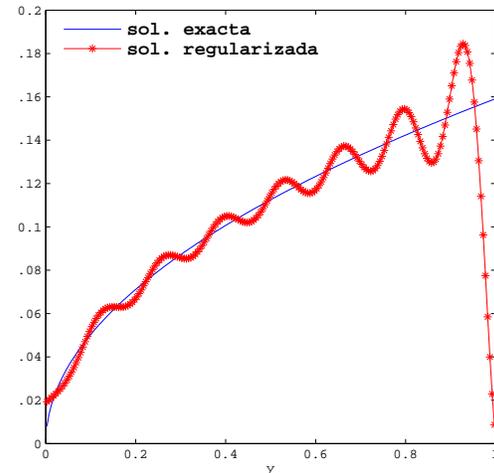
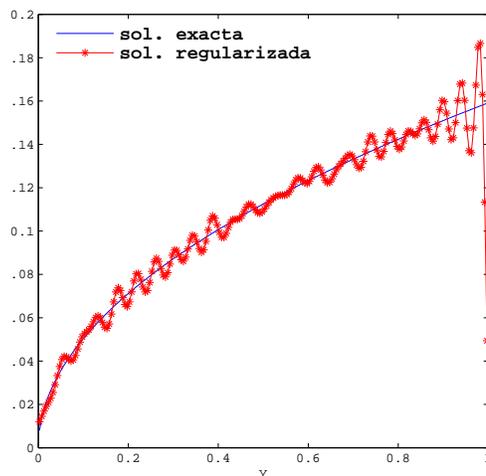
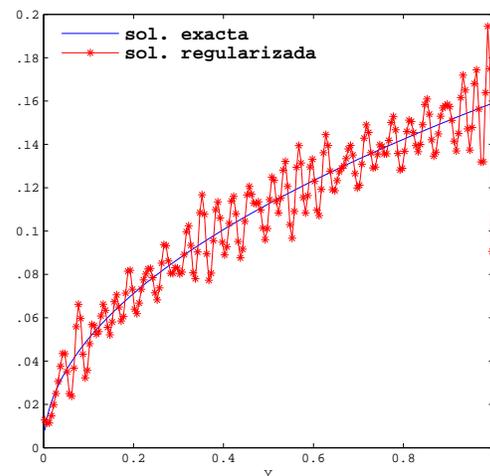
$$\min_{\mathbf{x} \in \mathbb{R}^n} \|A_k \mathbf{x} - (\mathbf{b}_{\text{exacto}} + \boldsymbol{\epsilon})\|_2.$$

Con la SVD truncada, tratamos de escoger el nivel de truncamiento  $k$  de modo que atenúemos la influencia del error en la solución regularizadora  $\mathbf{x}_k$ .

**Ejemplo 3.2.** Retomemos el problema de la forma del vertedero (Ejemplo 1.3). Dada la función del caudal  $c(z) = z^2$ , resolvemos numéricamente la ecuación integral

$$16 \int_0^z \sqrt{z-y} f(y) dy = c(z), \quad 0 \leq z \leq 1.$$

para encontrar valores aproximados de la función  $f$  que nos da la forma del vertedero.

(a)  $k = 5$ (b)  $k = 15$ (c)  $k = 50$ (d)  $k = 75$ 

**Figura 3.4:** Para el problema inverso del Ejemplo 1.3 aproximamos la solución  $f(y) = \sqrt{y}/(2\pi)$  por la solución regularizada  $x_k$  del problema discreto mediante SVD truncada.

Discretizamos la ecuación integral con el método de colocación en la malla uniforme

$z_j = j/200$ ,  $j = 0, \dots, 200$  del intervalo  $[0, 1]$ . De este modo,

$$z_i^2 = \int_0^{z_i} \sqrt{z_i - y} f(y) dy = \sum_{j=1}^i \int_{z_{j-1}}^{z_j} \sqrt{z_i - y} f(y) dy$$

Por cuadratura de punto medio, se sigue que

$$\left(\frac{i}{200}\right)^2 = \sum_{j=1}^i \frac{16}{200^{3/2}} \sqrt{i - j + \frac{1}{2}} f\left(\frac{2j-1}{2(200)}\right)$$

Así que el problema es resolver la ecuación  $A\mathbf{x} = \mathbf{b}$ , donde

$$b_i = \left(\frac{i}{200}\right)^2, \quad a_{i,j} = \begin{cases} \frac{16}{200^{3/2}} \sqrt{i - j + \frac{1}{2}}, & \text{si } i \geq j, \\ 0 & \text{en otro caso,} \end{cases} \quad i, j = 1, \dots, 200.$$

Agregamos ruido aleatorio  $\boldsymbol{\epsilon}$  de distribución normal con  $E(\boldsymbol{\epsilon}) = \mathbf{0}$  y  $\text{Cov}(\boldsymbol{\epsilon}) = 10^{-4}I$  al vector  $\mathbf{b}$ . Queremos recuperar los valores de la función de forma  $f$  a partir del sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ . Vamos a usar la SVD troncada para regularizar este problema discreto mal planteado.

En la Figura 3.4 aproximamos los valores de la función solución

$$f(y) = \sqrt{y}/(2\pi)$$

sobre el intervalo  $[0, 1]$  con los valores que nos da la solución regularizadora  $\mathbf{x}_k$  obtenida por SVD troncada. Para  $k = 5$ , la solución regularizada oscila con altas amplitudes. Así que las discrepancias entre la función  $f$  y los valores de las  $\mathbf{x}_k$  son notables. Conforme aumentamos el parámetro, de  $k = 15$  a  $k = 50$  vemos como la solución regularizada oscila más y reduce su amplitud. Si continuamos aumentando el nivel de truncamiento, el ruido ejerce mayor influencia y las discrepancias aumentan como se ve en el caso  $k = 75$ .

### 3.1.2. SVD Selectiva

Una variante de la SVD troncada consiste en seleccionar los coeficientes de la solución de cuadrados mínimos de norma mínima de la ecuación  $A\mathbf{x} = \mathbf{b}$  que aporten información relevante. Lo que hacemos es usar los coeficientes  $\mathbf{u}_i^T \mathbf{b}$  de  $\mathbf{x}^\dagger$  de tamaño mayor que un umbral dado  $\tau > 0$ . De este modo, en vez de emplear  $\mathbf{x}_k$ , usamos

$$\mathbf{x}_\tau := \sum_{|\mathbf{u}_i^T \mathbf{b}| > \tau} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i.$$

Este método se conoce como *SVD selectiva*. El parámetro de regularización es  $\tau$  y la solución regularizada es  $\mathbf{x}_\tau$ . La elección pertinente del umbral  $\tau$  debe remover los coeficientes

$\mathbf{u}_i^T \mathbf{b}$  por debajo del nivel del ruido [101].

Cuando el vector de observaciones  $\mathbf{b}$  tiene error aditivo aleatorio  $\boldsymbol{\epsilon}$ , vemos la influencia de este error en los coeficientes  $\mathbf{u}_i^T \mathbf{b}$  de  $\mathbf{x}_\tau$  para elegir el parámetro  $\tau$ . En ese caso,  $\mathbf{b}$  es un vector aleatorio.

Considere el modelo de regresión

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

donde  $E(\boldsymbol{\epsilon}) = \mathbf{0}$  y  $\text{Cov}(\boldsymbol{\epsilon}) = \eta^2 I$ . El vector  $\mathbf{b}$  es aleatorio. En cambio,  $\mathbf{u}_i^T A\mathbf{x}$  no lo es. Entonces

$$\text{var}(\mathbf{u}_i^T \mathbf{b}) = \text{var}(\mathbf{u}_i^T A\mathbf{x} + \mathbf{u}_i^T \boldsymbol{\epsilon}) = \text{var}(\mathbf{u}_i^T \boldsymbol{\epsilon})$$

Puesto que

$$\text{var}(\mathbf{u}_i^T \boldsymbol{\epsilon}) = \mathbf{u}_i^T \text{Cov}(\boldsymbol{\epsilon}) \mathbf{u}_i,$$

se sigue que

$$\text{var}(\mathbf{u}_i^T \mathbf{b}) = \mathbf{u}_i^T \text{Cov}(\boldsymbol{\epsilon}) \mathbf{u}_i$$

Luego,  $\|\mathbf{u}_i\|_2 = 1$  y  $\text{Cov}(\boldsymbol{\epsilon}) = \eta^2 I$  implican que

$$\text{var}(\mathbf{u}_i^T \mathbf{b}) = \eta^2.$$

Por eso el umbral que escogemos es  $\tau = \eta$ .

#### Observaciones 3.4:

☞ Para evitar incluir coeficientes al nivel del ruido, incluimos un factor  $\nu > 0$  como sugiere Hansen [54]. Así que tomamos  $\tau = \nu\eta$ .

**Ejemplo 3.3.** En el Ejemplo 2.1, usamos el método de Galerkin con regla de trapecio para discretizar la Ecuación de Phillips

$$\int_{-6}^6 k(s-t)f(t)dt = g(s), \quad |s| \leq 6 \quad (3.3)$$

con núcleo

$$k(s) = \begin{cases} 1 + \cos\left(\frac{s\pi}{3}\right), & \text{si } |s| < 3, \\ 0, & \text{si } |s| \geq 3, \end{cases}$$

y observaciones

$$g(s) = (6 - |s|) \left( 1 + \frac{1}{2} \cos\left(\frac{s\pi}{3}\right) \right) + \frac{9}{2\pi} \sin\left(\frac{|s|\pi}{3}\right),$$

La aproximación de la solución  $f$  es la combinación lineal de  $n$  funciones sombrero

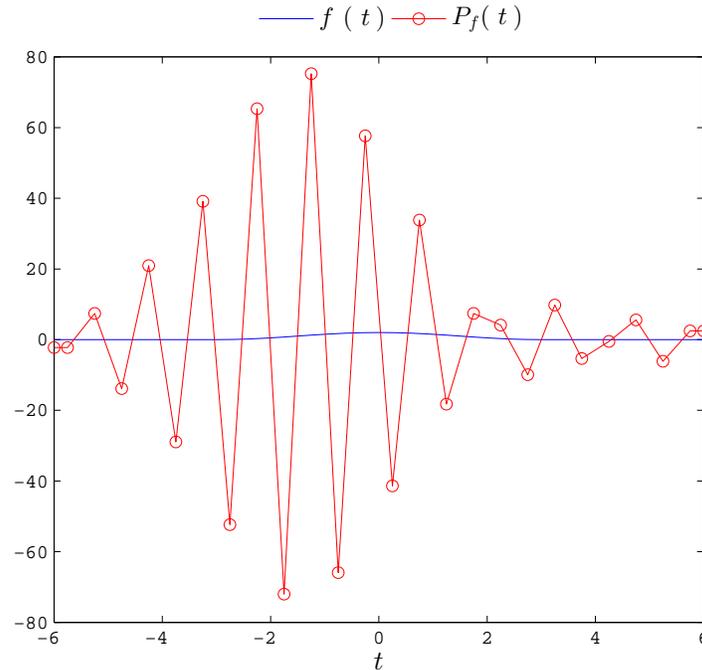
$$\varphi_j(s) = \begin{cases} \sqrt{\frac{n}{12}}, & \text{si } s_j \leq s < s_{j+1}, \\ 0, & \text{en otro caso,} \end{cases} \quad j = 1, \dots, n.$$

Los coeficientes  $x_j$  se obtienen de la ecuación  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ , donde

$$b_i = \frac{1}{2} \sqrt{\frac{12}{n}} (g(s_i) + g(s_{i+1})),$$

$$a_{i,j} = \frac{3}{n} (k(s_i - s_j) + k(s_{i+1} - s_j) + k(s_i - s_{j+1}) + k(s_{i+1} - s_{j+1})),$$

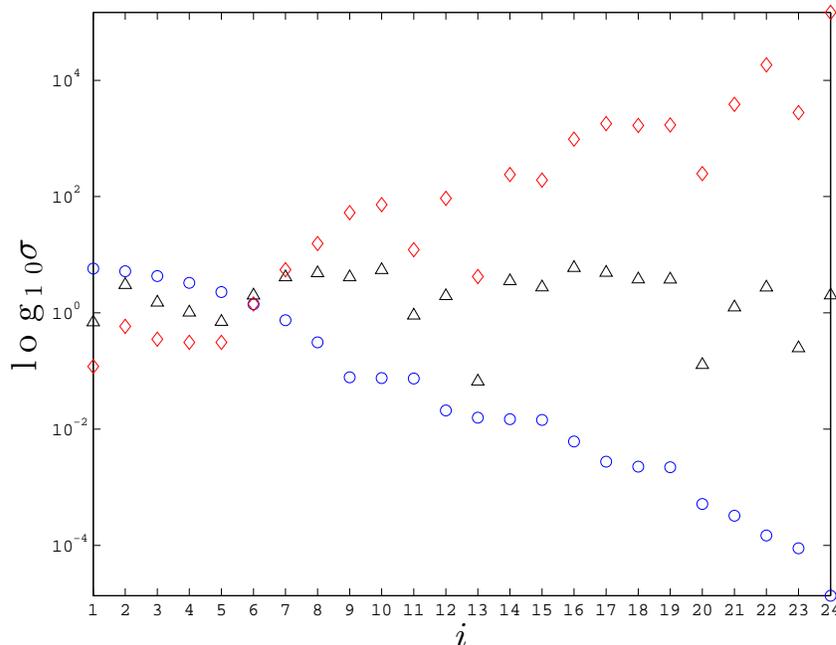
$\boldsymbol{\epsilon} = D\boldsymbol{\epsilon}'$ , donde  $D$  es una diagonal con elementos entre  $10^{-3}$  y  $10^{-2}$  y  $\boldsymbol{\epsilon}' \sim N(\mathbf{0}, I)$ .



**Figura 3.5:** Gráficas de la solución  $f$  de la ecuación de Phillips y la expansión  $P_f$  por funciones 24 sombrero cuando hay ruido gaussiano  $\boldsymbol{\epsilon}$  en observaciones

En la Figura 3.5 comparamos la solución analítica  $f$  de la Ecuación de Phillips con su aproximación  $P_f$  en  $n = 24$  funciones sombrero  $\varphi_j$ . Los coeficientes son las componentes del estimador  $\mathbf{x}_{LS}$  de de cuadrados mínimos de la ecuación  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ . El ruido gaussiano  $\boldsymbol{\epsilon}$  en las observaciones ocasiona que  $P_f$  presente oscilaciones. En este caso el valor singular más grande de  $A$  es  $\sigma_1 = 5.786$ , el más pequeño es  $\sigma_{24} = 1.3631 \times 10^{-5}$ , y por consiguiente el número de condición es  $\kappa_2(A) = 4.2447 \times 10^5$ .

La gráfica de Picard del problema se muestra en Figura 3.6. Observamos que la mayoría de los valores absolutos de los coeficientes  $\mathbf{u}_i^T(\mathbf{b} + \boldsymbol{\epsilon})/\sigma_i$  del estimador  $\mathbf{x}_{LS}$  van aumentando a partir del subíndice  $i \geq 5$ . Por lo que la condición discreta de Picard se cumple con los 5 valores singulares más grandes. Esto nos sugiere regularizar el problema



**Figura 3.6:** Gráfica de Picard para el problema discreto  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$  de la Ecuación de Phillips obtenido por método de Galerkin con 24 funciones sombrero. Mostramos en escala logarítmica a los valores singulares  $\sigma_i$  de  $A$  ( $\circ$ ), los coeficientes  $|\mathbf{u}_i^T(\mathbf{b} + \boldsymbol{\epsilon})|$  de  $\mathbf{b} + \boldsymbol{\epsilon}$  en base de vectores singulares de derecha de  $A$  ( $\triangle$ ), y los coeficientes  $|\mathbf{u}_i^T(\mathbf{b} + \boldsymbol{\epsilon})|/\sigma_i$  del estimador  $\mathbf{x}_{LS}$  ( $\diamond$ ).

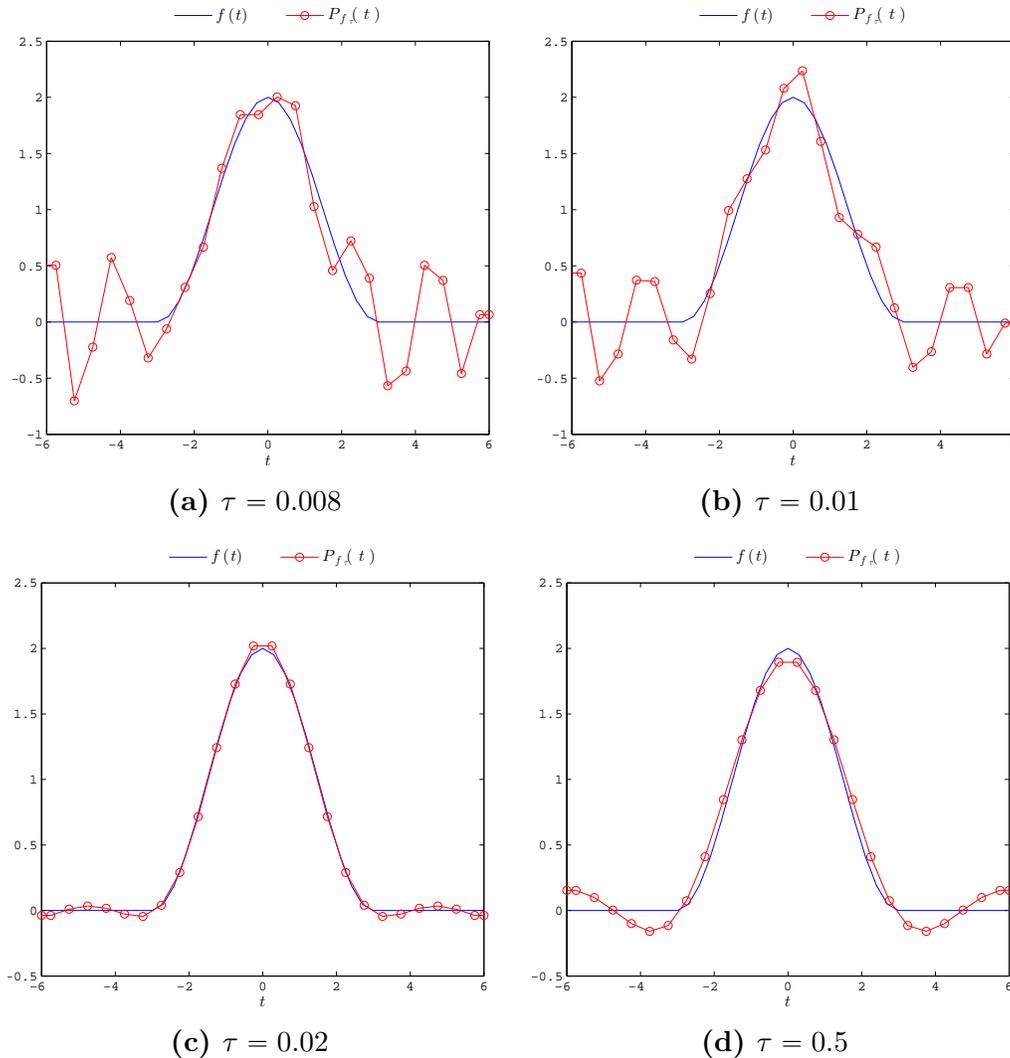
Usamos la SVD selectiva para obtener soluciones regularizadas  $\mathbf{x}_\tau$  de  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ . Las componentes  $\mathbf{x}_\tau(1), \dots, \mathbf{x}_\tau(24)$  de  $\mathbf{x}_\tau$  son los coeficientes de la combinación lineal de funciones sombrero

$$f_\tau(t) = \sum_{j=1}^{24} \mathbf{x}_\tau(j) \phi_j(t).$$

24 funciones sombrero		$\kappa_2(A) = 4.2447 \times 10^5$	
SVD selectiva		SVD truncada	
$\tau$	$\ f - f_\tau\ _\infty$	$k$	$\ f - f_k\ _\infty$
0.008	0.70258	5	0.159478
0.01	0.52457	10	0.070222
0.02	0.05322	15	0.682979
0.5	0.15863	20	5.57299
100 funciones sombrero		$\kappa_2(A) = 2.68399 \times 10^8$	
SVD selectiva		SVD truncada	
$\tau$	$\ f - f_\tau\ _\infty$	$k$	$\ f - f_k\ _\infty$
0.008	$9.22099 \times 10^4$	20	0.957077
0.01	$8.60705 \times 10^4$	50	33.9469
0.02	$4.01476 \times 10^{-2}$	75	$5.05458 \times 10^2$
0.5	0.145833	90	$6.36816 \times 10^3$

**Tabla 3.1:** Errores de las soluciones regularizadoras  $f_\tau$  y  $f_k$  obtenidas por SVD selectiva y truncada en el problema de la Ecuación de Phillips para 24 y 100 funciones sombrero.

En la Figura 3.7 mostramos las gráficas correspondientes a cuatro valores distintos de  $\tau$ . Con  $\tau = 0.02$  tenemos la menor discrepancia de las cuatro expansiones  $f_\tau$ .



**Figura 3.7:** La solución  $f = k$  de la Ecuación de Phillips y la combinación lineal  $f_\tau$  de 24 funciones sombrero. Los coeficientes de  $f_\tau$  son las componentes de  $\mathbf{x}_\tau$  y el parámetro de regularización es  $\tau$ .

En la Tabla 3.1 para  $n \in \{24, 100\}$  mostramos la discrepancia  $\|f - f_\tau\|_\infty$  para cuatro valores diferentes del umbral  $\tau$  además de la discrepancia  $\|f - f_k\|_\infty$  para el problema regularizado por SVD truncada para cuatro distintos niveles de truncamiento  $k$ . Observamos que cuando tomamos 100 en vez de 24 funciones sombrero,  $\|f - f_k\|_\infty$  aumenta del orden de  $10^{-1}$  a  $10^3$  cuando  $k$  aumenta de 20 hasta 90. En cambio,  $\|f - f_\tau\|_\infty$  es del orden de  $10^4$  para  $\tau \leq 0.01$ ; mientras que  $\|f - f_\tau\|_\infty$  es del orden de  $10^{-2}$  para  $\tau = 0.02$ .

## 3.2. Factores Filtro

La estimación estadística y la construcción de modelos son problemas inversos donde deseamos hacer inferencias sobre un fenómeno a partir de información parcial o completa. A menudo, los valores de los estimadores de cuadrados mínimos, de máxima verosimilitud o de distancia mínima son sensibles a pequeñas perturbaciones en los datos. Por lo que tratamos con problemas inversos mal planteados. Los algoritmos para resolver estos problemas inversos se conocen como *algoritmos de inversión*. El principio básico que siguen es buscar una solución que sea consistente tanto con los datos observados como con la información *a priori* sobre el comportamiento del fenómeno bajo estudio. Los métodos de regularización implementan este principio.

Considere el problema discreto mal planteado dado por  $\mathbf{b} = A\mathbf{x} + \boldsymbol{\epsilon}$ , donde  $A \in \mathbb{R}^{m \times n}$  de rango  $n$  con  $m \geq n$ . Sea  $A = U\Sigma V^T$  una SVD de  $A$ . Tanto la SVD truncada y la SVD selectiva usan los cocientes  $(\mathbf{u}_i^T \mathbf{b})/\sigma_i$  del estimador de cuadrados mínimos de  $\mathbf{x}$ . Cuando aplicamos estos métodos, filtramos los coeficientes ruidosos de la solución del problema.

Introducimos pesos  $\varphi_1, \dots, \varphi_n \in [0, 1]$  en la solución de cuadrados mínimos de norma mínima dada por la SVD de modo que la solución regularizada sea

$$\mathbf{x}_{\text{reg}} := \sum_{i=1}^n \varphi_i \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i.$$

Llamamos *factores filtro* a los pesos  $\varphi_i$  [51]. Éstos dependen de un parámetro continuo o discreto. En el caso de la SVD truncada con nivel de truncamiento  $k \in \mathbb{N}$ , tenemos que

$$\varphi_i = \begin{cases} 1, & \text{si } i \leq k, \\ 0, & \text{de otro modo.} \end{cases}$$

En cambio, si usamos la SVD selectiva con un umbral  $\tau > 0$ , entonces

$$\varphi_i = \begin{cases} 1, & \text{si } |\mathbf{u}_i^T \mathbf{b}| > \tau, \\ 0, & \text{de otro modo.} \end{cases}$$

Mediante las matrices

$$\Phi = \begin{bmatrix} \varphi_1 & & 0 \\ & \ddots & \\ 0 & & \varphi_n \end{bmatrix} \quad \text{y} \quad \Sigma^\dagger = \begin{bmatrix} \frac{1}{\sigma_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_n} \\ \hline \mathbf{0}_{(m-n) \times n} \end{bmatrix}.$$

podemos escribir la solución regularizadora en aritmética exacta como

$$\mathbf{x}_{\text{reg}} = V\Phi\Sigma^\dagger U^T \mathbf{b}.$$

El objetivo del método de regularización por factores filtro es usar a  $\mathbf{x}_{\text{reg}}$  como un estimador de  $\mathbf{x}$ . Para medir lo cerca que está  $\mathbf{x}_{\text{reg}}$  de  $\mathbf{x}$ , usamos el *error en media cuadrática*:

$$\text{MSE}(\mathbf{x}_{\text{reg}}) = E(\|\mathbf{x}_{\text{reg}} - \mathbf{x}\|_2^2).$$

O'Sullivan [94] menciona que las características de un algoritmo de inversión pueden estudiarse mediante el error en media cuadrática, el cual puede separarse en sesgo y varianza. El sesgo mide errores del sistema, mientras que la varianza mide errores aleatorios. Usamos esta idea en la regularización por factores filtro.

En términos del sesgo  $\mathbf{x} - E(\mathbf{x}_{\text{reg}})$  y la matriz de covarianza de  $\mathbf{x}_{\text{reg}}$ , el error en media cuadrática se expresa como

$$\text{MSE}(\mathbf{x}_{\text{reg}}) = \text{Tr}(\text{Cov}(\mathbf{x}_{\text{reg}})) + \|\mathbf{x} - E(\mathbf{x}_{\text{reg}})\|_2^2. \quad (3.4)$$

Como el ruido  $\boldsymbol{\epsilon}$  que tratamos cumple con las condiciones de Gauss-Markov, podemos expresar  $\text{MSE}(\mathbf{x}_{\text{reg}})$  en términos de los factores filtro y valores singulares.

**Proposición 3.1.** *El error en media cuadrática del estimador  $\mathbf{x}_{\text{reg}}$  es*

$$\text{MSE}(\mathbf{x}_{\text{reg}}) = \eta^2 \sum_{i=1}^n \left( \frac{\varphi_i}{\sigma_i} \right)^2 + \sum_{i=1}^n (1 - \varphi_i)^2 (\mathbf{v}_i^T \mathbf{x})^2.$$

*Demostración.* El valor esperado de la solución regularizadora  $\mathbf{x}_{\text{reg}}$  es

$$E(\mathbf{x}_{\text{reg}}) = V\Phi\Sigma^\dagger U^T E(A\mathbf{x}) + V\Phi\Sigma^\dagger U^T E(\boldsymbol{\epsilon})$$

Dado que  $E(\boldsymbol{\epsilon}) = \mathbf{0}$  y no hacemos supuestos estadísticos sobre  $A\mathbf{x}$ , tenemos que

$$E(\mathbf{x}_{\text{reg}}) = V\Phi\Sigma^\dagger U^T A\mathbf{x} = V\Phi\Sigma^\dagger \Sigma V^T \mathbf{x}.$$

Puesto que  $A$  es de rango completo,  $\Sigma^\dagger \Sigma = I_{n \times n}$ . En consecuencia,

$$E(\mathbf{x}_{\text{reg}}) = V\Phi V^T \mathbf{x}.$$

Entonces

$$\mathbf{x} - E(\mathbf{x}_{\text{reg}}) = (I - V\Phi V^T) \mathbf{x}$$

Debido a que  $V$  es una matriz ortogonal, se sigue que

$$\mathbf{x} - E(\mathbf{x}_{\text{reg}}) = V(I - \Phi)V^T \mathbf{x}$$

Así que

$$\|\mathbf{x} - E(\mathbf{x}_{\text{reg}})\|_2^2 = \sum_{i=1}^n (1 - \varphi_i)^2 (\mathbf{v}_i^T \mathbf{x})^2 \quad (3.5)$$

Por otra parte,

$$\mathbf{x}_{\text{reg}} - E(\mathbf{x}_{\text{reg}}) = V\Phi\Sigma^\dagger U^T \boldsymbol{\epsilon}.$$

Por lo que

$$\text{Cov}(\mathbf{x}_{\text{reg}}) = E\left((V\Phi\Sigma^\dagger U^T \boldsymbol{\epsilon})(V\Phi\Sigma^\dagger U^T \boldsymbol{\epsilon})^T\right)$$

esto es,

$$\text{Cov}(\mathbf{x}_{\text{reg}}) = V\Phi\Sigma^\dagger U^T E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) U\Sigma^\dagger \Phi^T V^T.$$

Puesto que la matriz  $U$  es ortogonal y

$$E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \eta^2 I,$$

se sigue que

$$\text{Cov}(\mathbf{x}_{\text{reg}}) = \eta^2 V\Phi\Sigma^\dagger \Sigma^{\dagger T} F^T V^T,$$

de donde

$$\text{Cov}(\mathbf{x}_{\text{reg}}) = \eta^2 \sum_{i=1}^n \left(\frac{\varphi_i}{\sigma_i}\right)^2 \mathbf{v}_i \mathbf{v}_i^T.$$

Por consiguiente,

$$\text{Tr}(\text{Cov}(\mathbf{x}_{\text{reg}})) = \eta^2 \sum_{i=1}^n \left(\frac{\varphi_i}{\sigma_i}\right)^2 \text{Tr}(\mathbf{v}_i \mathbf{v}_i^T).$$

Debido a que

$$\text{Tr}(\mathbf{v}_i \mathbf{v}_i^T) = \mathbf{v}_i^T \mathbf{v}_i = 1,$$

tenemos que

$$\text{Tr}(\text{Cov}(\mathbf{x}_{\text{reg}})) = \eta^2 \sum_{i=1}^n \left(\frac{\varphi_i}{\sigma_i}\right)^2. \quad (3.6)$$

En consecuencia, al sustituir las fórmulas (3.5) y (3.6) en la Identidad (3.4), obtenemos la identidad deseada. 

Sin importar como elijamos los factores filtro, el error en media cuadrática de  $\mathbf{x}_{\text{reg}}$  tiene un valor mínimo respecto a los  $\varphi_i$  [19]. Para ver esto, sea  $M : [0,1]^n \rightarrow \mathbb{R}$  la función dada por

$$M(\varphi_1, \dots, \varphi_n) = \text{MSE}(\mathbf{x}_{\text{reg}}).$$

A partir de la expresión (3.2) que nos da el Teorema 3.1, tenemos que

$$\frac{\partial M}{\partial \varphi_i} = \frac{2}{\sigma_i^2} [(\eta^2 + (\mathbf{v}_i^T \mathbf{x})^2) \varphi_i - (\mathbf{v}_i^T \mathbf{x})^2], \quad i = 1, \dots, n.$$

Entonces los factores filtro óptimos son

$$\varphi_i^{\text{opt}} = \frac{(\mathbf{v}_i^T \mathbf{x})^2}{\eta^2 + (\mathbf{v}_i^T \mathbf{x})^2}, \quad i = 1, \dots, r.$$

Por lo que el valor mínimo de  $\text{MSE}(\mathbf{x}_{\text{reg}})$  es

$$M(\varphi_1^{\text{opt}}, \dots, \varphi_n^{\text{opt}}) = \eta^2 \sum_{i=1}^n \frac{(\mathbf{v}_i^T \mathbf{x})^2}{\sigma_i^2 (\eta^2 + (\mathbf{v}_i^T \mathbf{x})^2)}.$$

Una manera para que este valor mínimo sea pequeño es que los sumandos sean menores que un umbral  $\delta > 0$ :

$$\frac{\eta^2 (\mathbf{v}_i^T \mathbf{x})^2}{\sigma_i^2 (\eta^2 + (\mathbf{v}_i^T \mathbf{x})^2)} \leq \delta, \quad i = 1, \dots, n,$$

de donde

$$(\mathbf{v}_i^T \mathbf{x})^2 \leq \frac{\delta \eta^2 \sigma_i^2}{\eta^2 / \sigma_i^2 - \delta}, \quad i = 1, \dots, n.$$

Como  $A$  es de rango completo por columnas, tenemos que

$$\mathbf{u}_i^T A \mathbf{x} = \mathbf{u}_i^T U \Sigma V^T \mathbf{x} = \sigma_i \mathbf{v}_i^T \mathbf{x}, \quad i = 1, \dots, n.$$

En consecuencia,

$$\frac{(\mathbf{u}_i^T A \mathbf{x})^2}{\sigma_i^2} \leq \frac{\delta \eta^2 \sigma_i^2}{\eta^2 / \sigma_i^2 - \delta}, \quad i = 1, \dots, n.$$

Así,  $(\mathbf{u}_i^T A \mathbf{x})^2 / \sigma_i^2$  está acotado por arriba por una cota que se hace más pequeña conforme  $i$  aumenta. De aquí, los coeficientes  $|\mathbf{u}_i^T A \mathbf{x}|$  deben decaer a cero más rápido que los valores singulares para que el valor mínimo de  $\text{MSE}(\mathbf{x}_{\text{reg}})$  sea pequeño.

### Observaciones 3.5:

 Si reemplazamos a  $A \mathbf{x}$  por  $\mathbf{b} = A \mathbf{x} + \boldsymbol{\epsilon}$  en el cociente  $(\mathbf{u}_i^T A \mathbf{x})^2 / \sigma_i^2$ , entonces lo que estamos pidiendo es que  $|\mathbf{u}_i^T \mathbf{b}|$  debe decaer a cero más rápido que  $\sigma_i$ , esto es, que se cumpla con la DPC.

## 3.3. Regularización de Tikhonov

Un método clásico para regularizar un problema discreto mal planteado es penalizar el Problema de Cuadrados Mínimos (2.6) con el tamaño de la solución. Este método se conoce como *regularización de Tikhonov* [113],[114]. Las soluciones de cuadrados mínimos del problema discreto mal planteado están dominadas por coeficientes asociados a altas frecuencias. Cuando incluimos el término  $\|\mathbf{x}\|_2$  en el problema de cuadrados mínimos, intentamos suprimir el efecto de estas altas frecuencias.

La regularización de Tikhonov consiste en resolver

$$\min_{\mathbf{x} \in \mathbb{R}^n} \{ \|\mathbf{A} \mathbf{x} - \mathbf{b}\|_2^2 + \lambda^2 \|\mathbf{x}\|_2^2 \}. \quad (3.7)$$

para un factor  $\lambda > 0$  que es nuestro parámetro de regularización [30]. La solución regularizada se denota por  $\mathbf{x}_\lambda$ . El balance entre el tamaño del residuo  $\mathbf{r}_\lambda = A \mathbf{x}_\lambda - \mathbf{b}$  y el tamaño de la solución  $\mathbf{x}_\lambda$  está controlado por  $\lambda$ .

**Observaciones 3.6:**

☞ Mientras más grande sea  $\lambda$ , el tamaño de la solución regularizada disminuye:

$$\|\mathbf{x}_\lambda\|_2 \rightarrow 0 \quad \text{cuando} \quad \lambda \rightarrow \infty.$$

Por otra parte, mientras más pequeña sea  $\lambda$ , la solución regularizadora se acerca más al estimador de cuadrados mínimos  $\mathbf{x}^\dagger$  del problema discreto mal planteado:

$$\|\mathbf{x}_\lambda - \mathbf{x}^\dagger\|_2 \rightarrow 0 \quad \text{cuando} \quad \lambda \rightarrow 0.$$

Notamos que

$$\left\| \begin{bmatrix} A \\ \lambda I \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} A\mathbf{x} - \mathbf{b} \\ \lambda\mathbf{x} \end{bmatrix} \right\|_2^2 = \|A\mathbf{x} - \mathbf{b}\|_2^2 + \|\lambda\mathbf{x}\|_2^2 \quad \forall \mathbf{x} \in \mathbb{R}^n.$$

Por lo que resolver (3.7) equivale a resolver el problema de cuadrados mínimos amortiguados

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\| \begin{bmatrix} A \\ \lambda I \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\|_2^2.$$

Los puntos críticos son las soluciones de la ecuación

$$\begin{bmatrix} A \\ \lambda I \end{bmatrix}^T \begin{bmatrix} A \\ \lambda I \end{bmatrix} \mathbf{x} = \begin{bmatrix} A \\ \lambda I \end{bmatrix}^T \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix},$$

A partir de esta ecuación, obtenemos las **ecuaciones normales regularizadas**

$$(A^T A + \lambda^2 I) \mathbf{x} = A^T \mathbf{b}. \quad (3.8)$$

El parametro  $\lambda$  es un desplazamiento con el que intentamos pasar de columnas colineales a columnas ortogonales.

Mediante la SVD de  $A$  podemos conocer la solución de las ecuaciones normales regularizadas.

**Teorema 3.2.** *Sea  $A \in \mathbb{R}^{m \times n}$  con  $m \geq n$ . Sean  $\sigma_1, \dots, \sigma_n$  sus valores singulares, posiblemente algunos iguales a cero. Sean  $\mathbf{u}_1, \dots, \mathbf{u}_m$  y  $\mathbf{v}_1, \dots, \mathbf{v}_n$  sus vectores singulares de izquierda y derecha, respectivamente. Entonces la solución de las ecuaciones normales regularizadas (3.8) es*

$$\mathbf{x}_\lambda = \sum_{i=1}^n \frac{\sigma_i}{\sigma_i^2 + \lambda^2} (\mathbf{u}_i^T \mathbf{b}) \mathbf{v}_i.$$

*Demostración.* La matriz  $A$  tiene una SVD

$$A = \underbrace{[\mathbf{u}_1 \ \cdots \ \mathbf{u}_m]}_{U_{m \times m}} \underbrace{\begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_n & \\ \hline & & & \mathbf{0} \end{bmatrix}}_{\Sigma_{m \times n}} \underbrace{[\mathbf{v}_1 \ \cdots \ \mathbf{v}_n]^T}_{V_{n \times n}^T}$$

Dado que

$$U^T U = I_m \quad \text{y} \quad V V^T = I_n,$$

tenemos que

$$A^T A + \lambda^2 I = V \Sigma \Sigma^T V^T + \lambda^2 V V^T,$$

es decir,

$$A^T A + \lambda^2 I = V \text{diag}(\sigma_1^2 + \lambda^2, \dots, \sigma_n^2 + \lambda^2) V^T.$$

Notamos que la matriz diagonal tiene los valores propios de  $A^T A + \lambda^2 I$  y los vectores singulares de derecha son sus respectivos vectores propios. Como  $\lambda \neq 0$ , entonces los valores propios de  $A^T A + \lambda I$  son positivos. Por lo que esta matriz es positiva definida, de hecho, su inversa es

$$(A^T A + \lambda^2 I)^{-1} = V \text{diag} \left( \frac{1}{\sigma_1^2 + \lambda^2}, \dots, \frac{1}{\sigma_n^2 + \lambda^2} \right) V^T.$$

Por consiguiente, la solución de las Ecuaciones Normales Regularizadas (3.8) es

$$\mathbf{x}_\lambda = V \text{diag} \left( \frac{1}{\sigma_1^2 + \lambda^2}, \dots, \frac{1}{\sigma_n^2 + \lambda^2} \right) V^T V \Sigma^T U^T \mathbf{b}.$$

Como

$$V^T V = I_n \quad \text{y} \quad \Sigma^T U^T \mathbf{b} = \begin{bmatrix} \sigma_1 \mathbf{u}_1^T \mathbf{b} \\ \vdots \\ \sigma_n \mathbf{u}_n^T \mathbf{b} \end{bmatrix},$$

entonces

$$\mathbf{x}_\lambda = [\mathbf{v}_1 \quad \dots \quad \mathbf{v}_n] \text{diag} \left( \frac{1}{\sigma_1^2 + \lambda^2}, \dots, \frac{1}{\sigma_n^2 + \lambda^2} \right) \begin{bmatrix} \sigma_1 \mathbf{u}_1^T \mathbf{b} \\ \vdots \\ \sigma_n \mathbf{u}_n^T \mathbf{b} \end{bmatrix}.$$

De este modo, conseguimos la expansión deseada de  $\mathbf{x}_\lambda$ . ♣

### Observaciones 3.7:

 La regularización de Tikhonov usa los factores filtro

$$\varphi_i(\lambda) = \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2}, \quad i = 1, \dots, n.$$

 Bajo condiciones de Gauss-Markov, el sesgo de  $\mathbf{x}_\lambda$  es

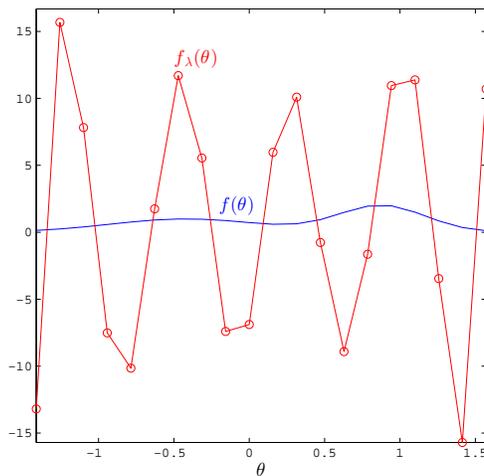
$$\|\mathbf{x} - E(\mathbf{x}_\lambda)\|_2^2 = \sum_{i=1}^n \left( \frac{\lambda^2}{\sigma_i^2 + \lambda^2} \right)^2 (\mathbf{v}_i^T \mathbf{x})^2.$$

**Ejemplo 3.4.** El problema discreto de la reconstrucción del haz de luz (Ejemplo 2.5) es resolver la ecuación  $A\mathbf{x} = \mathbf{b}$ , donde

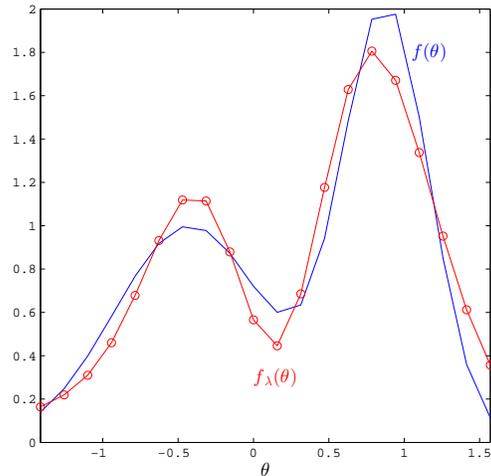
$b_i = g(\phi_i)$  dado por la Tabla 2.3

$$a_{i,j} = \frac{\pi}{20} (\cos \phi_i + \cos \phi_j)^2 \left( \frac{\sin(\pi(\sin \phi_i + \sin \phi_j))}{\pi(\sin \phi_i + \sin \phi_j)} \right)^2, \quad i, j = 1, \dots, 20.$$

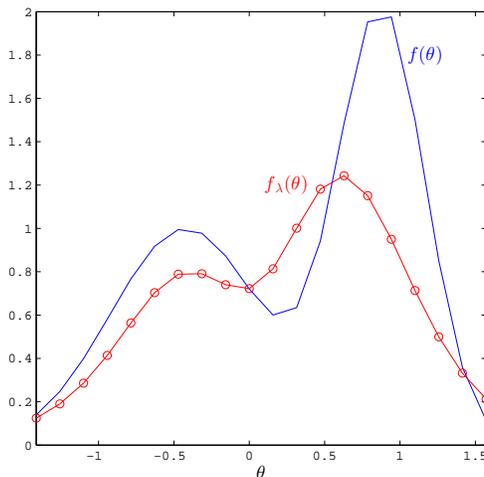
$$\phi_i = (i - 1/2)\pi/20 - \pi/2,$$



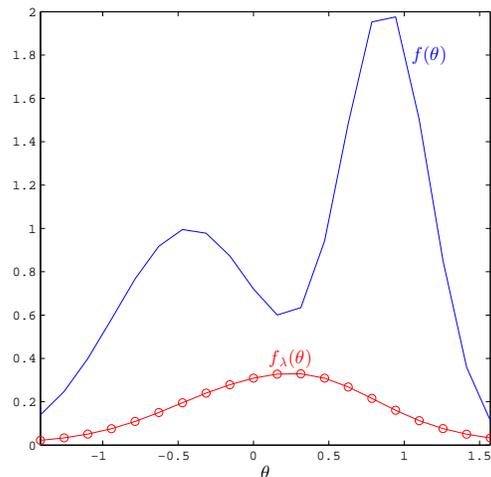
(a)  $\lambda = 10^{-4}$



(b)  $\lambda = 0.1$



(c)  $\lambda = 1$



(d)  $\lambda = 5$

**Figura 3.8:** Poligonal  $f$  que nos da la intensidad de luz incidente en la reconstrucción 1D del haz en el Ejemplo 2.5 y la poligonal  $f_\lambda$  que toma los valores de las componentes de la solución regularizadora de Tikhonov  $\mathbf{x}_\lambda$ .

Sea  $\mathbf{x}^\dagger$  la solución de cuadrados mínimos de norma mínima de  $A\mathbf{x} = \mathbf{b}$ . Agregamos a  $\mathbf{b}$  ruido  $\boldsymbol{\epsilon}$  distribuido bajo una gaussiana con  $E(\boldsymbol{\epsilon}) = \mathbf{0}$  y  $\text{Cov}(\boldsymbol{\epsilon}) = 10^{-4}I$ . A partir de la ecuación  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ , deseamos recuperar los valores de la poligonal  $f$  que en el ángulo  $\phi_i$  tiene el valor  $\mathbf{x}^\dagger(i)$ . Esta vez usamos regularización de Tikhonov.

Sea  $f_\lambda$  la poligonal sobre el intervalo  $[-\pi/2, \pi/2]$  que interpola las componentes de la solución regularizada  $\mathbf{x}_\lambda$  en los ángulos  $\phi_i$ . En la Figura 3.8(a) mostramos la función  $f$  junto con su aproximación  $f_\lambda$  para  $\lambda = 10^{-4}$ . Observamos las altas oscilaciones de  $f_\lambda$  a comparación de la función  $f$ . En la Figura 3.8(b) observamos que para  $\lambda = 0.1$ , las discrepancias entre  $f$  y  $f_\lambda$  son pequeñas. Si el parámetro  $\lambda$  se hace más grande, los valores de  $f_\lambda$  están cerca de cero. Como puede verse en la Figura 3.8(d). De las cuatro soluciones regularizadas tomamos la del parámetro  $\lambda = 0.1$ .

### Observaciones 3.8:

☞ En los ejemplos que hemos visto en este capítulo, es importante seleccionar un valor adecuado del parámetro de regularización para que la solución regularizadora obtenida por SVD truncada, SVD selectiva, Tikhonov o factores filtro aproxime a la solución del problema mal planteado con pequeños errores. Esta tarea se explica con detalle en la Sección 3.5.

#### 3.3.1. Aplicación en Imágenes Digitales

En el siguiente ejemplo tratamos un problema mal planteado del procesamiento de imágenes digitales.

**Deblurring:** Dada una imagen difuminada, obtener una imagen de la misma escena con menos degradaciones.

La idea de este problema es obtener un modelo para difuminar imágenes. Dada una imagen ideal  $\mathcal{F}$ , buscamos una transformación  $\mathcal{B}$  entre espacios de imágenes que la convierta en la imagen difuminada  $\mathcal{G}$ . El problema directo es dados  $\mathcal{F}$  y  $\mathcal{B}$ , obtener  $\mathcal{G} = \mathcal{B}(\mathcal{F})$ . El deblurring es el problema inverso, a saber, dados  $\mathcal{G}$  y  $\mathcal{B}$ , hallar  $\mathcal{F}$  tal que  $\mathcal{B}(\mathcal{F}) = \mathcal{G}$ .

Para dar una formulación matemática del problema, representamos una imagen digital en escala de grises mediante una matriz con elementos en el intervalo  $[0, 1]$ . Cada posición de la matriz indica una muestra o píxel de la imagen. Lo que hace la transformación  $\mathcal{B}$  es asignar a cada píxel un promedio ponderado de los valores de los píxeles que se encuentran alrededor. Los pesos están dados por una función que dispersa el brillo de un píxel a sus vecinos. Apilamos las columnas de las matrices de las imágenes en vectores. De ese modo,  $\mathcal{B}$  puede verse como una transformación lineal sobre un espacio euclideo. Por lo que el problema de deblurring se reduce a resolver un sistema algebraico de ecuaciones lineales.

En el Capítulo 4 explicamos con más detalle el problema de Deblurring. Mientras tanto vemos que el sistema de ecuaciones que obtenemos está mal condicionado. Esto ocasiona

que la imagen obtenida tenga más degradación. Mediante la regularización de Tikhonov, tratamos de obtener una mejor imagen.

### Observaciones 3.9:

 Aclaremos que no hay un consenso general para la traducción al español del término deblurring.

**Ejemplo 3.5.** Considere la Imagen desenfocada con ruido gaussiano  $\epsilon$  de media cero y varianza 0.01 del *Alma Mater* de la Universidad de la Habana de  $m \times n$  píxeles en formato JPG que mostramos en la Figura 3.9 con  $m = 1600$  y  $n = 1200$ . Los valores de los píxeles de la imagen desenfocada están en una matriz  $G$  de tamaño  $1600 \times 1200$ . Queremos hallar la matriz  $F$  del mismo tamaño que nos da la imagen restaurada.



**Figura 3.9:** Imagen difuminada con desenfoco gaussiano de desviación estándar 3.

En nuestro caso, la función que ocasiona el desenfoco es la gaussiana de media cero con desviación estándar 3

$$k : \{-m/2, \dots, m/2\} \times \{-n/2, \dots, n/2\} \rightarrow \mathbb{R}$$

$$k_{i,j} = \frac{1}{3\sqrt{2\pi}} \exp\left(-\frac{i^2 + j^2}{2(3)^2}\right)$$

En el modelo que consideramos, el valor de cada píxel de la imagen difuminada es una suma ponderada de los valores de los píxeles dentro de una región de la imagen original que rodea al píxel correspondiente:

$$g_{i,j} = \sum_{p=-(m/2)}^{m/2} \sum_{q=-n/2}^{n/2} k_{p,q} f_{i-p,j-q}, \quad \begin{array}{l} i = 1, \dots, m, \\ j = 1, \dots, n. \end{array}$$

Los valores de  $f_{i,j}$  que hagan falta se toman iguales a cero. Denotemos por  $\mathbf{x}$  y  $\mathbf{b}$  a los vectores de  $(1600)(1200)$  componentes con las columnas apiladas de  $F$  y  $G$ , respectivamente. Obtenemos el sistema de ecuaciones lineales

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

donde

$$A = \begin{bmatrix} A^{(0)} & A^{(-1)} & \cdots & \cdots & A^{(-n/2)} & \mathbf{0}_{m \times m} & \cdots & \mathbf{0}_{m \times m} \\ \vdots & \ddots & \ddots & & & & \ddots & \vdots \\ \vdots & & \ddots & \ddots & & & \ddots & \mathbf{0}_{m \times m} \\ A^{(n/2-1)} & & & \ddots & \ddots & & & A^{(-n/2)} \\ A^{(n/2)} & \ddots & & & \ddots & \ddots & & \vdots \\ \mathbf{0}_{m \times m} & \ddots & \ddots & & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & \ddots & A^{(-1)} \\ \mathbf{0}_{m \times m} & \cdots & \mathbf{0}_{m \times m} & A^{(n/2)} & A^{(n/2-1)} & \cdots & \cdots & A^{(0)} \end{bmatrix}$$

es una matriz de  $n \times n$  bloques con una estructura en banda tal que los bloques sobre una misma diagonal son iguales, y a su vez cada bloque

$$A^{(j)} = \begin{bmatrix} k_{0,j} & k_{-1,j} & \cdots & k_{-m/2,j} & \mathbf{0} \\ \vdots & \ddots & \ddots & & \ddots \\ k_{m/2-1,j} & & \ddots & \ddots & k_{-m/2,j} \\ k_{m/2,j} & \ddots & & \ddots & \vdots \\ \mathbf{0} & \ddots & \ddots & \ddots & k_{-1,j} \\ \mathbf{0} & & k_{m/2,j} & k_{m/2-1,j} & \cdots & k_{0,j} \end{bmatrix}$$

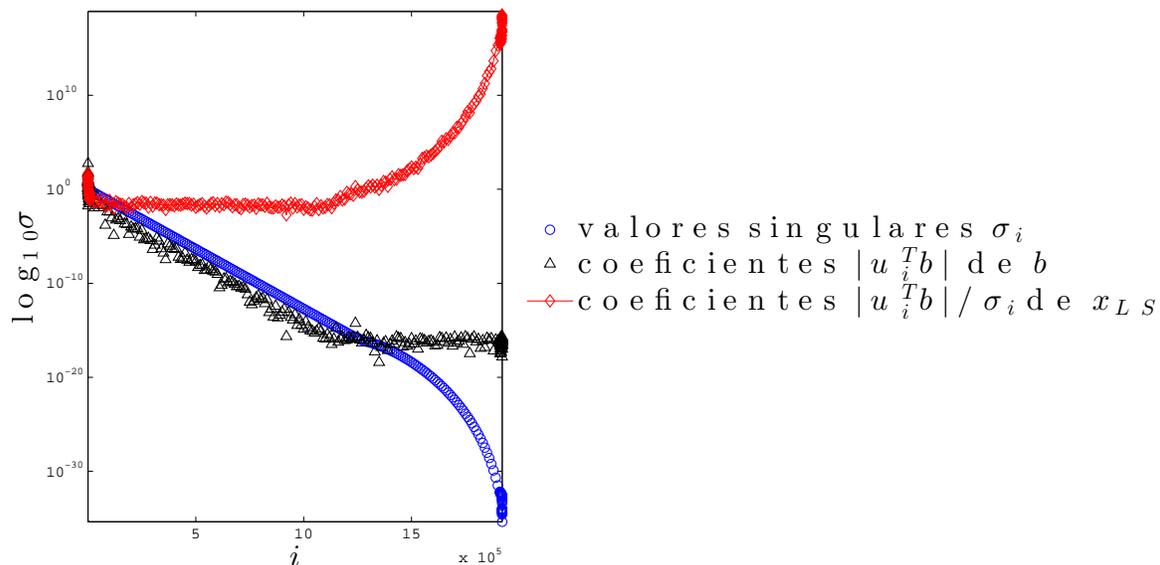
es una matriz de tamaño  $m \times m$  con la misma estructura en banda.



**Figura 3.10:** Imagen obtenida por solución de cuadrados mínimos.

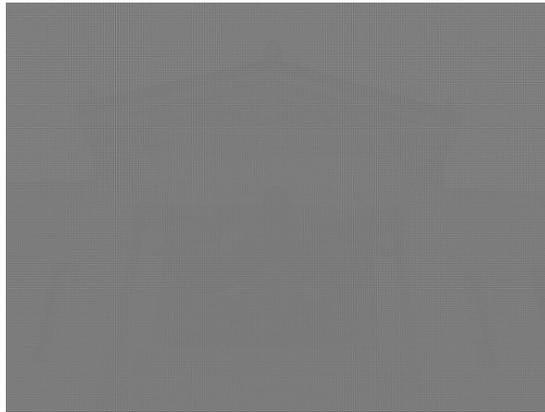
En §4.1 explicamos como reducir las dimensiones del problema para obtener la solución de cuadrados mínimos de norma mínima  $\mathbf{x}_{LS}$  de la Ecuación  $A\mathbf{x} = \mathbf{b}$  que mostramos en la Figura 3.10. En este momento nos interesa observar que los errores de redondeo en el cálculo de  $\mathbf{x}_{LS}$  ocasionan que la imagen se distorciona.

En la Gráfica de Picard 3.11 vemos como los valores singulares de  $A$  en orden decreciente decaen de  $\sigma_1 = 0.99995$  a  $\sigma_{(1600)(1200)} = 4.40426 \times 10^{-36}$ , por lo que  $\kappa_2(A) = 2.27042 \times 10^{35}$ , mientras que la poligonal con las medias geométricas móviles de los coeficientes  $|\mathbf{u}_i^T(\mathbf{b})|/\sigma_i$  de  $\mathbf{x}_{LS}$  oscila, y decrece en los primeros mil subíndices al orden de  $10^{-3}$ , se mantiene en ese orden hasta  $i = 10^5$  para crecer al orden de  $10^{18}$  en los últimos subíndices.



**Figura 3.11:** Gráfica de Picard para el problema de Deblurring  $A\mathbf{x} = \mathbf{b}$  con la imagen del *Alma Mater*. Mostramos en escala logarítmica los valores singulares de  $A$  en orden decreciente, los coeficientes  $|\mathbf{u}_i^T(\mathbf{b})|$  de  $\mathbf{b}$ , y la poligonal, marcada por (◇), con las medias geométricas de los coeficientes  $|\mathbf{u}_i^T(\mathbf{b})|/\sigma_i$  de la solución de norma mínima  $\mathbf{x}_{LS}$ . Dejamos un espacio de  $10^4$  entre los subíndices que mostramos, salvo en los primeros y últimos, donde dejamos un espacio de 200.

Usamos regularización de Tikhonov para obtener un vector  $\mathbf{x}_\lambda$  con las columnas apiladas de la matriz  $F_\lambda$  de la imagen restaurada. En la Figura 3.12(a), obtuvimos una imagen dominada por el ruido que generan los errores de redondeo con el parámetro de regularización  $\lambda = 10^{-17}$ . Si usamos  $\lambda = 10^{-15}$ , podemos apreciar características de la escena, aunque, todavía presenta artificios generados como puede verse en Figura 3.12(b). Cuando tomamos  $\lambda = 10^{-12}$ , en la imagen de la Figura 3.12(c) quitamos los artificios, podemos apreciar con más claridad detalles de la imagen. De seguir aumentado el parámetro hasta  $\lambda = 0.1$ , la imagen presenta más degradaciones. Este es el caso de la Figura 3.12(d). Así que conforme  $\lambda$  se acerca a cero, el ruido se propaga en la imagen restaurada por regularización de Tikhonov, mientras que si aumenta más de lo debido, la imagen se desenfoca.

(a)  $\lambda = 10^{-17}$ (b)  $\lambda = 10^{-15}$ (c)  $\lambda = 10^{-12}$ (d)  $\lambda = 0.1$ **Figura 3.12:** Imágenes obtenidas con la solución regularizadora de Tikhonov  $f_\lambda$ .

### 3.3.2. Aplicación en el Ajuste de Curvas

En problemas reales, los datos no aparecen en forma exacta, sino que están contaminados por ruido. Desde un punto de vista determinista, ese ruido se considera como una perturbación no deseada que aparece en la adquisición de datos. Por lo que no conviene interpolarlos, sino aproximarlos.

Si el problema es generar datos perturbados a partir de una función suave, el inverso es:

Dada una colección de observaciones con perturbaciones, reconstruir la curva de una función suave que se ajusta a éstas.

El inconveniente de resolver esto directamente es que la curva obtenida tiene altas oscilaciones aún con perturbaciones pequeñas en las observaciones. Por eso regularizamos.

Una idea que apoya a la teoría de regularización es que la solución regularizadora de un problema mal planteado puede obtenerse mediante un principio variacional que hace

uso de los datos, así como de la información a priori de la suavidad de la solución. Varias aplicaciones de la regularización variacional están influenciadas por los aportes de David Marr y James Gibson a la teoría de visión artificial [83].

Buscamos una familia de problemas que además de estar mejor condicionados, tengan soluciones más suaves y más cercanas a los datos suministrados. En estos problemas minimizamos una combinación lineal de funcionales  $E$  e  $I$  sobre un espacio de funciones suaves. Las discrepancias aparecen en  $E$ , mientras que  $I$  nos da la suavidad. Para controlar la suavidad usamos el parámetro de regularización  $\lambda$ .

Hemos visto que para medir la suavidad de una función, podemos usar la integral de su segunda derivada. Si  $I$  está dado de esta manera, podemos relacionar la solución regularizadora con la teoría de Splines.

En el siguiente ejemplo mostramos el enfoque variacional de la regularización de Tikhonov para reconstruir una curva suave que ajusta a unos datos.

**Ejemplo 3.6 (Spline de suavizamiento [27]).** En la Figura 3.13 mostramos las temperaturas diarias de la ciudad de Montreal del 1 de Enero de 1961 al 31 de Diciembre de 1962. Disponemos de  $m = 730$  datos  $(x_i, y_i)$ ,  $i = 1, \dots, m$  tomados de [98], donde  $x_i$  es el número de día e  $y_i$  es la temperatura en grados Celsius.

**Problema:** Encontrar la función más suave sobre el intervalo  $[x_1, x_m]$  de modo que las discrepancias entre sus valores y las temperaturas suministradas sean pequeñas.

Damos las medidas de suavidad y discrepancias mediante funcionales. Para medir las discrepancias empleamos el funcional  $E : C^2[x_1, x_m] \rightarrow \mathbb{R}^+$  dado por

$$E[f] := \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2$$

y para medir la suavidad de una función usamos el funcional  $I : C^2[x_1, x_m] \rightarrow \mathbb{R}^+$  dado por

$$I[f] := \int_{x_1}^{x_m} [f''(x)]^2 dx.$$

El funcional  $E$  asocia a la función  $f$  el promedio de los cuadrados de sus discrepancias con los datos, mientras que el funcional  $I$  nos da la curvatura de  $f$ .

Si resolvemos

$$\min_{f \in C^2[x_1, x_m]} I[f],$$

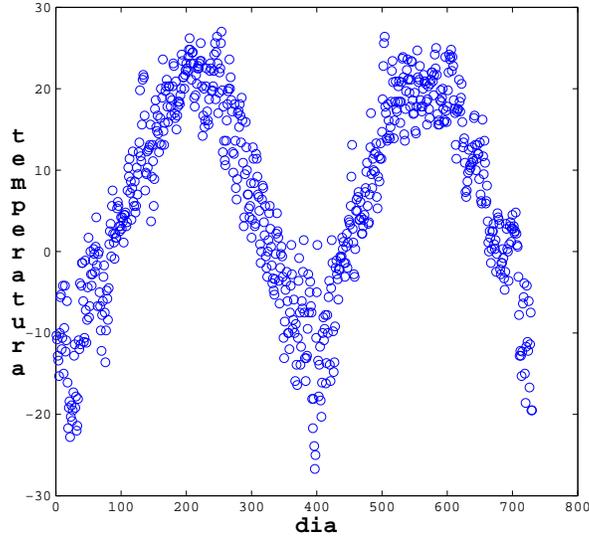
obtenemos una función suave de  $C^2[x_1, x_m]$ . Entre las funciones de este espacio vectorial, se puede probar que el polinomio cúbico por tramos  $S$  que cumple las condiciones de

frontera

$$S''(x_1) = S''(x_m) = 0$$

e interpola los datos  $(x_i, y_i)$ ,  $i = 1, \dots, m$ , minimiza el funcional  $I$ :

$$I[S] \leq I[f] \quad \forall f \in C^2[x_1, x_m] \quad (3.9)$$



**Figura 3.13:** Temperaturas en grados Celsius de Montreal en el período 1961-1962

El polinomio por tramos  $S$  se conoce como *spline cúbico natural* sobre el intervalo  $[x_1, x_m]$ . La Desigualdad (3.9) es la *propiedad de norma mínima del spline cúbico natural* [5]. Físicamente, podemos pensar un spline como una varilla de acero flexible que pasa por unos nodos de modo que minimiza la energía de tensión.

Como deseamos que los valores de nuestra función tengan pequeñas discrepancias con las temperaturas, nos restringimos a las funciones de  $f \in C^2[x_1, x_m]$  tales que  $E[f] = 0$ . Por lo que el problema es

$$\min_{\substack{f \in C^2[x_1, x_m] \\ E[f]=0}} I[f].$$

En la práctica, damos una tolerancia  $\sigma > 0$  para las discrepancias. Así que resolvemos

$$\min_{\substack{f \in C^2[x_1, x_m] \\ E[f] \leq \sigma}} I[f].$$

Lo que hacemos es introducir un parámetro  $\lambda > 0$  que compense la aproximación de los datos con la suavidad de la función mediante el funcional

$$f \mapsto E[f] + \lambda^2 I[f].$$

Por consiguiente, nuestro problema es:

$$\min_{f \in C^2[x_1, x_m]} \{E[f] + \lambda^2 I[f]\}. \quad (3.10)$$

Si minimizamos el funcional  $E$  sobre  $C^2[x_1, x_m]$ , obtenemos la recta que mejor se ajusta a los datos. En cambio, si minimizamos el funcional  $I$ , conseguimos un spline cúbico que interpola los datos. Se puede probar que la solución del Problema (3.10) es nuevamente un spline cúbico natural. Sin embargo, este spline no interpola los datos  $(x_i, y_i)$ , sino que es una aproximación suave de los datos, que en la literatura se conoce como **spline de suavizamiento** [120]. Para valores pequeños de  $\lambda$  aproximamos datos con altas frecuencias, mientras que para valores más grandes damos mayor peso al suavizamiento.

Para obtener el spline cúbico que resuelve el Problema (3.10) usamos una base de splines cúbicos. Dividimos el intervalo  $[x_1, x_m]$  en  $n'$  subintervalos  $[t_0, t_1), \dots, [t_{n'-1}, t_n)$ , donde definimos

$$B_j^0(x) = \begin{cases} 1, & \text{si } t_j \leq x < t_{j+1}, \\ 0, & \text{en otro caso,} \end{cases}$$

y de manera recursiva

$$B_j^k(x) = \frac{x - t_j}{t_{j+k} - t_j} B_j^{k-1}(x) + \left(1 - \frac{x - t_{j+1}}{t_{j+k+1} - t_{j+1}}\right) B_{j+1}^{k-1}(x), \quad k = 1, 2, 3.$$

Estas funciones se llaman **B-splines**. En la Figura 3.14 mostramos los B-splines  $B_i^k$  para  $k = 0, 1, 2, 3$ . Conforme  $k$  aumenta, la suavidad del spline aumenta y se define sobre intervalos cada vez más grandes.

En particular, los B-splines

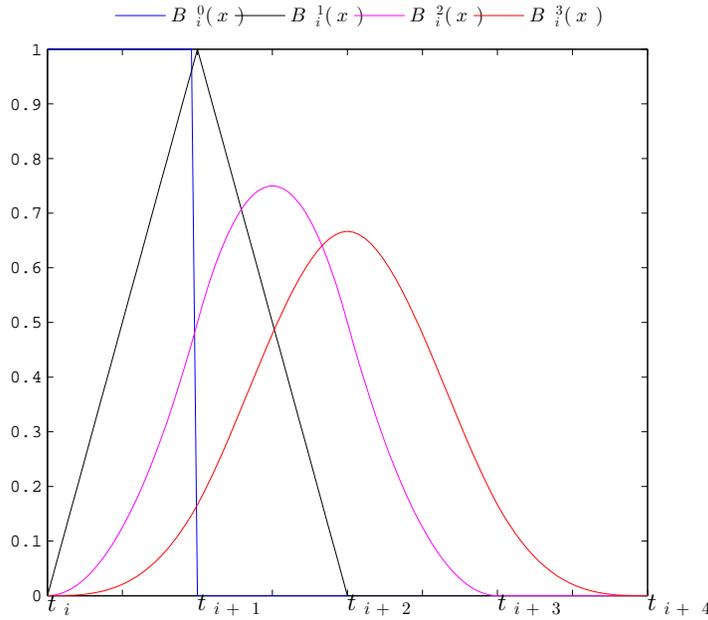
$$B_{-3}^3, \dots, B_{n'-1}^3$$

tienen las siguientes características [70]:

- \* Son funciones dos veces continuamente diferenciable sobre el intervalo  $[x_1, x_m]$  tales que sus evaluaciones, pendientes y segunda derivadas fuera del intervalo  $[t_j, t_{j+k})$  son iguales cero.
- \* Forman una base para el conjunto de splines cúbicos definidos sobre el intervalo  $[x_1, x_m]$ .

Tomamos nodos equidistantes  $t_i$ . Sea  $n = n' + 3$ , y sea  $S_\lambda$  el spline cúbico que resuelve el Problema (3.10). Expandimos  $S_\lambda$  como la combinación lineal

$$S_\lambda(x) = \sum_{j=1}^n a_j B_{j-4}^3(x).$$



**Figura 3.14:** B-splines  $B_i^0, B_i^1, B_i^2, B_i^3$ .

Para determinar los coeficientes  $a_j$ , reemplazamos  $f$  por  $S_\lambda$  en el problema

$$\min_{f \in C^2[x_1, x_m]} \left[ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda^2 \int_{x_1}^{x_m} [f''(x)]^2 dx \right].$$

De esa manera, resolvemos

$$\min_{\mathbf{a} \in \mathbb{R}^n} \left[ \frac{1}{m} \sum_{i=1}^m \left( \sum_{j=1}^n a_j B_{j-4}^3(x_i) - y_i \right)^2 + \lambda^2 \int_{x_1}^{x_m} \left[ \sum_{j=1}^n a_j \frac{d^2}{dx^2} B_{j-4}^3(x) \right]^2 dx \right]. \quad (3.11)$$

Eilers y Marx [27] prueban que

$$\int_{x_1}^{x_m} \left[ \sum_{j=1}^n a_j \frac{d^2}{dx^2} B_{j-4}^3(x) \right]^2 dx = 2 \sum_{j=3}^n c_j (a_j - 2a_{j-1} + a_{j-2})(a_{j-1} - 2a_{j-2} + a_{j-3}), \\ + \sum_{j=3}^n d_j (a_j - 2a_{j-1} + a_{j-2})^2,$$

donde las constantes  $c_j$  y  $d_j$  son integrales de productos de B-splines de grado uno en el

intervalo  $[x_1, x_m]$ . Ellos proponen resolver

$$\min_{\mathbf{a} \in \mathbb{R}^n} \left[ \frac{1}{m} \sum_{i=1}^m \left( \sum_{j=1}^n a_j B_{j-4}^3(x_i) - y_i \right)^2 + \lambda^2 \sum_{j=3}^n (a_j - 2a_{j-1} + a_{j-2})^2 \right].$$

en lugar del problema (3.11). En forma matricial, esto es,

$$\min_{\mathbf{a} \in \mathbb{R}^n} \frac{1}{m} \|\mathbf{B}\mathbf{a} - \mathbf{y}\|_2^2 + \lambda^2 \|\mathbf{L}\mathbf{a}\|_2^2,$$

donde

$$b_{i,j} = B_{j-4}^3(x_i), \quad i = 1, \dots, m, \quad j = 1, \dots, n.$$

y

$$L = \begin{bmatrix} 1 & -2 & 1 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & 1 & -2 & 1 \end{bmatrix}_{(n-2) \times n}$$

es la discretización por diferencias finitas centradas del operador que nos da la segunda derivada.

Identificamos el método de regularización de Tikhonov. En este problema, el parámetro de regularización  $\lambda$  compensa el error de la aproximación  $\|\mathbf{B}\mathbf{a} - \mathbf{y}\|_2^2$  con la suavidad de la solución  $\|\mathbf{L}\mathbf{a}\|_2^2$ . Las ecuaciones normales regularizadas

$$(\mathbf{B}^T \mathbf{B} + m\lambda^2 \mathbf{L}^T \mathbf{L})\mathbf{a} = \mathbf{B}^T \mathbf{y}.$$

nos dan los puntos críticos de la función objetivo

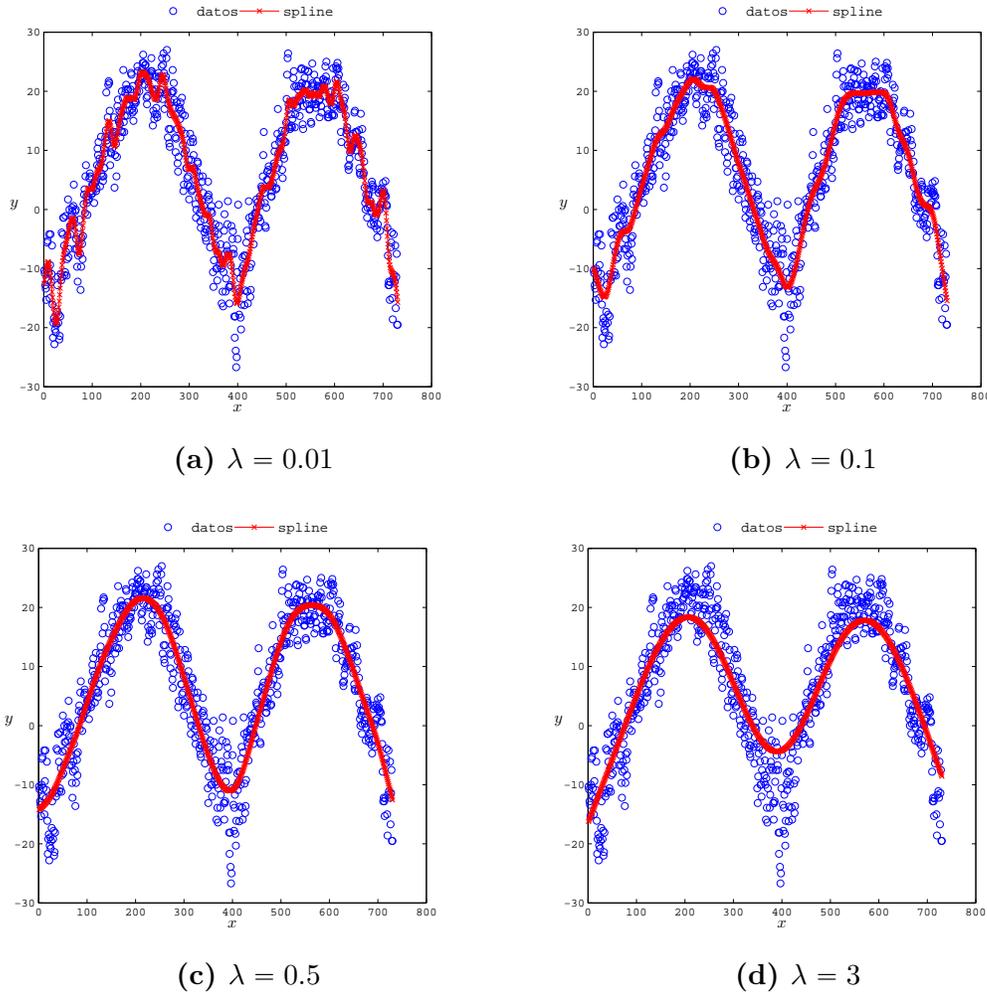
$$J(\mathbf{a}, \lambda) = \frac{1}{m} \|\mathbf{B}\mathbf{a} - \mathbf{y}\|_2^2 + \lambda^2 \|\mathbf{L}\mathbf{a}\|_2^2.$$

Como la matriz  $\mathbf{B}^T \mathbf{B} + \lambda^2 \mathbf{L}^T \mathbf{L}$  de este ejemplo es positiva definida, la solución regularizadora es

$$\mathbf{a}_\lambda = (\mathbf{B}^T \mathbf{B} + m\lambda^2 \mathbf{L}^T \mathbf{L})^{-1} \mathbf{B}^T \mathbf{y}.$$

El vector con las evaluaciones del spline cúbico  $S_\lambda$  en  $x_1, \dots, x_m$  es el producto  $\mathbf{B}\mathbf{a}_\lambda$ .

En la Figura 3.15 comparamos los datos  $(x_i, y_i)$  con la gráfica del spline  $S_\lambda$  para distintos valores del parámetro  $\lambda$ . Observamos que para valores pequeños de  $\lambda$ , el spline  $S_\lambda$  trata de interpolar los datos, mientras para  $\lambda = 0.5$ , el spline nos da una buena aproximación de los datos. Incrementar el valor de  $\lambda$ , nos conduce a una solución más suave a costa de aumentar el error en la aproximación.



**Figura 3.15:** Spline cúbico  $S_\lambda$  obtenido con regularización de Tikhonov

### 3.3.3. Aplicación en Regresión Múltiple

Una aplicación de la regularización de Tikhonov en la Estadística ocurre en modelos de regresión múltiple que presentan problemas de colinealidad.

En ocasiones, las variables predictivas de un modelo de regresión múltiple están fuertemente correlacionadas entre ellas. Cuando esto ocurre, los valores estimados de los coeficientes de regresión son muy sensibles a cambios pequeños en los datos o al quitar o agregar variables en la regresión, lo que afecta las inferencias y predicciones basadas en el modelo de regresión. En esta situación, ya no es válido interpretar un coeficiente de regresión como una medida del cambio en la variable explicativa cuando la correspondiente variable predictiva aumenta una unidad, mientras que las otras quedan fijas.

La existencia de una fuerte correlación lineal entre las variables predictoras del modelo de regresión se conoce como *Colinealidad*. Este problema puede que no se deba a un

error de modelación, sino a datos deficientes [18].

Para el estadístico, la presencia de colinealidad en un modelo de regresión infla las varianzas de los coeficientes de regresión y amplifica los errores en las variables de regresión. Una vez escaladas y centradas las variables predictorias, podemos hacer un diagnóstico de colinealidad con medidas como correlaciones y los factores de inflación de varianza. Véase Marquardt [77]. Por su parte, el análista numérico puede usar los valores singulares de la matriz de diseño y su número de condición para detectar colinealidad, además de emplear cotas de perturbación como las de §2.5 para predecir errores en los coeficientes de regresión. Sin embargo, esas cotas de perturbación en términos del número de condición son pesimistas en aplicaciones estadísticas como señala Stewart en [112], donde da una explicación clara de la conexión entre el número de condición y los factores de inflación de varianza y propone como medida de colinealidad a los cuadrados de los factores de inflación de varianza.

La presencia de pequeños valores propios en el producto de la matriz de diseño con su transpuesta indica un problema de colinealidad en el modelo de regresión. Cuando uno o más de estos valores propios son pequeños, el error en media cuadrática es grande, lo que sugiere una estimación de cuadrados mínimos imprecisa. Para abordar el problema de colinealidad, buscamos un estimador con componentes de menor variación que las del estimador de cuadrados mínimos. Una manera de hacer esto es aproximar el conjunto original de variables explicativas por uno ortogonal que permita un sesgo pequeño en el estimador del vector parámetros. Esto se conoce como *Ridge Regression*. El nombre fue dado por Arthur E. Hoerl [60] por parecido con un método gráfico que él había propuesto anteriormente.

Mediante la ridge regression tratamos de dar un estimador alternativo con error en media cuadrática más pequeño que el estimador de cuadrados mínimos. Lo que hacemos es poner una restricción cuadrática al método de cuadrados mínimos para permitir un estimador sesgado. Penalizamos con un multiplicador de Lagrange, de modo que la varianza del nuevo estimador sea una función decreciente respecto al multiplicador, mientras que su sesgo sea una función creciente [58].

La introducción del multiplicador en el método de cuadrados mínimos con restricción cuadrática, nos permite relacionar la ridge regression con la regularización de Tikhonov. como vemos en el siguiente ejemplo.

**Ejemplo 3.7 (Regresión múltiple [86]).** En la Tabla 3.2 mostramos información sobre la mano de obra de  $n = 17$  hospitales navales de E.U.A. [118], donde

$Y$	horas laborales mensuales
$X_1$	promedio diario del número de pacientes
$X_2$	número de exposiciones a rayos X en un mes
$X_3$	total de días-camas ocupadas en un mes
$X_4$	población en el área dada $\div 1000$
$X_5$	promedio de la permanencia de pacientes en días

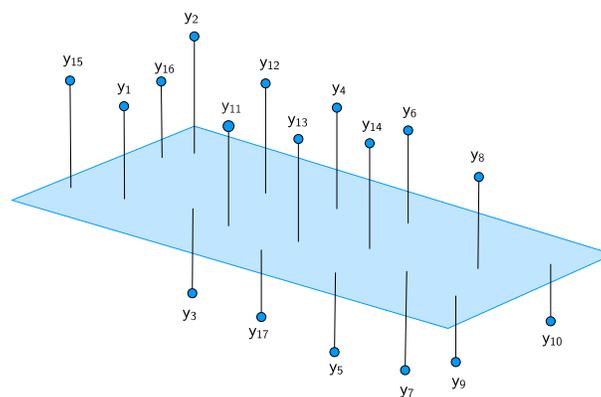
$i$	$Y$	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
1	566.52	15.57	2463	472.92	18	4.45
2	696.82	44.02	2048	1339.75	9.5	6.92
3	1033.15	20.42	3940	620.25	12.8	4.28
4	1603.62	18.74	6505	568.33	36.7	3.9
5	1611.37	49.2	5723	1497.6	35.7	5.5
6	1613.27	44.92	11520	1365.83	24	4.6
7	1854.17	55.48	5779	1687	43.3	5.62
8	2160.55	59.28	5969	1639.92	46.7	5.15
9	2305.58	94.39	8461	2872.33	78.7	6.18
10	3503.93	128.02	20106	3655.08	180.5	6.15
11	3571.89	96	13313	2912	60.9	5.88
12	3741.4	131.42	10771	3921	103.7	4.88
13	4026.52	127.21	15543	3865.67	126.8	5.5
14	10343.81	252.9	36194	7684.1	157.7	7
15	11732.17	409.2	34703	12446.33	169.4	10.78
16	15414.94	463.7	39204	14098.4	331.4	7.05
17	18854.45	510.22	86533	15524	371.6	6.35

**Tabla 3.2:** Datos de mano de obra del Hospital.

Vamos a obtener una ecuación que estime las horas laborales mensuales a partir de las otras variables. Para ello usamos un modelo de regresión multilínea donde el vector de observaciones  $Y$  depende de múltiples variables explicativas  $X_1, \dots, X_5$ :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5.$$

Geoméricamente, tratamos de ajustar nuestras observaciones con el hiperplano generado por las variables explicativas.



$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_5 X_5$$

**Figura 3.16:** Ajuste de observaciones con el hiperplano generado por las variables explicativas

A partir de la Regresión Multilineal, obtenemos el sistema de ecuaciones lineales

$$Y = X\boldsymbol{\beta},$$

donde

$$X = \left[ \begin{array}{c|c|c|c|c|c} 1 & & & & & \\ \vdots & X_1 & X_2 & X_3 & X_4 & X_5 \\ 1 & & & & & \end{array} \right]_{17 \times 6}.$$

Aplicamos el método de cuadrados mínimos. Por lo que resolvemos

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{17}} \|X\boldsymbol{\beta} - Y\|_2^2.$$

Como la matriz  $X$  tiene rango completo por columnas, entonces la matriz simétrica  $X^T X$  es positiva definida. Mediante su factorización de Cholesky, tenemos que la solución de las ecuaciones normales

$$X^T X\boldsymbol{\beta} = X^T Y$$

es

$$\boldsymbol{\beta}^\dagger = [1962.948 \quad -15.8517 \quad 0.05593 \quad 1.58962 \quad -4.21 \quad -394.314]^T.$$

Supongamos que nuestro vector de observaciones  $Y$  tiene un error aleatorio aditivo  $\boldsymbol{\epsilon}$  de componentes idénticamente distribuidas por una función gaussiana con  $E(\boldsymbol{\epsilon}) = 0$  y  $\text{Cov}(\boldsymbol{\epsilon}) = 0.01I$ . Por el Teorema de Gauss-Markov, sabemos que el estimador de cuadrados mínimos  $\boldsymbol{\beta}_{LS}$  es el estimador lineal insesgado de varianza mínima para el problema de regresión lineal

$$Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

¿Podemos encontrar un estimador lineal de  $\boldsymbol{\beta}^\dagger$  que tenga un sesgo pequeño y varianza menor en relación a la del estimador insesgado  $\boldsymbol{\beta}_{LS}$ ?

Lo que hacemos es buscar estimadores pequeños  $\hat{\boldsymbol{\beta}}$  de  $\boldsymbol{\beta}^\dagger$  que tengan un residuo

$$\mathbf{r}_{\hat{\boldsymbol{\beta}}} = Y - X\hat{\boldsymbol{\beta}}$$

de tamaño fijo. Al respecto, Hoerl y Kennard prueban en [60] que con el residuo

$$\mathbf{r}_{LS} = Y - X\boldsymbol{\beta}_{LS},$$

el tamaño de  $\mathbf{r}_{\hat{\boldsymbol{\beta}}}$  es

$$\|\mathbf{r}_{\hat{\boldsymbol{\beta}}}\|_2^2 = \|\mathbf{r}_{LS}\|_2^2 + \|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{LS})\|_2^2.$$

Así, todo estimador  $\hat{\beta}$  de  $\beta^\dagger$  con residuo de tamaño fijo  $r > \|\mathbf{r}_{LS}\|_2$ , se encuentra en el hiperlipsoide

$$\|X(\hat{\beta} - \beta_{LS})\|_2^2 = c_r,$$

donde

$$c_r^2 = r^2 - \|\mathbf{r}_{LS}\|_2^2.$$

Esto da lugar al problema de cuadrados mínimos con restricción cuadrática:

$$\min_{\|X(\hat{\beta} - \beta_{LS})\|_2^2 = c_r^2} \|\hat{\beta}\|_2^2.$$

Con el método de multiplicadores de Lagrange, penalizamos este problema con un parámetro  $\lambda \neq 0$ . Por lo que minimizamos el Lagrangiano

$$L(\hat{\beta}, \lambda) = \hat{\beta}^T \hat{\beta} + \lambda \left( c_r^2 - \|X(\hat{\beta} - \beta_{LS})\|_2^2 \right)$$

respecto al estimador  $\hat{\beta}$ . Los puntos críticos de  $L$  satisfacen

$$\frac{\partial L}{\partial \hat{\beta}} = \mathbf{0} \quad \text{y} \quad \frac{\partial L}{\partial \lambda} = 0.$$

Dado que

$$\frac{\partial L}{\partial \hat{\beta}} = 2\hat{\beta} - 2\lambda X^T X(\hat{\beta} - \beta_{LS}),$$

los puntos críticos del Lagrangiano cumplen la ecuación

$$\hat{\beta} - \lambda X^T X \hat{\beta} = -\lambda X^T X \beta_{LS}.$$

Dado que

$$X^T X \beta_{LS} = X^T Y,$$

se sigue que

$$\left( \frac{1}{\lambda} I - X^T X \right) \hat{\beta} = -X^T Y.$$

Si reemplazamos  $\lambda$  por  $-1/\lambda^2$ , entonces el mínimo de  $L$  cumple las ecuaciones normales regularizadas

$$(X^T X + \lambda^2 I) \hat{\beta} = X^T Y.$$

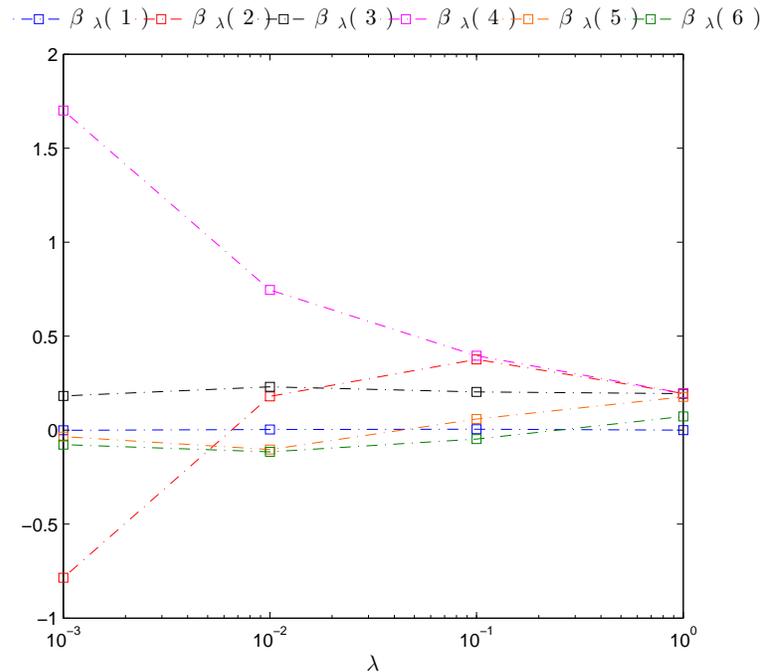
De aquí, el estimador buscado es la solución regularizada  $\hat{\beta} = \beta_\lambda$  y su sesgo respecto a  $\beta^\dagger$  con el vector de observaciones ideales  $Y_{\text{exacto}}$  se expresa mediante la SVD  $X = U \Sigma V^T$  como

$$\|\beta^\dagger - E(\beta_\lambda)\|_2^2 = \sum_{i=1}^n \left( \frac{\lambda^2}{\sigma_i^2 + \lambda^2} \right)^2 \left( \frac{\mathbf{u}_i^T Y_{\text{exacto}}}{\sigma_i} \right)^2$$

Adicionalmente, se puede verificar que

$$c_r^2 = \sum_{i=1}^n \left( \frac{\lambda^2}{\sigma_i^2 + \lambda^2} \right)^2 (\mathbf{u}_i^T (Y_{\text{exacto}} + \boldsymbol{\epsilon}))^2. \quad (3.12)$$

De esta manera modificamos el problema de cuadrados mínimos para permitir una estimación sesgada. Por lo que estamos aplicando ridge regression. En la Figura 3.17 graficamos las componentes  $\beta_\lambda(1), \dots, \beta_\lambda(6)$  del estimador  $\beta_\lambda$  contra el parámetro  $\lambda$  cuando este parámetro toma los valores  $1, 10^{-1}, 10^{-2}, 10^{-3}$ . Esta gráfica se llama *Ridge-Plot* [63]. Observamos que  $\beta_\lambda(4)$  y  $\beta_\lambda(5)$  no se van a cero cuando  $\lambda = 10^{-3}$ .



**Figura 3.17:** Componentes  $\beta_\lambda(1), \dots, \beta_\lambda(5)$  de estimador de ridge regression en función del parámetro  $\lambda = 1, 10^{-1}, 10^{-2}, 10^{-3}$ .

La introducción del sesgo en nuestro estimador se compensa con una reducción en la varianza. Esta compensación se mide con el error en media cuadrática. Por el Teorema 3.1, tenemos que  $\beta_{LS}$  es

$$\text{MSE}(\beta_{LS}) = \text{Tr}(\text{Cov}(\beta_{LS})) = \eta^2 \sum_{i=1}^6 \frac{1}{\sigma_i^2},$$

En este ejemplo,  $\text{MSE}(\beta_{LS}) = 185.66$ .

Por otra parte, por la Proposición 3.1, tenemos que el estimador sesgado  $\beta_\lambda$  es

$$\text{MSE}(\beta_\lambda) = \eta^2 \sum_{i=1}^6 \frac{\sigma_i^2}{(\sigma_i^2 + \lambda^2)^2} + \sum_{i=1}^6 \left( \frac{\lambda^2}{\sigma_i^2 + \lambda^2} \right)^2 \left( \frac{\mathbf{u}_i^T Y_{\text{exacto}}}{\sigma_i} \right)^2.$$

Comparamos los errores en media cuadrática para distintos valores del parámetro  $\lambda$ . Observamos en la Tabla 3.3 que reducimos el error en media cuadrática con el método de ridge regression. Entre  $\lambda$  se hace más pequeño, más grande se hace  $\text{MSE}(\beta_\lambda)$ .

$\lambda$	$\text{MSE}(\beta_\lambda)$
1	1.9906
0.1	1.9535
0.01	23.847
0.001	178.98

**Tabla 3.3:** Errores en media cuadrática del estimador  $\beta_\lambda$

### 3.3.4. Extensiones de la Regularización de Tikhonov

Considere el problema mal planteado dado por la Ecuación Integral de Fredholm

$$\int_a^b k(x, y) f(y) dy = g(x), \quad c \leq x \leq d.$$

Discretizamos esta ecuación con el método de colocación en un sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b}$ , donde  $A \in \mathbb{R}^{m \times n}$  y  $\mathbf{b} \in \mathbb{R}^m$ . Este sistema de ecuaciones lineales es un problema discreto mal planteado que regularizamos con el método de Tikhonov:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda^2 \|\mathbf{x}\|_2^2.$$

Con el término  $\|\mathbf{x}\|_2$  controlamos el tamaño de la solución regularizadora. Ahora, queremos controlar su suavidad. Para ello, vemos el tamaño de las derivadas de la función  $f$  en la norma-2 de  $L^2[a, b]$ . Si usamos una partición uniforme  $t_1 < \dots < t_n$  del intervalo  $[a, b]$  con incremento  $\Delta t$ , entonces

$$\int_a^b [f'(x)]^2 dx \approx \Delta t \sum_{i=1}^{n-1} [f'(t_i)]^2 \quad \text{y} \quad \int_a^b [f''(x)]^2 dx \approx \Delta t \sum_{i=2}^{n-1} [f''(t_i)]^2.$$

Mediante diferencias finitas aproximamos  $f'(t_i)$  y  $f''(t_i)$ . Por lo que

$$\int_a^b [f'(x)]^2 dx \approx \Delta t \sum_{i=1}^{n-1} \left[ \frac{f(t_{i+1}) - f(t_i)}{\Delta t} \right]^2 \quad \text{y} \quad \int_a^b [f''(x)]^2 dx \approx \Delta t \sum_{i=2}^{n-1} \left[ \frac{f(t_{i+1}) - 2f(t_i) + f(t_{i-1}))}{(\Delta t)^2} \right]^2.$$

De esta manera, con las matrices en banda

$$L_1 = \frac{1}{\Delta t} \begin{bmatrix} 1 & -1 & & & \mathbf{0} \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ \mathbf{0} & & & 1 & -1 \end{bmatrix}_{(n-1) \times n}, \quad L_2 = \frac{1}{(\Delta t)^3} \begin{bmatrix} 1 & -2 & 1 & & \mathbf{0} \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ \mathbf{0} & & & 1 & -2 & 1 \end{bmatrix}_{(n-2) \times n}$$

y el vector

$$\mathbf{x} = [f(t_1) \quad \cdots \quad f(t_n)]^T$$

podemos aproximar el tamaño de la primera y segunda derivada de  $f$  como

$$\|f'\|_2^2 \approx \|L_1 \mathbf{x}\|_2^2 \quad \text{y} \quad \|f''\|_2^2 \approx \|L_2 \mathbf{x}\|_2^2.$$

De modo que penalizamos el problema lineal de cuadrados mínimos con el término de suavizamiento  $\|L\mathbf{x}\|_2^2$ , en lugar del tamaño de la solución  $\|\mathbf{x}\|_2$ . Así que resolvemos

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda^2 \|L\mathbf{x}\|_2^2. \quad (3.13)$$

De este modo, la regularización de Tikhonov suaviza y amortigua las componentes de alta frecuencia. Los puntos críticos de la función objetivo del Problema (3.13) satisfacen las ecuaciones normales regularizadas

$$(A^T A + \lambda^2 L^T L)\mathbf{x} = A^T \mathbf{b}.$$

Para que estas ecuaciones tengan solución única  $\mathbf{x}_\lambda$ , supongamos que  $L$  tiene rango completo por renglones y que los espacios nulos de  $A$  y  $L$  se intersecan trivialmente.

Mediante cambios de variable podemos reducir el problema al caso donde penalizamos el tamaño de la solución. Par ver esto, notamos que el residual

$$\mathbf{r} = \begin{pmatrix} A \\ \lambda L \end{pmatrix} \mathbf{x} - \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}$$

tiene tamaño

$$\|\mathbf{r}\|_2^2 = \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda^2 \|L\mathbf{x}\|_2^2.$$

Lo que hacemos es buscar una matriz  $K$ , junto con vectores  $\mathbf{z}$  y  $\mathbf{c}$  tales que

$$\|\mathbf{r}\|_2^2 = \|K\mathbf{z} - \mathbf{c}\|_2^2 + \lambda^2 \|\mathbf{z}\|_2^2. \quad (3.14)$$

Comenzamos por calcular una factorización  $QR$  de  $L^T$ :

$$L^T = (V_{n \times p} \quad W_{n \times (n-p)}) \begin{pmatrix} R_{p \times p} \\ \mathbf{0}_{p \times (n-p)} \end{pmatrix}$$

A continuación, realizamos una Factorización  $QR$  de  $AW$ :

$$AW = (Q_{m \times (n-p)} \quad S_{m \times (m-n+p)}) \begin{pmatrix} U_{(n-p) \times (n-p)} \\ \mathbf{0}_{(m-n+p) \times (n-p)} \end{pmatrix}$$

Puesto que  $L$  tiene rango completo por renglones, las matrices  $R$  y  $U$  son invertibles. Tomamos

$$K = S^T A V R^{-T} \quad \text{y} \quad \mathbf{c} = S^T \mathbf{b}.$$

Entonces con los cambios de coordenadas

$$\mathbf{x} = V \mathbf{y}_1 + W \mathbf{y}_2 \quad \text{y} \quad \mathbf{z} = R^T \mathbf{y}_1$$

podemos verificar la relación (3.14). En consecuencia, el Problema (3.13) equivale a

$$\min_{\mathbf{z} \in \mathbb{R}^p} \{ \|\mathbf{Kz} - \mathbf{c}\|_2^2 + \lambda^2 \|\mathbf{z}\|_2^2 \}.$$

Denotamos al mínimo por  $\mathbf{z}_\lambda$ . Así,  $\mathbf{x}_\lambda$  se obtiene de la siguiente manera [30]:

$$\begin{aligned} \mathbf{y}_1 &= R^{-T} \mathbf{z}_\lambda, \\ \mathbf{y}_2 &= U^{-1} Q^T (\mathbf{b} - K V \mathbf{y}_1), \\ \mathbf{x}_\lambda &= V \mathbf{y}_1 + W \mathbf{y}_2. \end{aligned}$$

### Observaciones 3.10:

 Hasta ahora hemos manejado el tamaño de la solución en la regularización de Tikhonov con la norma euclidiana. En algunas aplicaciones conviene usar otras normas [54] como la norma-1 si queremos controlar la suavidad de una función con su variación total o con la norma Frobenius si la solución tiene las columnas apiladas de una matriz.

 Cuando  $L \neq I_{n \times n}$ , podemos emplear una generalización de la SVD para estudiar la regularización de Tikhonov como un método de regularización por factores filtro. Consulte el Apéndice.

En la regularización de Tikhonov, reemplazamos el problema mal planteado por otro mejor condicionado que incorpore información sobre la solución del problema. Bajo la hipótesis de que la solución sea suave, vimos que podemos incluir esta información en el problema de cuadrados mínimos de la Ecuación  $A\mathbf{x} = \mathbf{b}$  con una restricción  $\|L\mathbf{x}\|_2 \leq \Delta$ , donde  $L$  es una matriz que se obtiene de la discretización del tamaño de la primera o segunda derivada. Si buscamos solamente que el tamaño de la solución sea pequeño, podemos dar la misma restricción con  $L$  igual a la matriz identidad [99]. Esto da lugar a otra formulación de la regularización de Tikhonov donde debemos resolver un problema de cuadrados mínimos con restricciones cuadráticas:

$$\min_{\|L\mathbf{x}\|_2 \leq \Delta} \|A\mathbf{x} - \mathbf{b}\|_2^2.$$

Bajo cambios de variables, los reducimos a su forma estándar

$$\min_{\|\mathbf{x}\|_2 \leq \Delta} \|A\mathbf{x} - \mathbf{b}\|_2^2$$

conocido en optimización como *Subproblema de región de confianza* [22]. En la siguiente sección presentamos las características de este problema. Por lo mientras, vemos como podemos llevarlo a la formulación clásica de Tikhonov.

Con el método de multiplicadores de Lagrange, penalizamos el subproblema de región de confianza. De esa manera, incluimos la restricción cuadrática  $\|\mathbf{x}\|_2^2 \leq \Delta^2$  en la función objetivo dada por el tamaño del residuo  $\|A\mathbf{x} - \mathbf{b}\|_2^2$  con un multiplicador  $\mu > 0$  como

$$\min_{\mathbf{x} \in \mathbb{R}^n} \{ \|A\mathbf{x} - \mathbf{b}\|_2^2 + \mu(\|\mathbf{x}\|_2^2 - \Delta^2) \}.$$

Podemos omitir el término con  $\Delta^2$  en el problema penalizado porque minimizamos sobre el vector  $\mathbf{x}$ . Así que si tomamos  $\mu = \lambda^2$ , obtenemos el método de regularización de Tikhonov.

Otra manera de abordar el problema discreto mal planteado dado por la Ecuación  $A\mathbf{x} = \mathbf{b}$  es reducir el tamaño de la solución lo más que podamos y usar la información que tengamos sobre los errores en los datos. En este caso minimizamos  $\|\mathbf{x}\|_2$  y ponemos la restricción  $\|A\mathbf{x} - \mathbf{b}\|_2 \leq \Delta$ . Esto da lugar al problema de cuadrados mínimos con restricciones cuadráticas dado por

$$\min_{\|A\mathbf{x} - \mathbf{b}\|_2 \leq \Delta} \|\mathbf{x}\|_2^2.$$

Este es *el problema de la solución de norma mínima*, que consiste en buscar el vector más pequeño dentro o sobre el hiperelipsoide  $\|A\mathbf{x} - \mathbf{b}\|_2 = \Delta$ .

Nuevamente, con el método de multiplicadores de Lagrange podemos llevar el problema de la solución de la norma mínima a la formulación de Tikhonov. La diferencia es que cuando penalizamos, el multiplicador  $\mu$  acompaña al término  $\|A\mathbf{x} - \mathbf{b}\|_2^2 - \Delta^2$ , por lo que multiplicamos la función objetivo por  $1/\mu$  y hacemos  $\lambda^2 = 1/\mu$ .

## 3.4. El Subproblema de Región de Confianza

### 3.4.1. Método de Región de Confianza

Uno de los problemas de optimización sin restricciones es dada una función  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  acotada y dos veces continuamente diferenciable,

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

Lo que hacemos es generar una sucesión  $\{\mathbf{x}_k\}$  que converge a un mínimo  $\mathbf{x}_{\min}$  de  $f$ . A partir de una aproximación inicial  $\mathbf{x}_0$ , avanzamos un paso  $\mathbf{s}_k$  en cada iteración:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k.$$

Tomamos  $\mathbf{s}_k$  de modo que el tamaño del error  $f(\mathbf{x}_{\min}) - f(\mathbf{x}_k)$  en una vecindad de  $\mathbf{x}_k$  disminuya en cada iteración.

Una manera de elegir el paso  $\mathbf{s}_k$  de modo que la sucesión  $\{\mathbf{x}_k\}$  se aproxime al mínimo es mediante el *método de región de confianza* [22], [25]. La idea es construir un modelo aproximado  $\psi_k$  de la función  $f$  que en cada iteración sea válido en una vecindad de  $\mathbf{x}_k$  con radio  $\Delta_k$ , llamada región de confianza. Este método incorpora una estrategia local donde elegimos  $\mathbf{x}_{k+1}$  como un mínimo de  $\psi_k$  sobre la región de confianza y una estrategia global que tiene las siguientes características:

- \* En cada iteración, aceptamos ó rechazamos el paso. Para decidir, calculamos el cociente entre la reducción de la función objetivo y la reducción predecida por el modelo:

$$\rho_k = \frac{f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})}{\psi_k(\mathbf{0}) - \psi_k(\mathbf{s}_k)}.$$

Aceptamos el paso si  $\rho_k > 0$ , de otro modo lo rechazamos

- \* En cada iteración, reducimos el radio de la región de confianza cuando no tenemos una aproximación aceptable de  $\mathbf{x}_{k+1}$ . De tenerla, aumentamos el radio si  $\psi_k$  predice bien a la función  $f$ , lo disminuimos si la predicción de  $\psi_k$  es pobre, y lo dejamos igual en otro caso. Usamos  $\rho_k$  para medir la predicción.

### Observaciones 3.11:

 Para evitar que la región de confianza se expanda más de lo necesario, delimitamos su radio con el tamaño del gradiente:

$$\Delta_k \leq c \|\mathbf{g}_k\|_2, \quad c > 1.$$

Retomamos la estrategia local que incorpora el método de región de confianza. Sea  $H_k$  el Hessiano de  $f$  en  $\mathbf{x} = \mathbf{x}_k$ . Puesto que  $f \in \mathcal{C}^2(\mathbb{R}^n)$ , podemos expandir esta función en polinomios de Taylor de segundo orden como

$$f(\mathbf{x}) \approx f(\mathbf{x}_k) + \mathbf{g}_k^T (\mathbf{x} - \mathbf{x}_k) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_k)^T H_k (\mathbf{x} - \mathbf{x}_k)$$

para todo  $\mathbf{x}$  en la vecindad  $\|\mathbf{x} - \mathbf{x}_k\|_2 \leq \Delta_k$ . De esta manera, la Expansión de Taylor nos da el modelo cuadrático

$$\psi_k(\mathbf{s}) = \mathbf{g}_k^T \mathbf{s} + \frac{1}{2} \mathbf{s}^T H_k \mathbf{s}$$

en la región de confianza  $\|\mathbf{s}\| \leq \Delta_k$ . Con este modelo,

$$\psi_k(\mathbf{x} - \mathbf{x}_k) \approx f(\mathbf{x}) - f(\mathbf{x}_k) \quad \text{si} \quad \|\mathbf{x} - \mathbf{x}_k\|_2 \leq \Delta_k.$$

Tomamos el paso  $\mathbf{s}_k$  como el vector que minimiza el modelo  $\psi_k$  en la región de confianza. Esto da lugar al *Subproblema de Región de Confianza (TRS)*:

$$\min_{\substack{\mathbf{s} \in \mathbb{R}^n \\ \|\mathbf{s}\|_2 \leq \Delta_k}} \psi_k(\mathbf{s})$$

### 3.4.2. Características del TRS

En el resto de la sección trabajamos con el TRS. Por lo que omitimos el subíndice  $k$ . De modo que el modelo cuadrático es

$$\psi(\mathbf{s}) = \mathbf{g}^T \mathbf{s} + \frac{1}{2} \mathbf{s}^T H \mathbf{s}$$

y el TRS es

$$\min_{\substack{\mathbf{s} \in \mathbb{R}^n \\ \|\mathbf{s}\|_2 \leq \Delta}} \psi(\mathbf{s}).$$

#### Observaciones 3.12:

☞ El TRS tiene solución porque toda función continua (el modelo cuadrático) sobre un conjunto compacto (la restricción) alcanza sus óptimos.

☞ El hessiano  $H$  es una matriz simétrica porque  $f \in \mathcal{C}^2(\mathbb{R}^n)$ .

Podemos verificar que los puntos críticos del modelo cuadrático  $\psi$  sin restricciones cumplen la ecuación

$$H\mathbf{s} = -\mathbf{g}.$$

Cuando ponemos la restricción  $\|\mathbf{s}\|_2 \leq \Delta$  al modelo  $\psi$ , usamos un multiplicador de Lagrange  $\mu$  de modo que el mínimo de  $\psi$  en la región de confianza cumple la ecuación

$$(H + \mu I)\mathbf{s} = -\mathbf{g}.$$

Esta ecuación da lugar a condiciones necesarias y suficientes para la solución del TRS.

**Teorema 3.3** ([25]). *El vector  $\mathbf{s}$  es solución del TRS si y sólo si  $H$  es una matriz simétrica positiva semidefinida y existe un parámetro  $\mu \geq 0$  tal que  $\mathbf{s}$  cumple la ecuación*

$$(H + \mu I)\mathbf{s} = -\mathbf{g}$$

*y la relación  $\mu(\|\mathbf{s}\|_2 - \Delta) = 0$ .*

#### Observaciones 3.13:

☞ Si  $H$  es positiva definida, el mínimo del TRS (3.4.2) es único.

☞ Si  $\mu > 0$ , entonces la solución óptima del TRS está en la frontera  $\|\mathbf{s}\|_2 = \Delta$ .

Cuando la matriz  $H + \mu I$  es positiva definida, la ecuación  $(H + \mu I)\mathbf{s} = -\mathbf{g}$  tiene solución única  $\mathbf{s}_\mu$ . Si ésta se encuentra en la frontera de la región de confianza, el valor buscado de  $\mu$  es cero de

$$\phi(\mu) = 1/\Delta - 1/\|\mathbf{s}_\mu\|_2.$$

### 3.4.3. Métodos para el TRS

Podemos usar el Método de Newton para hallar un cero de  $\phi$ . El Algoritmo 3.1 actualiza  $\mu$  con este método.

---

#### Algoritmo 3.1

---

Dados  $\mu \geq 0$ ,  $H + \mu I \in \mathbb{R}^{n \times n}$  positiva definida,  $\mathbf{g} \in \mathbb{R}^n$ ,  $\Delta \geq 0$ .

1. Calcula factorización de Cholesky  $H + \mu I = R^T R$ ;
  2. Resuelve  $R^T R \mathbf{p} = -\mathbf{g}$ ;
  3. Resuelve  $R^T \mathbf{q} = \mathbf{p}$ ;
  4.  $\mu \leftarrow \mu + (\|\mathbf{p}\|_2 / \|\mathbf{q}\|_2)^2 (\|\mathbf{p}\|_2 - \Delta) / \Delta$ .
- 

Entre los métodos computacionales clásicos para resolver el TRS están el gancho que busca  $\mu$  tal que  $\|\mathbf{s}_\mu\| \approx \Delta$  y actualiza la estrategia global con paso  $\mathbf{s}_\mu$  en la iteración correspondiente; la pata de perro de Powell, que realiza una aproximación lineal por tramos de la curva descrita por  $\mathbf{s}$  en función de  $\mu$  y actualiza la estrategia global con el vector sobre esa curva que tiene discrepancia de tamaño  $\Delta$  con el de la anterior iteración; la doble pata de perro de Dennis y Mei [25]; y el algoritmo de Moré [84].

Sea  $\mu_1$  el valor propio más pequeño de  $H$ . Moré propone que además de actualizar  $\mu$  con el Algoritmo 3.1, delimitemos su valor a un intervalo que reduzca su tamaño en cada iteración. Dados  $\gamma_1, \gamma_2 \in (0, 1)$ , una aproximación inicial  $\mathbf{s}_0$  de la solución del TRS, y un valor inicial  $\mu_0 > 0$  del multiplicador de Lagrange tal que  $H + \mu_0 I$  sea positiva definida, buscamos una solución aproximada  $\mathbf{s}$  del TRS con el Algoritmo 3.1 que cumpla

$$\psi(\mathbf{s}) - \psi^* \leq \gamma_1(2 - \gamma_1) \max\{|\psi^*|, \gamma_2\} \quad \text{y} \quad \|\mathbf{s}\|_2 \leq (1 + \gamma_1)\Delta,$$

donde  $\psi^*$  es el valor mínimo de  $\psi$  en la región de confianza.

Para calcular  $\mathbf{s}$  con el Algoritmo 3.1, buscamos un vector unitario  $\mathbf{z}$  tal que  $\|R\mathbf{z}\|_2$  sea pequeña, y calculamos  $\tau \in \mathbb{R}$  tal que  $\mathbf{p} + \tau\mathbf{z}$  éste en la frontera de la región de confianza. Detenemos las iteraciones cuando

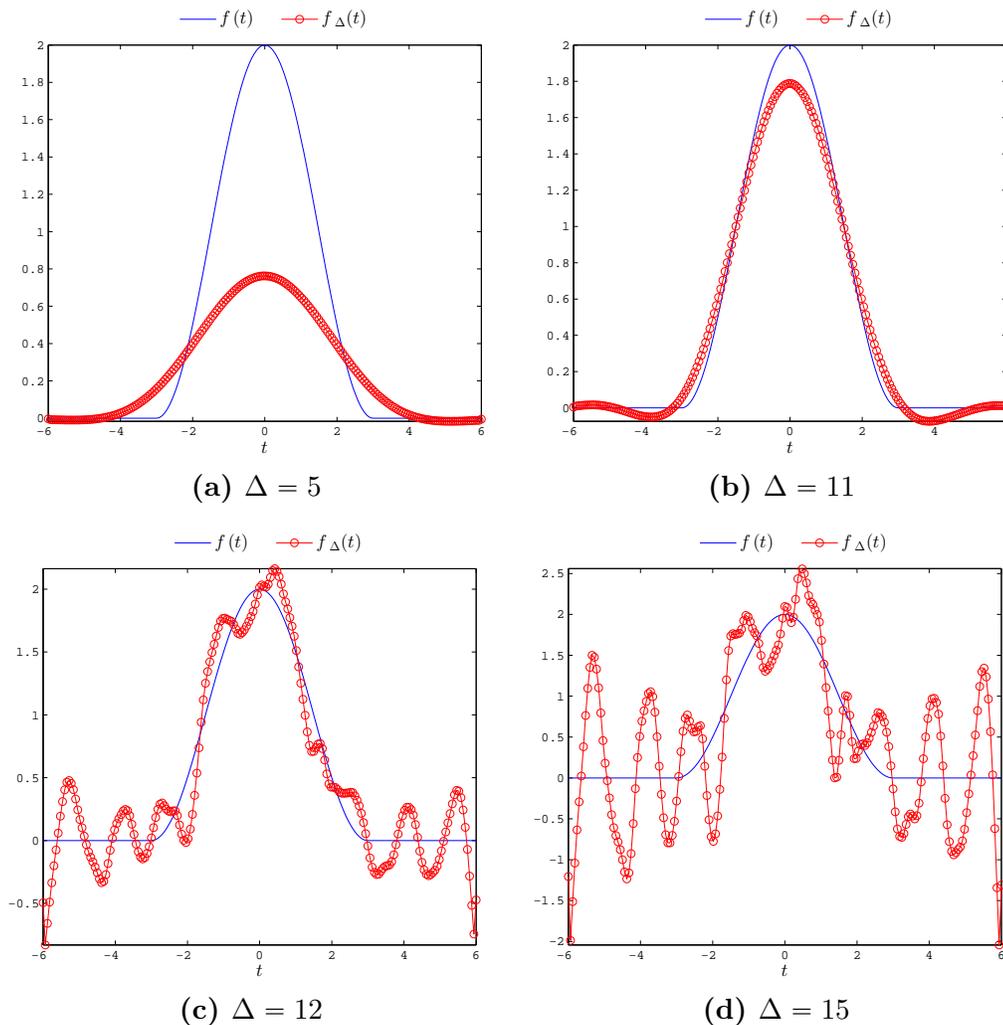
$$|\Delta - \|\mathbf{p}\|_2| \leq \gamma_1 \Delta \quad \text{o} \quad \|R(\tau\mathbf{z})\|_2^2 \leq \gamma_1(2 - \gamma_1) \max\{\gamma_2, \|\mathbf{R}\mathbf{p}\|_2^2 + \mu\Delta^2\}$$

En el primer caso, tomamos  $\mathbf{s} = \mathbf{p}$ , y en el otro,  $\mathbf{s} = \mathbf{p} + \tau\mathbf{z}$ .

**Ejemplo 3.8.** En el Ejemplo 2.1 retomamos la ecuación de Phillips (2.4). Su discretización por método de colocación con regla compuesta del trapecio en 201 puntos nos da el sistema de ecuaciones lineales

$$\frac{6}{200} \begin{bmatrix} k(s_1 - s_1) & 2k(s_1 - s_2) & \cdots & 2k(s_1 - s_{201}) & k(s_1 - s_{201}) \\ \vdots & \vdots & & \vdots & \vdots \\ k(s_{201} - s_1) & 2k(s_{201} - s_2) & \cdots & 2k(s_{201} - s_{201}) & k(s_{201} - s_{201}) \end{bmatrix} \begin{bmatrix} f(s_1) \\ \vdots \\ f(s_{201}) \end{bmatrix} = \begin{bmatrix} g(s_1) \\ \vdots \\ g(s_{201}) \end{bmatrix}.$$

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



**Figura 3.18:** En la ecuación de Phillips del Ejemplo 2.1, aproximamos la función  $f$  con la poligonal  $f_\Delta$  que tiene los valores de la solución del TRS con radio  $\Delta$  en partición uniforme de 201 puntos de  $[-6, 6]$ .

Perturbamos  $\mathbf{b}$  con ruido gaussiano  $\boldsymbol{\epsilon}$  de  $E(\boldsymbol{\epsilon}) = \mathbf{0}$  y  $\text{Cov}(\boldsymbol{\epsilon}) = 0.25I$ . Queremos recuperar una solución aproximada de la ecuación  $A\mathbf{x} = \mathbf{b}$  a partir del modelo de Gauss-Markov  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ . Para ello, resolvemos el TRS para cuatro valores distintos del radio  $\Delta$  de la región de confianza. Usamos el método de Moré con  $\gamma_1 = 0.1$ ,  $\gamma_2 = 0$ ,  $\mu_0 = 0$  y  $\mathbf{s} = \mathbf{0}$ . Formamos la poligonal  $f_\Delta$  que toma los valores del mínimo calculado  $\mathbf{x}_\Delta$  en la partición uniforme de 201 puntos del intervalo  $[-6, 6]$ .

Para  $\Delta = 5$ ,  $f_\Delta$  no alcanza el valor máximo de  $f$ , como se muestra en la Figura 3.18(a). Con  $\Delta = 11$ , vemos en la Figura 3.18(b) que  $f_\Delta$  se eleva cerca del valor máximo de  $f$ ; aun así hay oscilaciones en los extremos. En la Figura 3.18(c), aumentamos el radio a  $\Delta = 12$ . Se presentan perturbaciones en la poligonal  $f_\Delta$ . Para un radio más grande, la amplitud de las oscilaciones aumenta. Véase la Figura 3.18(d) con radio  $\Delta = 15$ . Así, cuando  $\Delta$  es pequeño,  $f_\Delta$  es suave y de amplitud pequeña; mientras que si  $\Delta$  es grande,  $f_\Delta$  oscila.

### 3.4.4. TRS y Regularización

Considere el problema discreto mal planteado dado por el modelo de Gauss-Markov  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ , donde  $A \in \mathbb{R}^{m \times n}$  con  $m \geq n$ ,  $\mathbf{b} \in \mathbb{R}^m$  y  $\boldsymbol{\epsilon}$  es ruido gaussiano. Tratamos este problema mediante el método de cuadrados mínimos con una restricción cuadrática en el tamaño de la solución:

$$\min_{\substack{\mathbf{x} \in \mathbb{R}^n \\ \|\mathbf{x}\|_2 \leq \Delta}} \frac{1}{2} \|A\mathbf{x} - (\mathbf{b} + \boldsymbol{\epsilon})\|_2^2. \quad (3.15)$$

Podemos dar la solución de este problema.

**Teorema 3.4** ([38]). *Sea  $\mathbf{x}_{LS}$  la solución de cuadrados mínimos de norma mínima de la ecuación  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ . Supongamos que  $\Delta < \|\mathbf{x}_{LS}\|_2$ . Entonces el Problema de Minimización (3.15) tiene única solución*

$$\mathbf{x}_{\mu_\Delta} = (A^T A + \mu_\Delta I)^{-1} A^T (\mathbf{b} + \boldsymbol{\epsilon})$$

donde  $\mu_\Delta > 0$  y cumple que  $\|\mathbf{x}_{\mu_\Delta}\|_2 = \Delta$ .

#### Observaciones 3.14:

☞ El Teorema 3.4 nos dice que el Problema de Cuadrados Mínimos (3.15) con restricciones cuadráticas equivale a la regularización de Tikhonov.

Nuestra intención es emplear los métodos para resolver el TRS en la regularización de problemas discretos mal planteados. Sean  $H = A^T A$  y  $\mathbf{g} = -A^T (\mathbf{b} + \boldsymbol{\epsilon})$ . El problema de minimización (3.15) se trata como el TRS

$$\min_{\|\mathbf{x}\|_2 \leq \Delta} \left( \frac{1}{2} \mathbf{x}^T H \mathbf{x} + \mathbf{g}^T \mathbf{x} \right). \quad (3.16)$$

**Observaciones 3.15:**

☞ La restricción cuadrática  $\|\mathbf{x}\|_2^2 \leq \Delta^2$  nos da la región de confianza, y el parámetro  $\Delta$  es el radio de esa región.

☞ La Ecuación  $(H + \mu I)\mathbf{s} = -\mathbf{g}$ , es equivalente a las ecuaciones normales regularizadas

$$(A^T A + \mu I)\mathbf{x} = A^T(\mathbf{b} + \boldsymbol{\epsilon}).$$

La Observaciones 3.15 junto con los teoremas 3.3 y 3.4 nos dicen que la regularización de Tikhonov equivale a resolver el TRS (3.16).

**Ejemplo 3.9 (Problema inverso de la ecuación de calor[16], [31]).** En el problema de la ecuación de calor queremos conocer la distribución de la temperatura en el interior de un cuerpo a partir su temperatura superficial. El problema inverso de la ecuación del calor consiste en estimar la temperatura superficial de un cuerpo que conduce calor a partir de las temperaturas medidas en posiciones de su interior. Algunas de las aplicaciones son la estimación del calor en barriles de pistolas y en la punta de cohetes espaciales que entran en la atmosfera.

Consideramos una barra cilíndrica de longitud infinita y área de sección transversal uno con superficie lateral aislada de modo que el flujo de calor que pasa por la barra solo se mueve a lo largo y se distribuye uniformemente en cada sección transversal. La temperatura  $u$  depende tanto de la posición  $0 \leq x < \infty$  como del tiempo  $0 \leq t < \infty$ . No hay fuentes externas de calor y la temperatura inicial es cero.

Con nuestras hipótesis, la evolución de la temperatura está descrita por la ecuación del calor

$$u_t = \kappa^2 u_{xx},$$

donde  $\kappa$  es el coeficiente de difusión que en nuestro caso tomamos como una constante positiva. Así que tenemos el problema de valores iniciales y de frontera en un cuadrante del plano:

$$\begin{cases} u_t = \kappa^2 u_{xx}, & 0 < x, t < \infty, \\ u(x, 0) = 0, & 0 \leq x < \infty, \\ u(0, t) = f(t), & 0 \leq t < \infty. \end{cases}$$

Como el problema es lineal, una manera de encontrar la temperatura  $u$  con cualquier condición de frontera que varia en el tiempo es resolver el mismo problema con la condición de frontera en  $x = 0$  dada por el impulso unitario  $\delta$  en vez de  $f$ . Puesto que la solución  $k$

del problema con  $u(0, t) = \delta(t)$  está dada por

$$k(x, t) = \frac{x}{2\sqrt{\pi\kappa t^{3/2}}} \exp\left(-\frac{x}{4\kappa^2 t}\right), \quad 0 < x, t < \infty.$$

mediante la transformada de Laplace se puede verificar [41] que la solución acotada del problema original es

$$u(x, t) = \int_0^t k(x, t - \tau) f(\tau) d\tau, \quad 0 \leq x, t < \infty.$$

En el problema inverso, deseamos calcular la temperatura  $u(0, t) = f(t)$  cuando disponemos de una medida de temperatura  $u(1, t) = g(t)$ . La idea para resolverlo, es reformularlo como la ecuación integral de Volterra de primera clase

$$\int_0^t k(t - \tau) f(\tau) d\tau = g(t), \quad 0 \leq t < \infty. \quad (3.17)$$

La solución que  $u$  que obtuvimos del problema directo nos da esta ecuación cuando nos fijamos en la posición  $x = 1$ .

### Observaciones 3.16:

 La función  $g$  es la convolución de  $k$  con  $f$ . Así que el problema inverso de la ecuación del calor se reduce a una deconvolución.

 Existen otras maneras de abordar el tema. Nosotros seguimos el enfoque de Carasso [16] y utilizamos la rutina `heat` del paquete `REGUTOOLS` de Hansen [50]. El lector interesado puede consultar Beck, Blackwell, Clair [7].

En el intervalo  $[0, 1]$ , tenemos  $n = 100$  observaciones de  $g$  obtenidas con la rutina `heat`, a las que agregamos ruido aditivo gaussiano  $\epsilon$  con  $E(\epsilon) = \mathbf{0}_{n \times 1}$  y  $\text{Cov}(\epsilon) = 0.001I_{100 \times 100}$ . Resolvemos numéricamente la Ecuación de Volterra para obtener los valores de  $f$ .

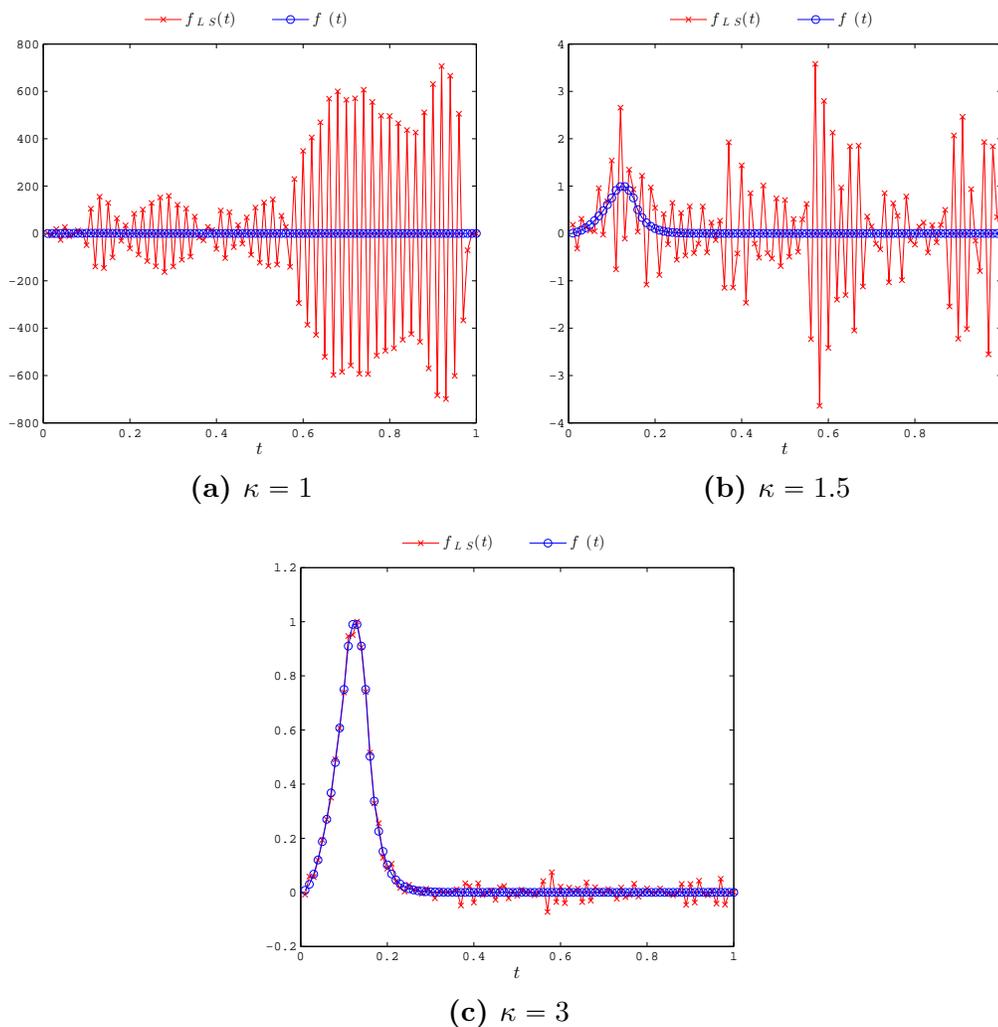
Discretizamos la ecuación (3.17) por colocación en 100 puntos  $t_i = i/100$ ,  $i = 1, \dots, 100$ . Aproximamos la integral por cuadratura compuesta del punto medio. De ese modo, formamos el sistema de ecuaciones lineales

$$\frac{1}{100} \begin{bmatrix} k_1 & & & 0 \\ k_2 & k_1 & & \\ \vdots & & \ddots & 0 \\ k_{100} & k_{99} & \cdots & k_1 \end{bmatrix} \begin{bmatrix} f(t_1) \\ \vdots \\ f(t_{100}) \end{bmatrix} = \begin{bmatrix} g(t_1) \\ \vdots \\ g(t_{100}) \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{100} \end{bmatrix},$$

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

donde  $k_i = k(t_{i-1/2})$  para  $i = 1, \dots, 100$ . La matriz  $A$  tiene diagonales constantes (de Toeplitz) y es triangular inferior.

Calculamos la solución  $\mathbf{x}_{LS}$  de cuadrados mínimos de norma mínima de la ecuación  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ . Formamos la poligonal  $f_{LS}$  que une los puntos  $(t_i, x_{LS}(i))$ ,  $i = 1, \dots, 100$ . En la Figura 3.19 comparamos  $f_{LS}$  con la aproximación  $f$  de la solución que nos da Hansen [50].  $f_{LS}$  depende del coeficiente de difusión  $\kappa$ . Para  $\kappa = 3$ , notamos pequeñas perturbaciones en  $f_{LS}$ . Cuando  $\kappa = 1.5$ , las oscilaciones de  $f_{LS}$  se amplifican. Si  $\kappa = 1$ , las perturbaciones ocasionan cambios drásticos en  $f_{LS}$ .



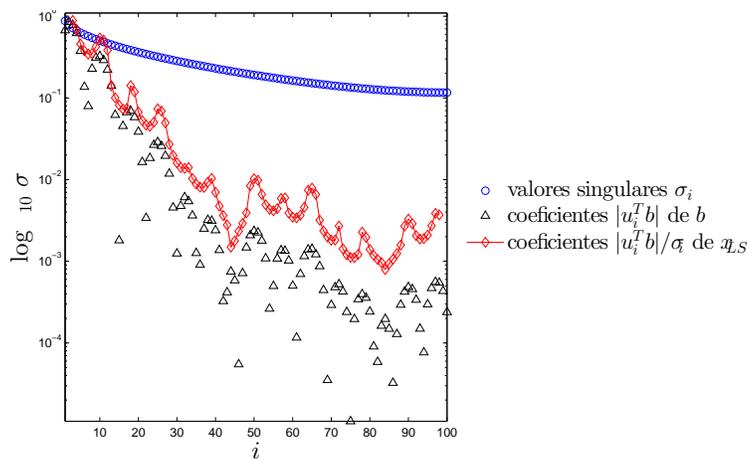
**Figura 3.19:** Para distintos valores de  $\kappa$  mostramos la solución  $f_{LS}$  de cuadrados mínimos de norma mínima de la ecuación (3.17) discretizada en 100 puntos de colocación, y la comparamos con la solución  $f$ .

Examinamos el condicionamiento del problema inverso discreto. Con los valores singula-

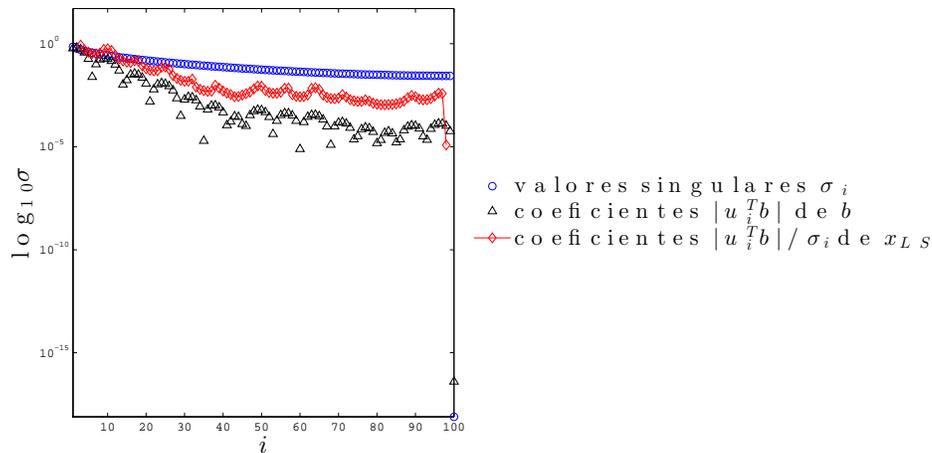
res  $\sigma_1, \dots, \sigma_n$  de  $A$  en orden decreciente, vemos en la Tabla 3.4 como cambia gradualmente el rango de la matriz y su número de condición decrece significativamente. El problema pasa de estar bien condicionado a ser de rango deficiente conforme  $\kappa$  disminuye.

$k$	$\sigma_n$	rankA	cond A
5	0.11618	100	7.54188
3	$7.8053 \times 10^{-19}$	99	$9.17928 \times 10^{17}$
1.5	$1.38177 \times 10^{-25}$	98	$3.70936 \times 10^{24}$

**Tabla 3.4:** Número de condición y rango de matriz  $A$  para distintos coeficientes de difusión  $k$ .



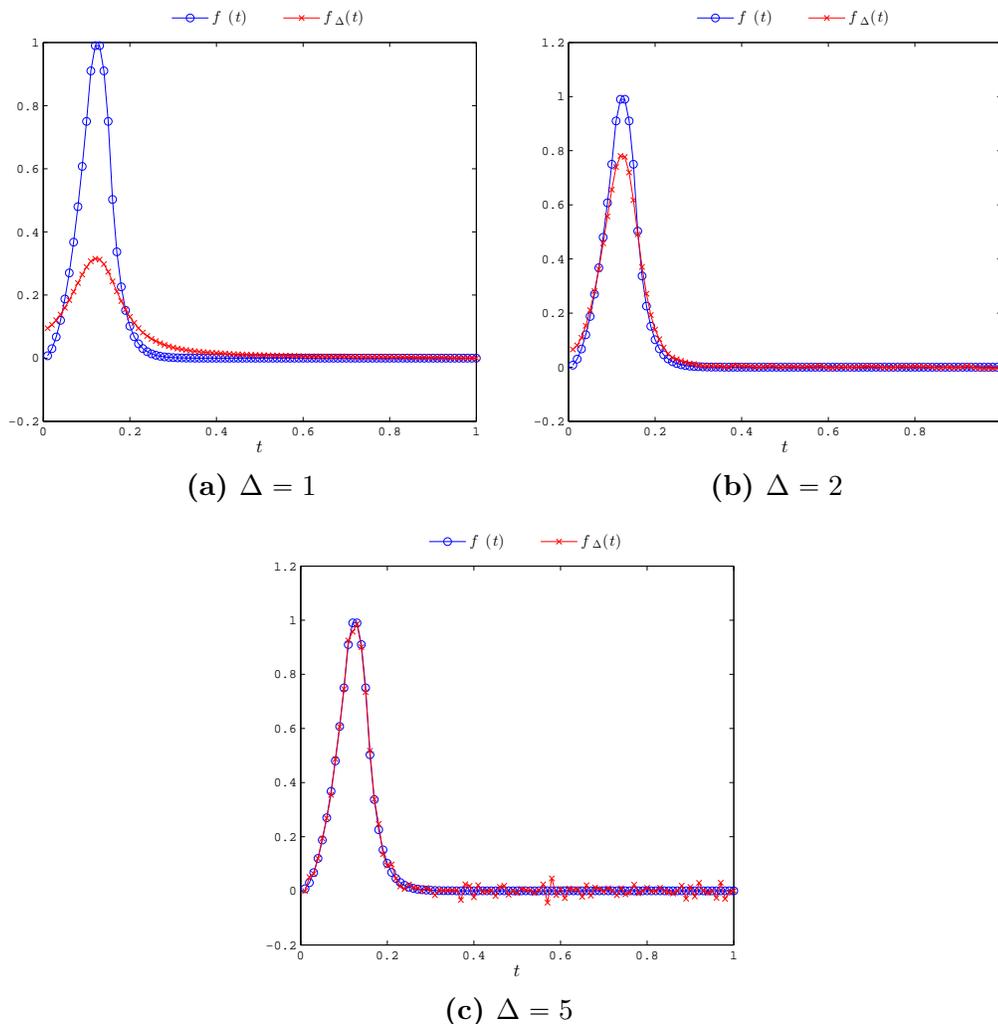
**Figura 3.20:** Gráfica de Picard del problema inverso de calor discreto  $Ax = b$  para  $\kappa = 5$ .



**Figura 3.21:** Gráfica de Picard del problema inverso de calor discreto  $Ax = b$  para  $\kappa = 3$ .

Mostramos las gráficas de Picard del problema discreto  $Ax = b$  para  $\kappa = 5$  en la Figura 3.20 y  $\kappa = 3$  en la Figura 3.21. En ambas la sucesión con la media geométrica móvil

de los valores absolutos de los coeficientes  $\mathbf{u}_i^T/\sigma_i$  de  $\mathbf{x}_{LS}$  no es monótona decreciente. Con los primeros 10 valores singulares se cumple la DPC. Para  $\kappa = 3$ , notamos un salto de  $\sigma_{99} \approx 10^{-2}$  a  $\sigma_{100} \approx 10^{-19}$ , por lo que el problema es de rango deficiente.



**Figura 3.22:** Solución  $f$  del problema inverso del calor y poligonal  $f_{\Delta}$  con valores de las componentes de las soluciones del TRS con  $\kappa = 3$  y distintos valores del radio  $\Delta$ .

Para obtener una mejor aproximación de los valores de la solución con las observaciones  $\mathbf{b} + \boldsymbol{\epsilon}$ , regularizamos la ecuación  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$  siguiendo el enfoque del TRS con diferentes radios de confianza  $\Delta$ . Resolvemos el TRS por el método de Moré con parámetros  $\gamma_1 = 0.1$ ,  $\gamma_2 = 0$ , multiplicador inicial  $\mu_0 = 0$  y aproximación inicial  $\mathbf{s}_0 = \mathbf{0}$ . Denotamos por  $f_{\Delta}$  a la poligonal con los valores de la solución aproximada del TRS con radio  $\Delta$ .

La aproximación  $f_{\Delta}$  es suave con  $\kappa = 3$  pero de baja amplitud con  $\Delta = 1$  como se muestra en Figura 3.22(a). Si  $\Delta = 2$ , la poligonal  $f_{\Delta}$  de la Figura 3.22(b) aumenta su valor máximo sin generar oscilaciones. Cuando tomamos  $\Delta = 5$  aparecen pequeñas



la literatura como el algoritmo libre de matrices de Santos y Sorensen [103] y el uso del gradiente conjugado preconditionado que propone Steihaug [111].

### 3.4.5. Gradiente Conjugado en el TRS

Para resolver problemas de cuadrados mínimos asociados a ecuaciones lineales  $A\mathbf{x} = \mathbf{b}$ , se puede ocupar el método de gradiente conjugado (CG) para resolver las ecuaciones normales cuando la matriz  $A^T A$  es positiva definida. Lo que hace el método es generar un sucesión de vectores que en cada iteración resuelven el problema de cuadrados mínimos restringido al subespacio de Krylov  $\{A^T \mathbf{b}, (A^T A)A^T \mathbf{b}, \dots, (A^T A)^{k-1} A^T \mathbf{b}\}$ . En algunas aplicaciones, este subespacio aproxima al generado por los vectores singulares de derecha asociados con los  $k$  valores singulares más grandes de  $A$ . En ese caso, el CG tiene un efecto de regularización en el problema de cuadrados mínimos [99]. Sin embargo, cuando la iteración aproxima al subespacio de Krylov que corresponde con a los valores singulares más pequeños, el ruido del vector de observaciones se propaga en la iteración. Esto provoca que la sucesión se aleje de la solución de cuadrados mínimos [82], [45].

Para acelerar la convergencia del CG en la solución de las ecuaciones normales, podemos usar un preconditionador. El objetivo del preconditionamiento es reducir el número de condición de una matriz, acumulando sus valores propios en una región o hacerlos cercanos a uno. Recordamos que tratamos con problemas discretos mal planteados, por lo que debemos tener en cuenta el mal condicionamiento. Cuando modificamos el espectro de la matriz  $A^T A$  con el preconditionador, no distinguimos entre los valores propios más grandes y más pequeños. Como los valores propios de  $A^T A$  son los cuadrados de los valores singulares positivos de  $A$ , el preconditionador puede reunir los valores singulares más grandes con los más pequeños, lo que provoca que la aproximación calculada por el CG preconditionado no converja a la solución.

En un problema mal condicionado, el preconditionador no debe cambiar todo el espectro de la matriz, sino que debe modificar los valores propios más grandes y dejar intactos a los valores propios más pequeños como señala Rojas [99], Hanke y Hansen, [46]. El preconditionamiento de problemas mal planteados de gran escala es tanto un área difícil y en desarrollo que no incluimos en esta obra. El lector interesado puede consultar [91],[20] para ver algunos de los métodos de preconditionamiento propuestos en este escenario.

## 3.5. Elección del Parámetro de Regularización

Una de las tareas de la regularización es elegir el valor del parámetro. En esta sección presentamos métodos para determinar el valor adecuado de éste:  $k \in \mathbb{N}$  para SVD truncada,  $\lambda > 0$  para regularización de Tikhonov, etc. En los ejemplos anteriores vimos que cuando el parámetro es pequeño, los errores de redondeo, truncamiento y en observaciones se propagan y amplifican en la solución regularizadora, por lo que debemos evitar que esté

cerca de cero; mientras que si el parámetro es grande, la solución regularizadora es más suave, pero no hay que suavizar más de lo necesario.

Considere el problema discreto mal planteado dado por el modelo lineal  $\mathbf{b} = A\mathbf{x} + \boldsymbol{\epsilon}$ . Denotamos por  $\mathbf{x}_{\text{reg}}$  a la solución regularizadora. Para elegir el parámetro de regularización, minimizamos el tamaño del error  $\mathbf{x}_{\text{reg}} - \mathbf{x}$ . Con la SVD  $A = U\Sigma V^T$  y los factores filtro  $\varphi_i$  vimos que

$$\mathbf{x}_{\text{reg}} = V\Phi\Sigma^\dagger U^T \mathbf{b}$$

Así que el error en la solución regularizada es

$$\begin{aligned} \mathbf{x}_{\text{reg}} - \mathbf{x} &= V\Phi\Sigma^\dagger U^T A\mathbf{x} + V\Phi\Sigma^\dagger U^T \boldsymbol{\epsilon} - \mathbf{x} \\ &= (V\Phi\Sigma^\dagger U^T A - I)\mathbf{x} - V\Phi\Sigma^\dagger U^T \boldsymbol{\epsilon} \\ &= (V\Phi\Sigma^\dagger U^T U\Sigma V^T - VV^T)\mathbf{x} + V\Phi\Sigma^\dagger U^T \boldsymbol{\epsilon} \\ &= V(\Phi\Sigma^\dagger \Sigma - I)V^T \mathbf{x} + V\Phi\Sigma^\dagger U^T \boldsymbol{\epsilon} \end{aligned}$$

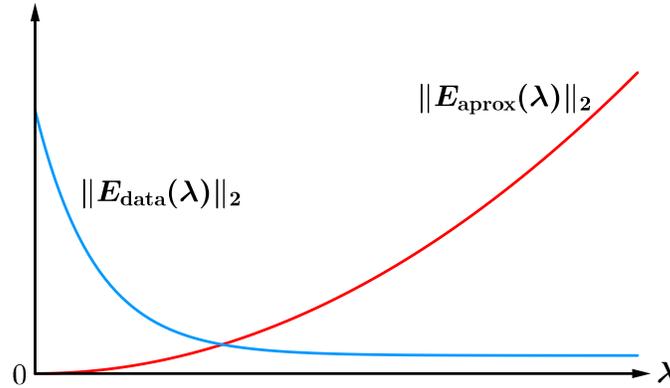
Entonces  $\mathbf{x}_{\text{reg}} - \mathbf{x}$  es el error de los datos

$$E_{\text{data}} = V\Phi\Sigma^\dagger U^T \boldsymbol{\epsilon}$$

más el error de aproximación

$$E_{\text{aprox}} = V(\Phi\Sigma^\dagger \Sigma - I)V^T \mathbf{x}$$

Para ruido aleatorio  $\boldsymbol{\epsilon}$  que cumple con las condiciones de Gauss-Markov, identificamos a  $E_{\text{aprox}}$  con el sesgo de  $\mathbf{x}_{\text{reg}}$  y a  $E_{\text{data}}$  con la propagación del error en el modelo.



**Figura 3.24:** Tamaños de  $E_{\text{data}}$  y  $E_{\text{aprox}}$  en función del parámetro de regularización  $\lambda$ .

### Observaciones 3.18:

☞ En la regularización de Tikhonov, tenemos que  $\|E_{\text{aprox}}\|_2$  es una función monótona creciente en el parámetro  $\lambda$  tal que  $\|E_{\text{aprox}}\|_2 \rightarrow 0$  cuando  $\lambda \rightarrow 0$ ; mientras que  $\|E_{\text{data}}\|_2$  es una función monótona decreciente en el parámetro  $\lambda$  tal que  $\|E_{\text{data}}\|_2 \rightarrow 0$  cuando  $\lambda \rightarrow \infty$ . Véase la Figura 3.24.

El objetivo de elegir un parámetro de regularización es balancear el error de los datos con el error de aproximación. Vamos a examinar tres criterios para elegir el parámetro de regularización.

### 3.5.1. Principio de Discrepancia

Supongamos que conocemos el nivel del ruido  $\|\boldsymbol{\epsilon}\|_2$  en las observaciones. Una manera para determinar el valor adecuado del parámetro de regularización es pedir que el residuo  $A\mathbf{x}_{\text{reg}} - \mathbf{b}$  tenga un tamaño comparable a la discrepancia en los datos. Este criterio, atribuida a Morozov [85], se conoce como **Principio de Discrepancia**. Para poder aplicarlo, necesitamos de una cota para el tamaño de la discrepancia entre los datos.

#### Aplicado a SVD Truncada

En el caso de la SVD truncada, podemos elegir un nivel de truncamiento adecuado si conocemos el nivel de error en las observaciones.

**Teorema 3.5.** Sea  $A \in \mathbb{R}^{m \times n}$  de rango  $r$  con SVD dada por  $A = U\Sigma V^T$ . Si  $\mathbf{y} \in \text{Col}(A)$  y  $\mathbf{y}^\delta \in \mathbb{R}^m$  cumplen

$$\|\mathbf{y} - \mathbf{y}^\delta\|_2 \leq \delta \leq \|\mathbf{y}^\delta\|_2,$$

entonces existe  $k \in \{1, \dots, r\}$  tal que

$$\mathbf{x}_k = \sum_{j=1}^k \frac{1}{\sigma_j} (\mathbf{u}_j^T \mathbf{y}^\delta) \mathbf{v}_j$$

satisface  $\|A\mathbf{x}_k - \mathbf{y}^\delta\|_2 \leq \delta$ , más aún,  $\|\mathbf{x}_k - A^\dagger \mathbf{y}\|_2 \rightarrow 0$  cuando  $\delta \rightarrow 0$ .

*Demostración.* Sea  $\psi : \{0, \dots, r\} \rightarrow \mathbb{R}$  la función dada por

$$\psi(p) = \|A\mathbf{x}_p - \mathbf{y}^\delta\|_2^2 - \delta^2.$$

Notamos que

$$\begin{aligned} A\mathbf{x}_p - \mathbf{y}^\delta &= U\Sigma V^T \left( \frac{\mathbf{u}_1^T \mathbf{y}^\delta}{\sigma_1} \mathbf{v}_1 + \dots + \frac{\mathbf{u}_p^T \mathbf{y}^\delta}{\sigma_p} \mathbf{v}_p \right) - UU^T \mathbf{y}^\delta \\ &= U \begin{pmatrix} \frac{\mathbf{u}_1^T \mathbf{y}^\delta}{\sigma_1} \\ \vdots \\ \frac{\mathbf{u}_p^T \mathbf{y}^\delta}{\sigma_p} \\ \mathbf{0}_{(m-p) \times 1} \end{pmatrix} - UU^T \mathbf{y}^\delta. \end{aligned}$$

Para  $p = r = m$  tenemos que  $A\mathbf{x}_p - \mathbf{y}^\delta = \mathbf{0}$ , y por consiguiente  $\psi(p) = -\delta^2$ .

Para  $p < m$  tenemos que

$$A\mathbf{x}_p - \mathbf{y}^\delta = -U \begin{pmatrix} \mathbf{0}_{k \times 1} \\ \frac{\mathbf{u}_{p+1}^T \mathbf{y}^\delta}{\sigma_{p+1}} \\ \vdots \\ \frac{\mathbf{u}_m^T \mathbf{y}^\delta}{\sigma_m} \end{pmatrix}.$$

Luego, como  $U$  es una matriz ortogonal,

$$\psi(p) = \sum_{j=p+1}^m (\mathbf{u}_j^T \mathbf{y}^\delta)^2 - \delta^2.$$

Entonces

- $\psi$  es monótona decreciente
- $\psi(0) = \sum_{j=1}^m (\mathbf{u}_j^T \mathbf{y}^\delta)^2 - \delta^2 = \|U^T \mathbf{y}^\delta\|_2^2 - \delta^2 = \|\mathbf{y}^\delta\|_2^2 - \delta^2 \geq 0$ .
- Dado que  $\mathbf{y} \in \text{Col}(A)$ ,  $\mathbf{u}^T \mathbf{y} = 0$  para todo  $\mathbf{u}$  ortogonal a  $\text{Col}(A)$ . A su vez, como  $\mathbf{u}_{r+1}, \dots, \mathbf{u}_m$  generan todos los vectores ortogonales a  $\text{Col}(A)$ , se sigue que  $\mathbf{u}_j^T \mathbf{y} = 0$  para  $j = r + 1, \dots, m$ . En consecuencia,

$$\begin{aligned} \psi(r) &= \sum_{j=r+1}^m (\mathbf{u}_j^T \mathbf{y}^\delta)^2 - \delta^2 \\ &= \sum_{j=r+1}^m (\mathbf{u}_j^T (\mathbf{y}^\delta - \mathbf{y}))^2 - \delta^2 \\ &\leq \sum_{j=1}^m (\mathbf{u}_j^T (\mathbf{y}^\delta - \mathbf{y}))^2 - \delta^2 \\ &= \|U^T (\mathbf{y}^\delta - \mathbf{y})\|_2^2 - \delta^2 \\ &= \|\mathbf{y}^\delta - \mathbf{y}\|_2^2 - \delta^2 \\ &\leq 0. \end{aligned}$$

Por consiguiente, existe  $k \in \{1, \dots, r\}$  tal que  $\psi(k) \leq 0$ , es decir,

$$\|A\mathbf{x}_k - \mathbf{y}^\delta\|_2 \leq \delta.$$

De entre estos enteros seleccionamos el menor que cumple con esto.

Ahora, veamos que  $\mathbf{x}_k \rightarrow A^\dagger \mathbf{y}$  cuando  $\delta \rightarrow 0$ . En efecto, dado que

$$A^\dagger A \mathbf{v}_j = V \Sigma^\dagger U^T U \Sigma V^T \mathbf{v}_j = V \begin{pmatrix} I_{r \times r} & \\ & \mathbf{0}_{(n-r) \times (n-r)} \end{pmatrix} V^T \mathbf{v}_j = \mathbf{v}_j, \quad j = 1, \dots, r,$$

tenemos que  $A^\dagger A \mathbf{x}_k = \mathbf{x}_k$ . Entonces

$$\begin{aligned} \|\mathbf{x}_k - A^\dagger \mathbf{y}\|_2 &= \|A^\dagger (A \mathbf{x}_k - \mathbf{y})\|_2 \\ &\leq \|A^\dagger\|_2 \|A \mathbf{x}_k - \mathbf{y}\|_2 \\ &\leq \|A^\dagger\|_2 (\|A \mathbf{x}_k - \mathbf{y}^\delta\|_2 + \|\mathbf{y} - \mathbf{y}^\delta\|_2). \end{aligned}$$

Luego,

$$\|A \mathbf{x}_k - \mathbf{y}^\delta\|_2 \leq \delta \quad \text{y} \quad \|\mathbf{y} - \mathbf{y}^\delta\|_2 \leq \delta$$

implican que  $\|\mathbf{x}_k - A^\dagger \mathbf{y}\|_2 \rightarrow 0$  cuando  $\delta \rightarrow 0$  ♣

### Observaciones 3.19:

 Para  $\mathbf{y}^\delta = \mathbf{b}$  e  $\mathbf{y} = A \mathbf{x}$ , el nivel de truncamiento seleccionado es el índice  $k$  más grande tal que

$$\|A \mathbf{x}_k - \mathbf{b}\|_2 \leq \|\boldsymbol{\epsilon}\|_2.$$

 Para evitar que el ruido domine a la solución regularizadora, Hansen [54] introduce un factor  $\omega > 1$  en el tamaño de la discrepancia  $\|\boldsymbol{\epsilon}\|_2$  de modo que busquemos el primer índice  $k$  tal que

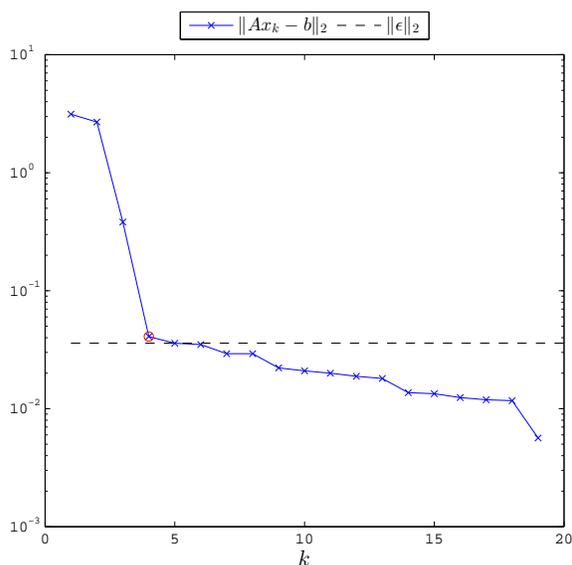
$$\|A \mathbf{x}_k - \mathbf{b}\|_2 \leq \omega \|\boldsymbol{\epsilon}\|_2 < \|A \mathbf{x}_{k+1} - \mathbf{b}\|_2.$$

**Ejemplo 3.10.** Retomemos el problema de la reconstrucción del haz de luz (Ejemplo 2.5). Su discretización por colocación y cuadratura compuesta de punto medio nos da el sistema de ecuaciones  $A \mathbf{x} = \mathbf{b}$ , donde

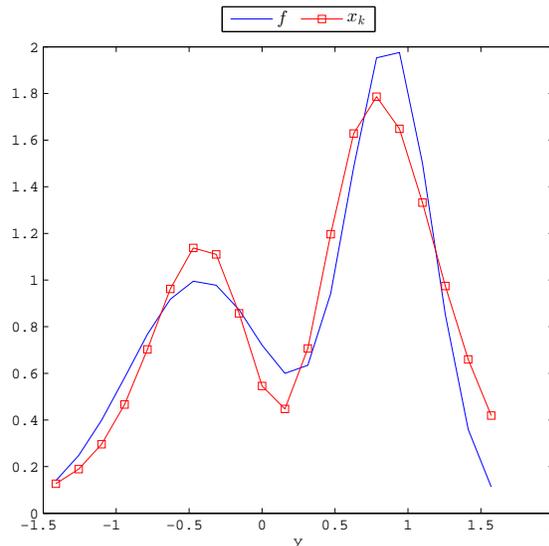
$$\begin{aligned} b_i &= g(\phi_i) \text{ dado por la Tabla 2.3} \\ a_{i,j} &= \frac{\pi}{20} (\cos \phi_i + \cos \phi_j)^2 \left( \frac{\sin(\pi(\sin \phi_i + \sin \phi_j))}{\pi(\sin \phi_i + \sin \phi_j)} \right)^2, \quad i, j = 1, \dots, 20. \\ \phi_i &= (i - 1/2)\pi/20 - \pi/2, \end{aligned}$$

Introducimos errores aditivos  $\epsilon_i$  idénticamente distribuidos bajo una gaussiana de media cero y desviación estándar 0.1 en las observaciones. Mediante la SVD truncada de  $A$ , obtenemos una aproximación de la solución de cuadrados mínimos de norma mínima de la ecuación  $A \mathbf{x} = \mathbf{b}$  a partir de la ecuación perturbada  $A \mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ .

Con ayuda del criterio de discrepancia obtenemos el nivel de truncamiento  $k = 4$ . En la Figura 3.25 mostramos la gráfica de la norma residual  $\|\mathbf{b} - A \mathbf{x}_k\|_2$  en función del parámetro  $k$ . Este nivel de truncamiento está cerca del nivel del error  $\|\boldsymbol{\epsilon}\|_2$ . En la Figura 3.26 comparamos la solución exacta  $f$  con la solución regularizada  $\mathbf{x}_k$  para  $k = 4$ .



**Figura 3.25:** Gráfica de la norma residual  $\|\mathbf{b} - A\mathbf{x}_k\|_2$  del problema discreto regularizado del Ejemplo 2.5 en función de  $k$ . Distinguimos el nivel de truncamiento  $k = 4$  por arriba el nivel de error  $\|\epsilon\|_2$



**Figura 3.26:** Solución aproximada  $f$  del Ejemplo 2.5 sin ruido en observaciones y la poligonal con los valores de las componentes de solución regularizadora  $\mathbf{x}_k$  del problema con ruido para  $k = 4$ .

### Aplicado a Regularización de Tikhonov

Además de la SVD truncada, podemos usar el principio de discrepancia en la regularización de Tikhonov para hallar un valor adecuado del parámetro  $\lambda$  a partir del nivel de error en las observaciones.

**Teorema 3.6** ([71]). Sea  $A \in \mathbb{R}^{m \times n}$  con SVD dada por  $A = U\Sigma V^T$ . Si  $\mathbf{y} \in \text{Col}(A)$  y  $\mathbf{y}^\delta \in \mathbb{R}^m$  cumplen

$$\|\mathbf{y} - \mathbf{y}^\delta\|_2 \leq \delta \leq \|\mathbf{y}^\delta\|_2,$$

entonces existe un único escalar positivo  $\lambda = \lambda(\delta)$  tal que

$$\mathbf{x}_\lambda = \sum_{j=1}^n \frac{\sigma_j}{\lambda^2 + \sigma_j^2} (\mathbf{v}_j^T \mathbf{y}^\delta) \mathbf{u}_j$$

satisface

$$\|A\mathbf{x}_\lambda - \mathbf{y}^\delta\|_2 = \delta,$$

más aún,  $\|\mathbf{x}_\lambda - A^\dagger \mathbf{y}\|_2 \rightarrow 0$  cuando  $\delta \rightarrow 0$ .

**Observaciones 3.20:**

☞ De acuerdo con el Teorema 3.6, escogemos el parámetro de regularización  $\lambda > 0$  al hacer coincidir el tamaño de la discrepancia con el tamaño del error en las observaciones. Para  $\mathbf{y}^\delta = \mathbf{b}$ ,  $\mathbf{y} = A\mathbf{x}$ ,  $\delta = \|\boldsymbol{\epsilon}\|_2$ , esto es,

$$\|A\mathbf{x}_\lambda - \mathbf{b}\|_2 = \|\boldsymbol{\epsilon}\|_2.$$

☞ Para evitar que el ruido domine a la solución regularizadora, Hansen [54] introduce un factor  $\omega > 1$ . Así, elegimos  $\lambda$  tal que

$$\|A\mathbf{x}_\lambda - \mathbf{b}^\delta\|_2 = \omega\|\boldsymbol{\epsilon}\|_2.$$

En términos de la SVD, debemos hallar un cero de

$$\phi(\lambda) = \lambda^4 \sum_{j=1}^r \left( \frac{\mathbf{u}_j^T \mathbf{b}}{\sigma_j^2 + \lambda^2} \right)^2 + \sum_{j=r+1}^m (\mathbf{u}_j^T \mathbf{b})^2 - \omega^2 \boldsymbol{\epsilon}^T \boldsymbol{\epsilon}.$$

**3.5.2. Criterio de la L-Curva**

Otra manera para ver como la solución regularizada  $\mathbf{x}_{\text{reg}}$  depende de nuestro parámetro de regularización es observar la norma de esta solución  $\|\mathbf{x}_{\text{reg}}\|_2$  y la norma del residuo correspondiente  $\|\mathbf{b} - A\mathbf{x}_{\text{reg}}\|_2$ . Nos enfocamos en la regularización de Tikhonov. En este caso el parámetro es  $\lambda > 0$  y  $\mathbf{x}_{\text{reg}} = \mathbf{x}_\lambda$ .

Sean  $\eta : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  y  $\rho : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  las funciones dadas por

$$\eta(\lambda) = \|\mathbf{x}_\lambda\|_2^2 \quad \text{y} \quad \rho(\lambda) = \|\mathbf{b} - A\mathbf{x}_\lambda\|_2^2.$$

Una manera de elegir el parámetro  $\lambda$  es buscar el punto de la curva

$$\Gamma(\lambda) = (\rho(\lambda), \eta(\lambda)) : \mathbb{R}^+ \rightarrow \mathbb{R}^2$$

con mayor curvatura. Esta estrategia para elegir el parámetro de regularización se conoce como **Criterio de la L-curva** [52]. Buscamos una expresión para la curvatura  $c$  de  $\Gamma$ .

Sea  $\phi$  el ángulo tangencial de  $\Gamma$  en  $\lambda$ , y sea  $s$  su longitud de arco. La curvatura de la curva es

$$c(\lambda) = \frac{d\phi}{ds},$$

esto es,

$$c(\lambda) = \frac{\frac{d\phi}{d\lambda}}{\frac{ds}{d\lambda}}.$$

Dado que

$$\frac{d\phi}{d\lambda} = \frac{1}{1 + \tan^2 \phi} \cdot \frac{d}{d\lambda}(\tan \phi) \quad \text{y} \quad \tan \phi = \frac{d\eta}{d\rho},$$

la regla de la cadena

$$\frac{d\eta}{d\rho} = \frac{\eta'}{\rho'} \tag{3.18}$$

implica que

$$\frac{d\phi}{d\lambda} = \frac{\rho'\eta'' - \eta'\rho''}{(\rho')^2 + (\eta')^2}.$$

Luego, como

$$\frac{ds}{d\lambda} = \sqrt{(\eta')^2 + (\rho')^2},$$

se sigue que

$$c(\lambda) = \frac{\rho'(\lambda)\eta''(\lambda) - \eta'(\lambda)\rho''(\lambda)}{((\rho'(\lambda))^2 + (\eta'(\lambda))^2)^{\frac{3}{2}}}.$$

Puesto que  $\eta$  y  $\rho$  pueden tomar tanto valores grandes como pequeños, empleamos una escala logarítmica en las dos variables. Sean

$$\eta_l(\lambda) = \log_{10} \sqrt{\eta(\lambda)} \quad \text{y} \quad \rho_l(\lambda) = \log_{10} \sqrt{\rho(\lambda)}.$$

Sea  $c_l$  la curvatura de  $\Gamma$  en escala logarítmica. Elegimos como parámetro de regularización al máximo de la curvatura  $c_l$ :

$$\max_{\lambda > 0} c_l(\lambda).$$

Esto equivale a resolver

$$\min_{\lambda > 0} [-c_l(\lambda)].$$

Puesto que

$$c_l(\lambda) = \frac{\rho'_l(\lambda)\eta''_l(\lambda) - \eta'_l(\lambda)\rho''_l(\lambda)}{([\rho'_l(\lambda)]^2 + [\eta'_l(\lambda)]^2)^{\frac{3}{2}}},$$

podemos expresar  $c_l$  en términos de  $\eta$  y  $\rho$ . Por propiedades del logaritmo, tenemos que

$$\eta_l(\lambda) = \frac{1}{2 \ln(10)} \ln(\eta(\lambda)),$$

de donde

$$\begin{aligned} \eta'_l(\lambda) &= \frac{1}{2 \ln(10)} \cdot \frac{\eta'(\lambda)}{\eta(\lambda)}, \\ \eta''_l(\lambda) &= \frac{1}{2 \ln(10)} \cdot \frac{\eta(\lambda)\eta''(\lambda) - [\eta'(\lambda)]^2}{[\eta(\lambda)]^2}, \end{aligned}$$

Mutatis mutandis (cambiando lo que se deba cambiar), tenemos que

$$\begin{aligned}\rho_l(\lambda) &= \frac{1}{2 \ln(10)} \ln(\rho(\lambda)), \\ \rho'_l(\lambda) &= \frac{1}{2 \ln(10)} \cdot \frac{\rho'(\lambda)}{\rho(\lambda)}, \\ \rho''_l(\lambda) &= \frac{1}{2 \ln(10)} \cdot \frac{\rho(\lambda)\rho''(\lambda) - [\rho'(\lambda)]^2}{[\rho(\lambda)]^2}.\end{aligned}$$

Una vez que sustituimos estas fórmulas en  $c_l(\lambda)$  y simplificamos, obtenemos

$$c_l(\lambda) = 2 \ln(10) \cdot \frac{\rho(\lambda)\eta(\lambda)}{\eta'(\lambda)} \cdot \frac{\lambda^2 \rho(\lambda)\eta'(\lambda) + 2\lambda \rho(\lambda)\eta(\lambda) + \lambda^4 \eta(\lambda)\eta'(\lambda)}{(\lambda^4[\eta(\lambda)]^2 + [\rho(\lambda)]^2)^{\frac{3}{2}}}.$$

### Observaciones 3.21:



$$\rho'(\lambda) = -\lambda^2 \eta'(\lambda), \quad \lambda > 0. \quad (3.19)$$

En efecto, en términos de los factores filtro

$$\varphi_j(\lambda) = \sigma_j^2 / (\sigma_j^2 + \lambda^2)$$

tenemos que

$$\eta(\lambda) = \sum_{j=1}^r \left[ (\mathbf{u}_j^T \mathbf{b})^2 \left( \frac{\varphi_j}{\sigma_j} \right)^2 \right], \quad (3.20)$$

$$\rho(\lambda) = \sum_{j=1}^r (\mathbf{u}_j^T \mathbf{b})^2 (1 - \varphi_j)^2 + \sum_{j=n+1}^m (\mathbf{u}_j^T \mathbf{b})^2. \quad (3.21)$$

Por lo que

$$\begin{aligned}\eta'(\lambda) &= 2 \sum_{j=1}^r \left[ (\mathbf{u}_j^T \mathbf{b})^2 \cdot \frac{\varphi_j}{\sigma_j^2} \cdot \frac{d\varphi_j}{d\lambda} \right], \\ \rho'(\lambda) &= -2 \sum_{j=1}^r \left[ (\mathbf{u}_j^T \mathbf{b})^2 \cdot (1 - \varphi_j) \cdot \frac{d\varphi_j}{d\lambda} \right].\end{aligned}$$

Como

$$\frac{d\varphi_j}{d\lambda} = -\frac{2\lambda\sigma_j^2}{(\sigma_j^2 + \lambda^2)^2}, \quad j = 1, \dots, n,$$

entonces

$$\eta'(\lambda) = -4\lambda \sum_{j=1}^r (\mathbf{u}_j^T \mathbf{b})^2 \frac{\sigma_j^2}{(\sigma_j^2 + \lambda^2)^3}, \quad (3.22)$$

$$\rho'(\lambda) = 4\lambda^3 \sum_{j=1}^r (\mathbf{u}_j^T \mathbf{b})^2 \frac{\sigma_j^2}{(\sigma_j^2 + \lambda^2)^3}, \quad (3.23)$$

A partir de estas fórmulas es inmediata la identidad (3.19)

 Por la identidad (3.19), tenemos que  $\eta$  decrece conforme  $\rho$  aumenta. Así que el punto donde la curva  $\Gamma$  en escala logarítmica tiene mayor curvatura corresponde a la esquina donde la curva desciende a una parte plana. Intuitivamente, alrededor de ese punto, la curva tiene forma de  $\mathbf{L}$ .

 Mediante las fórmulas de  $\eta$ ,  $\rho$  y  $\eta'$  dadas por (3.20), (3.21) y (3.22), respectivamente, podemos calcular la curvatura  $c_l$  con los valores singulares de la matriz  $A$ .

 La curva  $\Gamma$  es convexa si su curvatura no cambia de signo. Éste es el caso, ya que con la identidad (3.19), tenemos que

$$\rho''(\lambda) = -\lambda^2 \eta''(\lambda) - 2\lambda \eta'(\lambda),$$

y por consiguiente,

$$c(\lambda) = \frac{2\lambda[\eta'(\lambda)]^2}{((\rho'(\lambda))^2 + (\eta'(\lambda))^2)^{\frac{3}{2}}} > 0.$$

 En [43] proponen otra manera de calcular la esquina de la L-curva.

**Ejemplo 3.11.** En el problema de reconstrucción del haz de luz (Ejemplo 2.5), discretizamos por colocación y cuadratura compuesta de punto medio. Obtuvimos la ecuación  $A\mathbf{x} = \mathbf{b}$ , donde

$b_i = g(\phi_i)$  dado por la Tabla 2.3

$$a_{i,j} = \frac{\pi}{20} (\cos \phi_i + \cos \phi_j)^2 \left( \frac{\sin(\pi(\sin \phi_i + \sin \phi_j))}{\pi(\sin \phi_i + \sin \phi_j)} \right)^2, \quad i, j = 1, \dots, 20.$$

$$\phi_i = (i - 1/2)\pi/20 - \pi/2,$$

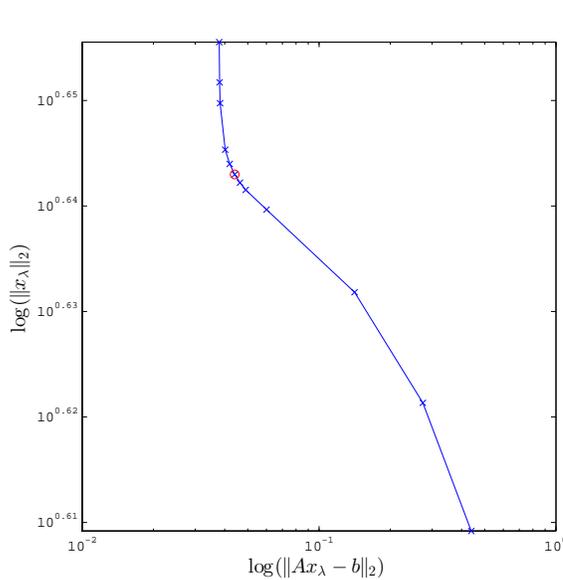
Esta vez agregamos ruido  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, 0.01I_{20 \times 20})$ . Queremos una aproximación de la solución de cuadrados mínimos de norma mínima  $\mathbf{x}^\dagger$  de  $A\mathbf{x} = \mathbf{b}$  a partir de  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$ .

Mediante regularización de Tikhonov tratamos de obtener una aproximación  $\mathbf{x}_\lambda$  de  $\mathbf{x}^\dagger$ . Esta vez elegimos el parámetro de regularización  $\lambda$  con el criterio de la L-curva. En la

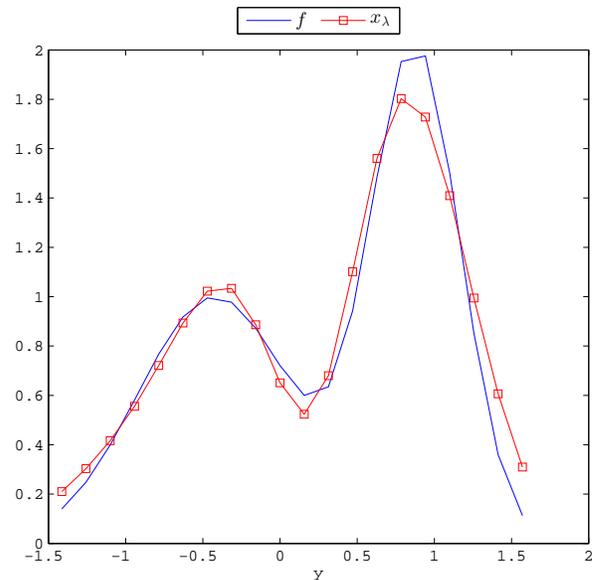
Figura 3.27 mostramos la gráfica de la L-curva

$$(\log_{10} \|\mathbf{b} - A\mathbf{x}_\lambda\|_2, \log_{10} \|\mathbf{x}_\lambda\|_2).$$

La esquina de la la L-curva, marcada por  $\circ$ , corresponde al valor  $\lambda = .0499$ . En la Figura 3.28 comparamos la poligonal  $f$  que tiene los valores de las componentes de  $\mathbf{x}^\dagger$  con la poligonal que tiene los valores de las componentes de  $\mathbf{x}_\lambda$  con  $\lambda = .0499$ . Notamos que la solución regularizadora con este valor del parámetro no oscila y se aproxima a  $f$ .



**Figura 3.27:** L-curva para el problema discreto regularizado de reconstrucción 1D del haz del Ejemplo 2.5. Se marca la esquina  $\circ$  en  $\lambda = .0499$ .



**Figura 3.28:** Poligonal  $f$  con los valores de las componentes de la solución de norma mínima de la ecuación  $A\mathbf{x} = \mathbf{b}$  en Ejemplo 2.5 y poligonal que tiene valores de solución regularizada de Tikhonov  $\mathbf{x}_\lambda$ , con  $\lambda = .0499$

### 3.5.3. Validación Cruzada Generalizada

Consideramos el modelo lineal  $\mathbf{b} = A\mathbf{x} + \boldsymbol{\epsilon}$ , donde  $A$  es una matriz de tamaño  $m \times n$  con  $m \geq n$  y  $\boldsymbol{\epsilon}$  es ruido con distribución gaussiana de varianza  $\eta^2$  que cumple las condiciones de Gauss-Markov. Vamos a buscar el valor del parámetro  $\lambda$  en la regularización de Tikhonov

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left[ \frac{1}{m} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda^2 \|\mathbf{x}\|_2^2 \right]$$

de modo que  $\mathbf{x}_{\text{reg}}$  prediga las observaciones lo mejor posible. La idea es reciclar los datos disponibles. Separamos las observaciones dadas en dos colecciones. Empleamos una de estas colecciones para ajustar un modelo reducido, y la otra para su evaluación. Calculamos

una solución del problema reducido que posteriormente usamos para predecir los elementos de la otra colección. Esta estrategia se llama *Validación Cruzada* [39].

Quitamos una componente del vector de observaciones. Denotemos por  $\mathbf{b}^{(k)}$  al vector que obtenemos al remover la  $k$ -ésima componente de  $\mathbf{b}$ , por  $A_k$  al vector columna que es el  $k$ -ésimo renglón de  $A$ , y por  $\mathbf{A}^{(k)}$  a la matriz que conseguimos al remover  $A_k^T$  de  $A$ . Consideramos el problema reducido

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|A^{(k)}\mathbf{x} - \mathbf{b}^{(k)}\|_2^2$$

Aplicamos regularización de Tikhonov a este problema. Por lo que resolvemos

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left[ \frac{1}{m} \|A^{(k)}\mathbf{x} - \mathbf{b}^{(k)}\|_2^2 + \lambda^2 \|\mathbf{x}\|_2^2 \right]. \quad (3.24)$$

La solución regularizada  $\mathbf{x}_\lambda^{(k)}$  es estimador de la solución de cuadrados mínimos del problema reducido. Usamos éste para predecir la componente removida  $b_k$ . Medimos la discrepancia con el error predictivo

$$P(\lambda) = \frac{1}{m} \sum_{k=1}^m \left( A_k^T \mathbf{x}_\lambda^{(k)} - b_k \right)^2$$

El criterio para elegir el parámetro de regularización consiste en elegir  $\lambda$  que minimize el error predictivo.

Para evitar resolver las  $m$  ecuaciones normales regularizadas que nos dan los estimadores  $\mathbf{x}_\lambda^{(k)}$ , damos otra expresión de  $P(\lambda)$ . Sea  $\mathbf{b}_\lambda^{(k)}$  el vector que obtenemos de  $\mathbf{b}$  al reemplazar  $b_k$  por  $A_k^T \mathbf{x}_\lambda^{(k)}$ . Como

$$\|A\mathbf{x} - \mathbf{b}_\lambda^{(k)}\|_2^2 = \left( A_k^T \mathbf{x} - A_k^T \mathbf{x}_\lambda^{(k)} \right)^2 + \|A^{(k)}\mathbf{x} - \mathbf{b}^{(k)}\|_2^2.$$

y  $\mathbf{x}_\lambda^{(k)}$  es el mínimo del problema (3.24), se sigue que  $\mathbf{x}_\lambda^{(k)}$  resuelve el problema

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left[ \frac{1}{m} \|A\mathbf{x} - \mathbf{b}_\lambda^{(k)}\|_2^2 + \lambda^2 \|\mathbf{x}\|_2^2 \right].$$

Sea

$$A(\lambda) = A(A^T A + m\lambda^2 I)^{-1} A^T.$$

y  $\mathbf{x}_\lambda$  la solución regularizadora de Tikhonov del problema original. Tenemos que

$$A\mathbf{x}_\lambda = A(\lambda)\mathbf{b} \quad \text{y} \quad A\mathbf{x}_\lambda^{(k)} = A(\lambda)\mathbf{b}_\lambda^{(k)},$$

de donde

$$A_k^T \mathbf{x}_\lambda - A_k^T \mathbf{x}_\lambda^{(k)} = A(\lambda)_{kk} \left( b_k - A_k^T \mathbf{x}_\lambda^{(k)} \right).$$

Por lo que

$$(1 - A(\lambda)_{kk}) (b_k - A_k^T \mathbf{x}_\lambda^{(k)}) = b_k - A_k^T \mathbf{x}_\lambda.$$

Sea

$$D(\lambda) = \text{diag} \left( \frac{1}{1 - A(\lambda)_{1,1}}, \dots, \frac{1}{1 - A(\lambda)_{m,m}} \right).$$

Entonces

$$\mathbf{b} - \begin{bmatrix} A_1^T \mathbf{x}_\lambda^{(1)} \\ \vdots \\ A_m^T \mathbf{x}_\lambda^{(m)} \end{bmatrix} = D(\lambda)(\mathbf{b} - A\mathbf{x}_\lambda) = D(\lambda)(I - A(\lambda))\mathbf{b}.$$

En consecuencia,

$$P(\lambda) = \frac{1}{m} \|D(\lambda)(I - A(\lambda))\mathbf{b}\|_2^2. \quad (3.25)$$

Como queremos reciclar los datos disponibles, nos conviene que las columnas de  $A$  estén acopladas. Mediante la **Validación Cruzada Generalizada (GCV)** [120] transformamos el modelo original en otro acoplado cuando hacemos validación cruzada. Para ver esto, realizamos un cambio de variable con las transformaciones ortogonales que nos ofrecen la SVD y la Transformada discreta de Fourier (DFT).

La matriz  $F_{DT}$  con elementos

$$f_{k,j} = \exp(-2\pi k j \mathbf{i} / m) \quad k, j = 1, \dots, m.$$

es simétrica, cumple que

$$\overline{F_{DT}} F_{DT} = mI.$$

y nos da la DFT de  $\mathbf{b}$  como  $F_{DT}\mathbf{b}$  [37]. Por lo que

$$W := 1/\sqrt{m} F_{DT}$$

es un matriz invertible tal que  $W^{-1} = \overline{W}^T$ .

Usamos  $W$  y la SVD  $A = U\Sigma V^T$  con los  $r$  valores singulares positivos de  $A$  para transformar  $\mathbf{b} = A\mathbf{x} + \boldsymbol{\epsilon}$  en el sistema de ecuaciones lineales

$$\hat{\mathbf{b}} = \hat{A}\mathbf{x} + \hat{\boldsymbol{\epsilon}},$$

donde

$$\hat{\mathbf{b}} = WU^T \mathbf{b}, \quad \hat{\boldsymbol{\epsilon}} = WU^T \boldsymbol{\epsilon}, \quad \hat{A} = W\Sigma V^T.$$

En este nuevo modelo

$$\hat{A}^T \hat{A} = F_{DT} \text{diag}(\mathbf{d}) F_{DT}^{-1},$$

donde

$$d_k = \begin{cases} \sigma_k^2, & \text{si } k \in \{1, \dots, r\} \\ 0, & \text{si } k \in \{r+1, \dots, m\}. \end{cases}$$

Por lo que  $\widehat{A}^T \widehat{A}$  tiene como primera columna a

$$\mathbf{c} = 1/m F_{DT} \mathbf{d}$$

y es de la forma

$$\widehat{A}^T \widehat{A} = \begin{bmatrix} c_1 & c_m & c_{m-1} & \cdots & c_2 \\ c_2 & c_1 & c_m & \cdots & c_3 \\ c_3 & c_2 & c_1 & \cdots & c_4 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_m & c_{m-1} & c_{m-2} & \cdots & c_1 \end{bmatrix}.$$

Las matrices con esta estructura se llaman *matrices circulares* [37].

### Observaciones 3.22:

👁️ A diferencia de  $A^T A$ , los renglones de  $\widehat{A}^T \widehat{A}$  están acoplados porque es circular.

Considere el problema de cuadrados mínimos

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\widehat{A}\mathbf{x} - \widehat{\mathbf{b}}\|_2^2, \quad (3.26)$$

La *estimación GCV* del parámetro  $\lambda$  en la regularización de Tikhonov del problema discreto mal planteado  $A\mathbf{x} = \mathbf{b}$  es el valor  $\lambda_{GCV}$  que minimiza el error predictivo  $\widehat{P}(\lambda)$  al remover cada componente de  $\widehat{\mathbf{b}}$  y predecirla por la solución regularizadora del problema (3.26) reducido correspondiente.

Sea

$$\widehat{A}(\lambda) := \widehat{A}(\widehat{A}^T \widehat{A} + m\lambda^2 I)^{-1} \widehat{A}^T.$$

De la misma manera en que obtuvimos  $P(\lambda)$ , tenemos que

$$\widehat{P}(\lambda) = m \frac{\|(I - \widehat{A}(\lambda))\widehat{\mathbf{b}}\|_2^2}{[\text{Tr}(I - \widehat{A}(\lambda))]^2}.$$

Por otra parte, si en  $D(\lambda)$  reemplazamos cada elemento de la diagonal principal de  $A(\lambda)$  por su promedio y sustituimos en la expresión (3.25) de  $P(\lambda)$ , obtenemos

$$V(\lambda) := m \frac{\|(I - A(\lambda))\mathbf{b}\|_2^2}{[\text{Tr}(I - A(\lambda))]^2}.$$

Podemos probar [39] que

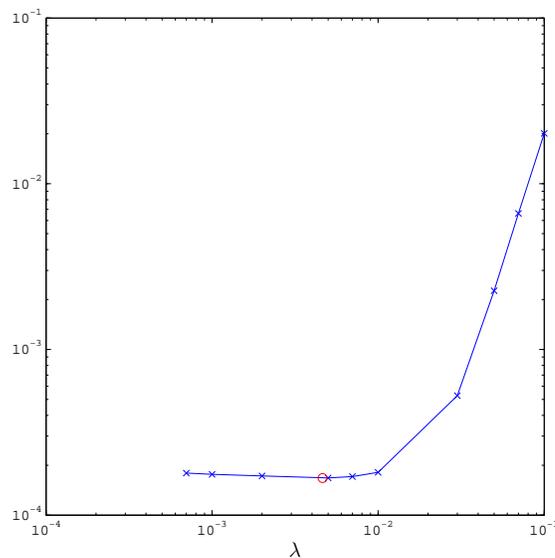
$$\widehat{P}(\lambda) = V(\lambda) = m \frac{\sum_{k=1}^r \left( \frac{m\lambda^2}{\sigma_k^2 + m\lambda^2} \right)^2 (\mathbf{u}_k^T \mathbf{b})^2 + \sum_{k=r+1}^m (\mathbf{u}_k^T \mathbf{b})^2}{\left[ \sum_{k=1}^r \left( \frac{m\lambda^2}{\sigma_k^2 + m\lambda^2} \right) + m - r \right]^2}. \quad (3.27)$$

Esto nos dice que la GCV es una forma de validación cruzada que es invariante bajo rotaciones (la matriz  $V$  de la SVD y  $W$  de la DFT).  $V$  se conoce como la *función GCV*.

**Observaciones 3.23:**

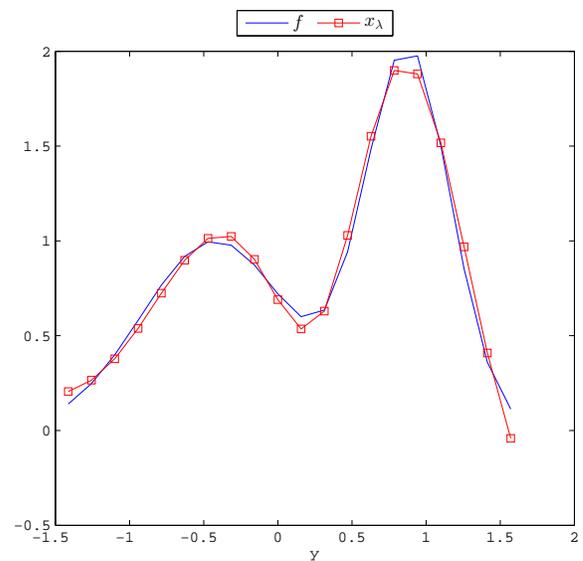
- ☞  $\lambda_{GCV}$  es mínimo de  $V(\lambda)$  para  $\lambda > 0$
- ☞ Con la expresión (3.27) podemos calcular la función GCV a partir de la SVD de  $A$ .
- ☞ Para un sistema de ecuaciones lineales indeterminado, Nguyen, Milanfar y Golub [91] verifican que la función GCV nuevamente está dada por (3.27).

**Ejemplo 3.12.** En el problema discreto de reconstrucción del haz de luz (Ejemplo 2.5) usamos regularización de Tikhonov para recuperar la solución  $\mathbf{x}^\dagger$  de cuadrados mínimos de norma mínima de  $A\mathbf{x} = \mathbf{b}$  a partir de la ecuación  $A\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}$  con  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, 0.01I)$ .



**Figura 3.29:** Función GCV para el problema discreto regularizado de reconstrucción 1D del haz del Ejemplo 2.5.

Marcamos por  $\circ$  el mínimo  
 $\lambda_{GCV} = 4.6305 \times 10^{-3}$ .



**Figura 3.30:** Solución aproximada  $f$  de la ecuación  $A\mathbf{x} = \mathbf{b}$  del Ejemplo 2.5 y poligonal con los valores de componentes de solución regularizada de Tikhonov  $\mathbf{x}_\lambda$  con  $\lambda = \lambda_{GCV}$ .

En esta ocasión usamos la estimación GCV para elegir el parámetro de regularización  $\lambda$ . En la Figura 3.29 mostramos la gráfica de la función GCV  $V$ . El mínimo que encontramos de  $V$  es  $\lambda_{GCV} = 4.6305 \times 10^{-3}$ . En la Figura 3.30 comparamos la poligonal  $f$  con los valores de las componentes de  $\mathbf{x}^\dagger$  con la poligonal que tiene valores iguales a la solución regularizadora  $\mathbf{x}_{\lambda_{GCV}}$ . Las discrepancias entre ambas poligonales son pequeñas.

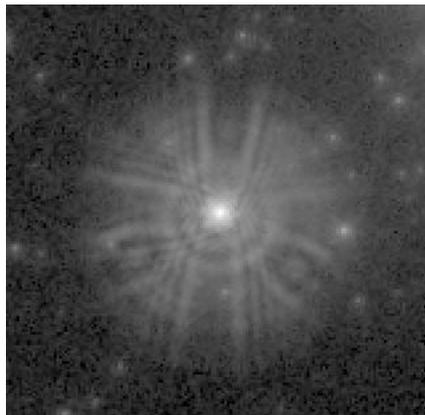


Estamos interesados en tratar con problemas computacionalmente intensivos del procesamiento de imágenes que involucren decenas de miles de incógnitas. Por ejemplo, si queremos reducir el ruido en una sucesión de imágenes de  $500 \times 500$  píxeles y ampliarlas en un factor de 4 en cada dimensión espacial, generamos una imagen que involucra  $2000 \times 2000 = 4 \times 10^6$  valores de píxeles desconocidos.

Los problemas que abordamos son de gran escala y están mal planteados. El inconveniente que tenemos en resolverlos con métodos directos de regularización como SVD truncada es que es costoso calcular la SVD de matrices de grandes dimensiones. Por eso nos interesa examinar la estructura del problema de gran escala, antes de regularizar.

Para empezar, damos un visión panorámica del procesamiento de imágenes. Una imagen puede pensarse como una distribución continua de colores o escala de grises sobre el plano. Ésta contiene información de una escena que nos interesa almacenar, transmitir y/o interpretar. Para tratar de acceder a esta información usamos dispositivos como cámaras digitales, microscopios y telescopios electrónicos. Estos recogen muestras que conocemos como píxeles y la información queda almacenada en bits. De ese modo, obtenemos imágenes digitales que podemos modificar en la PC.

En el procesamiento de imágenes, analizamos y manipulamos la información que contienen para extraer atributos o obtener una nueva imagen. El interés va desde retocar fotografías de celular, el reconocimiento facial en sistemas de vigilancia, identificar vías terrestres por satélite, hasta detectar nebulosas con la ayuda de sensores digitales.



**Figura 4.1:** Imagen de la estrella PC5 tomada por el telescopio Hubble en 1990.

A veces las imágenes captadas por telescopios, microscopios, cámaras de celular, entre otros, presentan degradaciones debidas a desenfoco del lente, movimiento del objeto, ruido, etc. Un caso famoso es el *Telescopio Espacial Hubble*. Poco después de su lanzamiento, 24 de Abril de 1990, el personal de la NASA descubrió que el telescopio tenía un lente defectuoso. Esto provocó que la luz de las escenas tomadas se propagara en gran parte de

las imágenes. Como la reparación del lente era complicada, el equipo de trabajo se vió en la necesidad de restaurar las imágenes tomadas por el Hubble [3].

En la actualidad, científicos e ingenieros usan el procesamiento de imágenes para comprender mejor los problemas de sus respectivas áreas, de hecho, se ha vuelto parte integral de su formación profesional. El potencial que tiene esta área, no se limita a los especialistas, pues gracias a los avances tecnológicos y computacionales, la gente puede manipular fácilmente algunas imágenes cuando hace veinte años no resultaba viable.



(a) imagen en espectro infrarrojo



(b) ultrasonido de un feto



(c) radiografía



(d) imagen en espectro visible

**Figura 4.2:** Imágenes digitales obtenidas por fuentes distintas

Las áreas de aplicación del procesamiento de imágenes son diversas. Una manera de clasificarlas es de acuerdo a la fuente que las genera. La principal fuente es la radiación del espectro electromagnético [40]. Las imágenes tomadas por cámaras comerciales están en el espectro visible, los médicos usan imágenes captadas por rayos x para hacer diagnósticos, los satélites usan radiación infrarroja para generar imágenes de una red de asentamientos humanos. Otras fuentes como ultrasonidos nos permiten obtener imágenes del estado de un feto. Véase Figura 4.2. Nosotros trabajamos con imágenes digitales captadas en el espectro visible.

El procesamiento de imágenes está repleto de problemas, que en un principio se pensaba que eran relativamente fáciles de resolver, pero hasta la fecha no revelan sus secretos. Conforme avanza la tecnología, algunas tareas se hacen más sencillas y a la vez surgen nuevos problemas. A grandes rasgos, los problemas que se presentan a menudo consisten en transformar las imágenes digitales o extraer sus atributos. De acuerdo a la aplicación, abordamos diferentes metodologías [40]:

**Compresión.** Se intenta reducir el almacenamiento en bits de la imagen o aumentar la capacidad transmisión. Esto es práctico cuando queremos descargar imágenes de una página web. De hecho, convivimos a diario con los formatos de compresión JPG y PNG.

**Procesamiento del color.** Lo que se pretende es discernir aspectos visuales de la imagen a partir de la naturaleza física del color. Mediante modelo del color, se intentan manipular características como la nitidez y la saturación

**Realce.** Nos interesa dar una mejor interpretación visual de la imagen. Con ese fin resaltamos atributos de interés o damos a conocer detalles escondidos en la imagen. Podemos manipular el contraste o sacar el negativo de una foto. Esto es útil en la Fotografía así como en el procesamiento de imágenes médicas.

**Restauración.** Se pretende recuperar una imagen ideal a partir de otra de la misma escena que tiene degradaciones. Hacemos uso del conocimiento a priori del fenómeno que genera las degradaciones. La idea es dar un modelo del proceso de degradación. Este es el caso cuando las imágenes están desenfocadas.

**Reconstrucción 3D.** A partir de imágenes de las proyecciones de un objeto tridimensional, reconstruimos su superficie. Las proyecciones pueden ser secciones transversales o sombras del objeto. La tomografía médica y sísmica hacen uso de estas reconstrucciones.

**Procesamiento morfológico.** Nos concentramos en extraer información relevante sobre la forma que tiene un objeto en una imagen. Tenemos diferentes grados de dificultad dependiendo de la forma, color o textura de los objetos presentes en la escena. La teoría de conjuntos da el fundamento teórico.

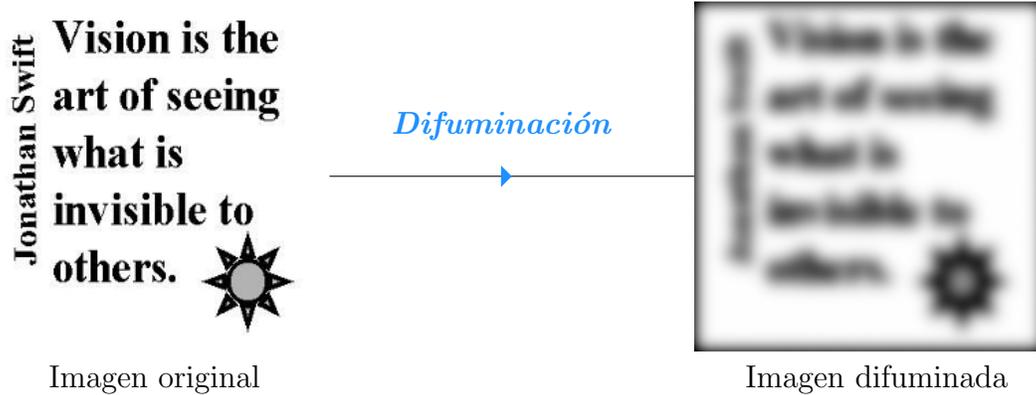
**Segmentación.** Separamos una imagen en regiones que corresponden a objetos presentes en la escena. Para separar, nos fijamos en los cambios de valores de los píxeles o usamos un criterio predefinido. Podemos aislar puntos, líneas y contornos.

Entre los diferentes problemas del procesamiento de imágenes, nosotros nos enfocamos en dos: *deblurring* y *super-resolución*.

### P1. Deblurring

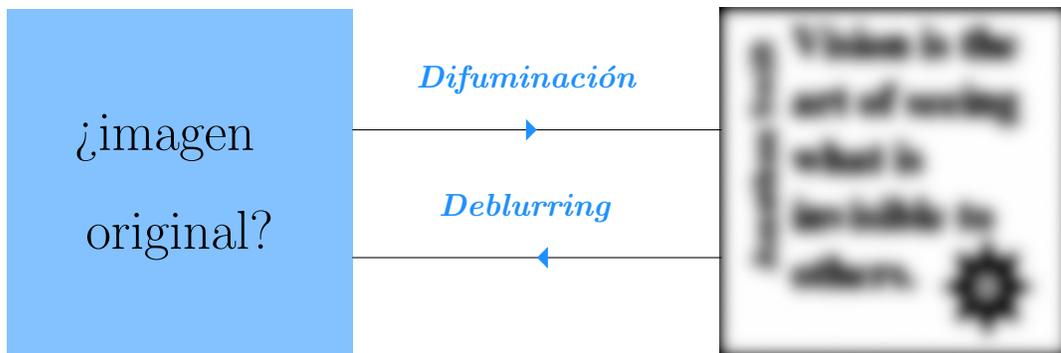
La difuminación de una imagen consiste en dispersar el brillo de una zona a sus alrededores. Esto genera degradaciones en la imagen. Véase Figura 4.3. Queremos ver la imagen con menos degradación. Esto nos conduce al problema de *Deblurring*:

*Dada una imagen difuminada  $\mathcal{G}$ , obtener una imagen  $\mathcal{F}$  de la misma escena con menos degradación.*



**Figura 4.3:** Difuminación de una imagen.

Si nuestro problema es difuminar una imagen, el Deblurring puede verse como el problema inverso de la difuminación. Véase Figura 4.4.



**Figura 4.4:** Deblurring como problema inverso de la difuminación.

## P2. Super-resolución

Queremos usar la información de varias imágenes de una misma escena para obtener una imagen con mejor resolución. Esto da lugar al problema de *Super-resolución*:

*Dada una colección finita de imágenes  $\mathcal{G}_1, \dots, \mathcal{G}_L$  de baja resolución de una misma escena, obtener una imagen  $\mathcal{F}$  de alta resolución.*



**Figura 4.5:** Super-resolución. A partir de dos imágenes desenfocadas  $\mathcal{G}_1$  y  $\mathcal{G}_2$  de una placa de automóvil de  $118 \times 90$  píxeles, obtener una imagen  $\mathcal{F}$  con menos degradaciones de  $236 \times 180$  píxeles

## 4.1. Deblurring

Si queremos resolver el deblurring, la idea que usamos es que éste es el problema inverso de la difuminación. Por eso buscamos un modelo que nos permita obtener una imagen difuminada a partir de la imagen ideal de la escena.

El modelo de difuminación que damos está descrito por un sistema que nos permite atenuar o suprimir características de una imagen. Este sistema se conoce como *filtro*. Pensemos en un sistema óptico que recibe a la imagen  $\mathcal{F}$  y devuelve la imagen difuminada  $\mathcal{G}$ . La imagen ideal  $\mathcal{F}$  se encuentra en un espacio de imágenes  $U$ , mientras que la imagen difuminada  $\mathcal{G}$  está en un espacio de imágenes  $V$ . El sistema transforma  $\mathcal{F}$  en  $\mathcal{G}$  mediante un operador  $\mathcal{B} : U \rightarrow V$ . Con este operador, la difuminación de imágenes consiste en lo siguiente:

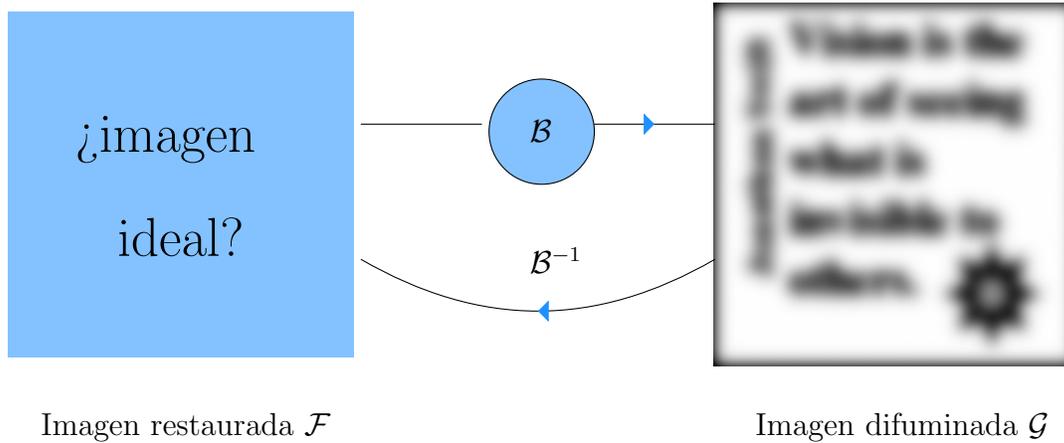
Dada la imagen  $\mathcal{F} \in U$  y el operador  $\mathcal{B} : U \rightarrow V$ , obtener  $\mathcal{G} = \mathcal{B}[\mathcal{F}]$ .

Así, el deblurring (el problema inverso) es

Dada la imagen difuminada  $\mathcal{G} \in V$  y el operador  $\mathcal{B} : U \rightarrow V$ , determinar la imagen  $\mathcal{F} \in U$  tal que  $\mathcal{G} = \mathcal{B}[\mathcal{F}]$ .

### Observaciones 4.1:

 El deblurring tiene solución si el operador  $\mathcal{B}$  es sobreyectivo. Su solución es única si  $\mathcal{B}$  es inyectivo. Por lo que si conocemos la inversa de  $\mathcal{B}$ , podemos restaurar la imagen difuminada. Véase Figura 4.6.



**Figura 4.6:** La inversa del operador  $\mathcal{B}$  para restaurar imágenes difuminadas en el problema de Deblurring.

#### 4.1.1. Modelo de difuminación

Vamos a dar un modelo continuo de difuminación. Pensemos en una imagen dada  $\mathcal{F}$  como una distribución continua de colores en el plano. Le vamos a asociar una función  $f$ . Si bien la imagen se delimita a un rectángulo  $R$ , podemos extenderla sobre todo  $\mathbb{R}^2$  mediante condiciones de frontera. Por ejemplo, que la imagen tenga fondo negro fuera de  $R$ . Así que tomamos a  $\mathbb{R}^2$  como el dominio de  $f$ . Por otra parte, nosotros manejamos imágenes en escala de grises. Para indicar el tono de gris, asignamos un número en el intervalo  $[0, 1]$  a cada punto de  $\mathbb{R}^2$ . El valor cero indica el color negro, mientras que el valor uno representa el color blanco. De esa manera, asociamos una función  $f : \mathbb{R}^2 \rightarrow [0, 1]$  a la imagen.

Los espacios  $U$  y  $V$  de imágenes que tratamos tienen la estructura de espacios vectoriales. Ambos contienen a todas las funciones definidas sobre  $\mathbb{R}^2$  con valores en el intervalo  $[0, 1]$ .

Una vez que conocemos la estructura del espacio de imágenes, buscamos la del operador de difuminación  $\mathcal{B}$ .

##### Hipótesis del modelo

La mayoría de los sistemas físicos reciben una señal en un espacio vectorial  $U$  y regresan una nueva señal en un espacio vectorial  $V$  mediante una transformación  $\mathcal{B} : U \rightarrow V$  que cumple con las siguientes propiedades:

\* **Linealidad.** Se verifica el principio de superposición:

$$\mathcal{B}[\alpha f_1 + f_2] = \alpha \mathcal{B}[f_1] + \mathcal{B}[f_2], \quad \forall \alpha \in \mathbb{R}, \forall f_1, f_2 \in U.$$

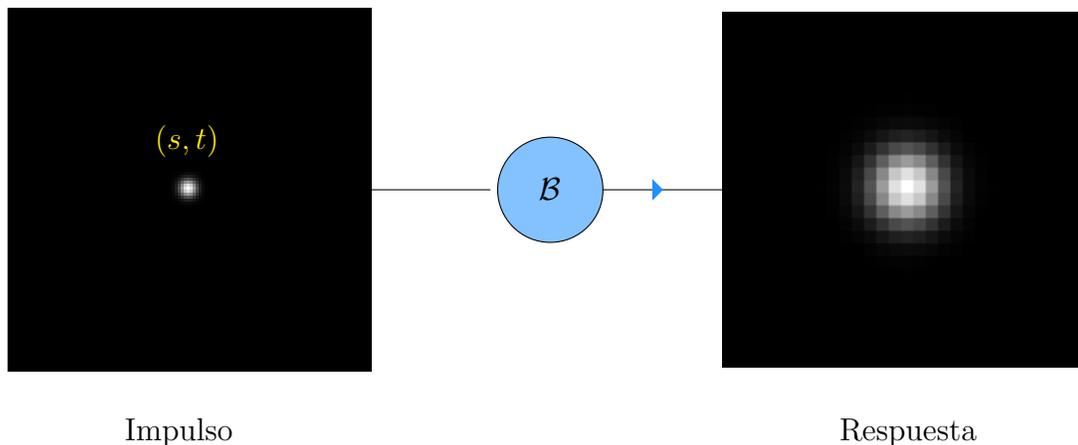
\* **Invarianza bajo traslación.** Un desplazamiento de la señal corresponde a un desplazamiento en la respuesta: Si  $\mathcal{B}[f] = g$ , entonces

$$\mathcal{B}[f](x + s, y + t) = g(x + s, y + t) \quad \forall x, y, s, t \in \mathbb{R}.$$

La clave es que si el filtro cumple con las propiedades anteriores, entonces podemos caracterizar el sistema con la respuesta que produce una imagen de color negro, salvo un punto brillante  $(s, t)$ . Esta imagen representa el impulso que recibe el sistema. Mediante la transformación  $\mathcal{B}$ , la luz que emite el punto brillante se dispersa a los puntos vecinos. De ese modo el sistema produce como respuesta la imagen difuminada. Véase la Figura 4.7.

Lo que hace el operador  $\mathcal{B}$  es “transformar” el impulso unitario  $\delta$  trasladado al punto  $(s, t)$  en una función  $k : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$  que nos da la respuesta del sistema:

$$\delta(x - s, y - t) \xrightarrow{\mathcal{B}} k((x, y), (s, t)) \quad (4.1)$$



**Figura 4.7:** El sistema dado por el operador  $\mathcal{B}$  hace que la luz del punto brillante  $(s, t)$  del impulso se disperse en la respuesta que produce.

#### Observaciones 4.2:

👉 Aclaremos que el impulso unitario no es una función sobre el plano, si no que es una distribución [126].

#### Función de Dispersión de punto

La respuesta  $k$  se conoce como **función de dispersión de punto (PSF)**. El punto brillante  $(s, t)$  donde se traslada el impulso es el **centro de la PSF**.



**Figura 4.8:** Descomposición de  $f$  en funciones elementales  $p_{i,j}$  que se transforman bajo el operador  $\mathcal{B}$  y se reintegran por linealidad en  $g = \mathcal{B}[f]$ .

Vamos a dar una expresión al operador  $\mathcal{B}$  a partir de la PSF. La idea es expandir  $f$  en funciones elementales. Como  $\mathcal{B}$  es lineal, podemos combinar las imágenes de esas funciones bajo  $\mathcal{B}$  para obtener la respuesta tal y como se muestra en la Figura 4.8. Al respecto, expandimos  $f$  mediante pulsos

$$p_{\Delta s, \Delta t}(x, y) = \begin{cases} \frac{1}{\Delta s \Delta t}, & \text{si } 0 \leq x < \Delta s, 0 \leq y < \Delta t, \\ 0, & \text{en otro caso} \end{cases}$$

como

$$f(x, y) = \lim_{\Delta s, \Delta t \rightarrow 0} \sum_{i, j \in \mathbb{Z}} f(i\Delta s, j\Delta t) p_{\Delta s, \Delta t}(x - i\Delta s, y - j\Delta t) \Delta s \Delta t.$$

En términos del impulso unitario, esto se representa simbólicamente como

$$f(x, y) = \int \int_{\mathbb{R}^2} f(s, t) \delta(x - s, y - t) ds dt. \quad (4.2)$$

La integral del lado derecho puede verse como una combinación lineal de impulsos unitarios trasladados, donde los valores de la función  $f$  son los coeficientes [106].

Extendemos la linealidad del operador  $\mathcal{B}$  [4], [125] de modo que

$$\mathcal{B} \left[ \int \int_{\mathbb{R}^2} f(s, t) \delta(x - s, y - t) ds dt \right] = \int \int_{\mathbb{R}^2} f(s, t) \mathcal{B}[\delta(x - s, y - t)] ds dt,$$

Luego, como  $\mathcal{B}$  transforma el impulso  $\delta$  en la PSF  $k$ , tenemos que

$$\mathcal{B} \left[ \int \int_{\mathbb{R}^2} f(s, t) \delta(x - s, y - t) ds dt \right] = \int \int_{\mathbb{R}^2} f(s, t) k((x, y), (s, t)) ds dt,$$

Por consiguiente, de la identidad (4.2) se sigue que  $\mathcal{B}$  es el operador integral

$$\mathcal{B}[f](x, y) = \int \int_{\mathbb{R}^2} f(s, t) k((x, y), (s, t)) ds dt, \quad \forall x, y \in \mathbb{R}.$$

Una condición suficiente para que el operador  $\mathcal{B}$  cumpla la hipótesis de invarianza bajo traslación es que

$$k((x, y), (s, t)) = k(x - s, y - t) \quad \forall x, y, s, t \in \mathbb{R},$$

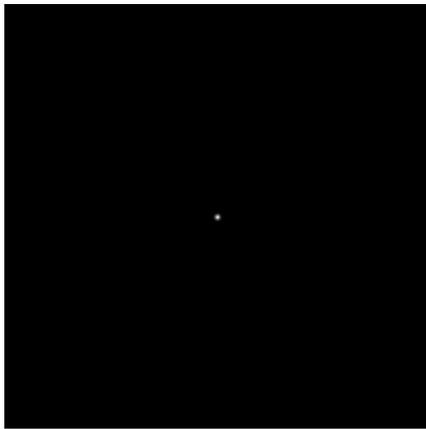
La PSF  $k$  dada de esta manera se conoce como *espacialmente invariante*. Usamos esta PSF con las siguientes restricciones:

- \*  $k(x, y) \geq 0$  para cualesquiera  $x, y \in \mathbb{R}$ . Esto se debe a las variables físicas como intensidad de luz solamente toman valores no negativos.
- \*  $\int_{\mathbb{R}^2} k(x, y) dx dy = 1$  para que la energía del sistema se conserve.

### Observaciones 4.3:

- ☞ La función núcleo del operador integral  $\mathcal{B}$  es la PSF  $k$ .
- ☞ En la PSF espacialmente invariante hacemos un abuso de notación al pasar de una función definida sobre  $\mathbb{R}^2 \times \mathbb{R}^2$  a otra definida sobre  $\mathbb{R}^2$ .
- ☞ Si desconocemos la PSF, podemos hacer una estimación de ésta. En la Literatura [6], se propone aproximarla por otras PSF's que sean espacialmente invariantes localmente.

**Ejemplo 4.1.** Mostramos algunas PSF espacialmente invariantes de la Literatura [55], [73]. Para ilustrar cada una, vemos como se difumina la imagen `cameraman.png` y una imagen de fondo negro con un píxel blanco en el centro.



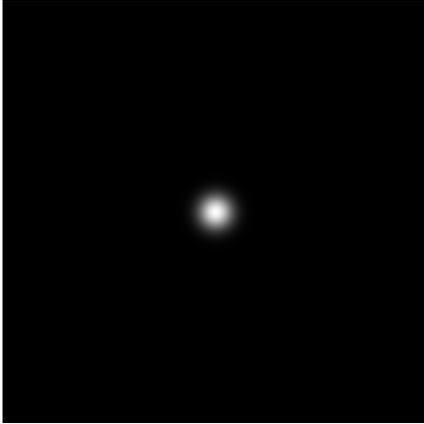
**Figura 4.9:** Imagen asociada al impulso unitario.



**Figura 4.10:** Imagen `cameraman`.

- \* La PSF gaussiana de varianza  $\sigma^2$  con media  $(s, t)$  es

$$k(x, y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-s)^2 + (y-t)^2}{2\sigma^2}\right).$$



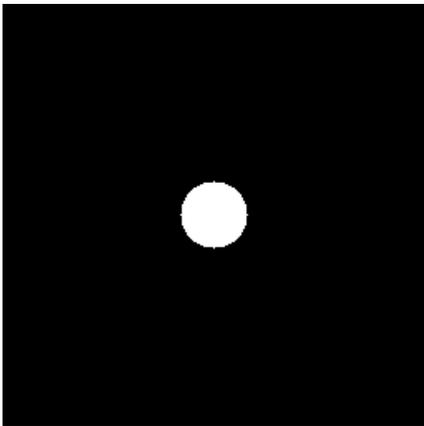
**Figura 4.11:** PSF gaussiana con desviación estándar  $\sigma = 7$  y valor esperado  $(s, t) = (128, 128)$ .



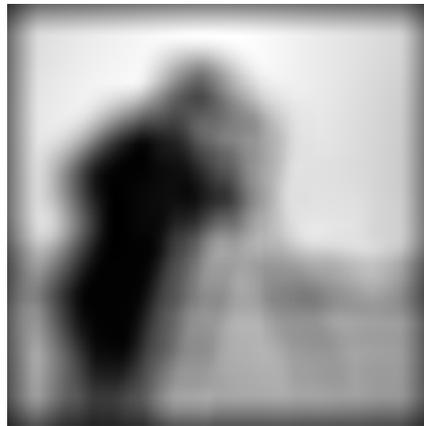
**Figura 4.12:** Imagen cameraman difuminada por PSF gaussiana.

- \* La PSF para desenfoque en un círculo de radio  $r$  con centro  $(s, t)$  es

$$k(x, y) = \begin{cases} \frac{1}{\pi r^2}, & \text{si } (x-s)^2 + (y-t)^2 \leq r, \\ 0, & \text{en otro caso.} \end{cases}$$



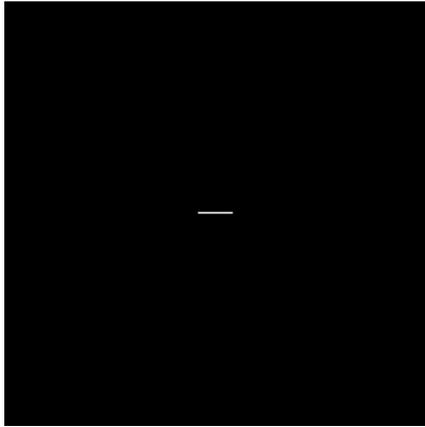
**Figura 4.13:** PSF de desenfoque con radio  $r = 20$ .



**Figura 4.14:** Imagen cameraman desenfocada.

\* La PSF de traslación horizontal a partir del punto  $(s, t)$  con desplazamiento  $L$  es

$$k(x, y) = \begin{cases} \frac{1}{L}, & \text{si } |x - s| \leq \frac{L}{2} \text{ e } y = t, \\ 0, & \text{en otro caso.} \end{cases}$$



**Figura 4.15:** PSF de movimiento horizontal con desplazamiento  $L = 20$ .



**Figura 4.16:** Imagen cameraman difuminada por PSF de movimiento horizontal.

### Convolución como Modelo de Difuminación

Mediante la PSF espacialmente invariante, nuestro operador de difuminación  $\mathcal{B} : U \rightarrow V$  está dado por

$$\mathcal{B}[f](x, y) = \int \int_{\mathbb{R}^2} k(x - s, y - t) f(s, t) ds dt. \quad \forall x, y \in \mathbb{R},$$

esto es,  $\mathcal{B}$  realiza la convolución

$$B[f] = k * f.$$

Vamos a pedir que la PSF sea cuadrado integrable:

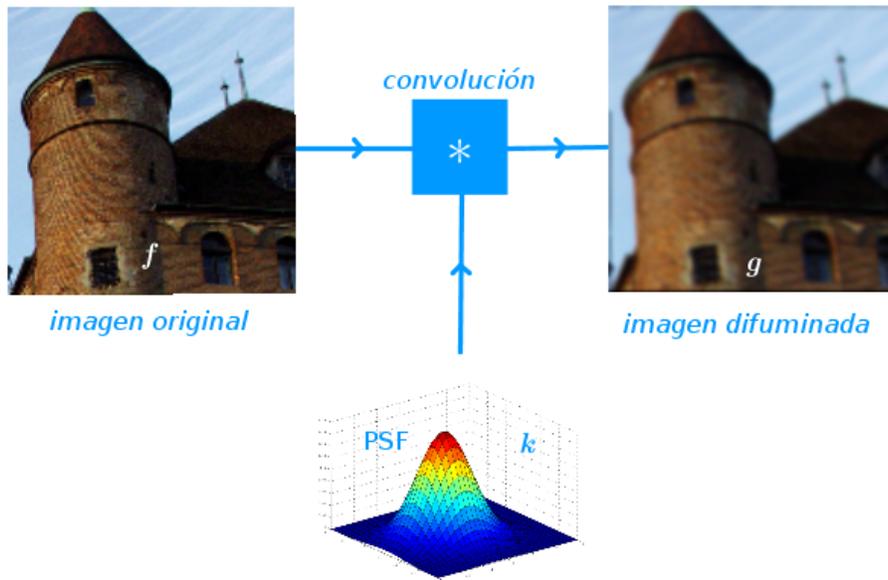
$$\int_{\mathbb{R}^2} \left( \int_{\mathbb{R}^2} [k((x, y), (s, t))]^2 dx dy \right) ds dt < \infty.$$

De ese modo,  $\mathcal{B}$  transforma funciones en  $L^2(\mathbb{R}^2)$  en otras del mismo espacio. Así que para nuestros fines, tomamos  $U = V = L^2(\mathbb{R}^2)$ .

Podemos pensar que la convolución es el desenfoque del lente de una cámara digital. En ese caso, la imagen ideal es la escena reflejada en uno de los lentes de la cámara, otro de los lentes hace el papel de la PSF, el mecanismo que desenfoca la imagen realiza la convolución y la imagen desenfocada se observa en la pantalla. Véase Figura 4.17.



**Figura 4.17:** Convención como el desenfoco del lente de una cámara



**Figura 4.18:** Convención como difuminación por PSF espacialmente invariante.

Así que nuestro modelo genera la imagen difuminada como la convención  $g = f * k$ . Veamos que hace geoméricamente la convención con un ejemplo en 1D.

**Ejemplo 4.2.** Sean  $f, k : \mathbb{R} \rightarrow \mathbb{R}$  las funciones dadas por

$$f(s) = \begin{cases} s, & \text{si } 0 < s < 2, \\ 0 & \text{en otro caso,} \end{cases} \quad \text{y} \quad k(s) = \begin{cases} \pi, & \text{si } 0 < s < 1, \\ 0 & \text{en otro caso.} \end{cases}$$

Para generar el valor en cada punto  $x$  de la convolución  $g = f * k$ , procedemos de la siguiente manera:

1. Reflejamos la función  $k$  respecto al eje de las ordenadas.
2. Trasladamos  $k(-s)$  a  $k(x - s)$ .
3. Multiplicamos  $k(x - s)$  por la función  $f$ ,
4. Calculamos el área bajo la curva de  $k(x - s)f(s)$  sobre toda la recta real.

Conforme aumentamos el valor de  $x$ , vamos generando la función  $g$ . En la Figura 4.19 mostramos esta construcción de  $g$ .

Sea  $I_x = [x - 1, x] \cap [0, 2]$ . Entonces

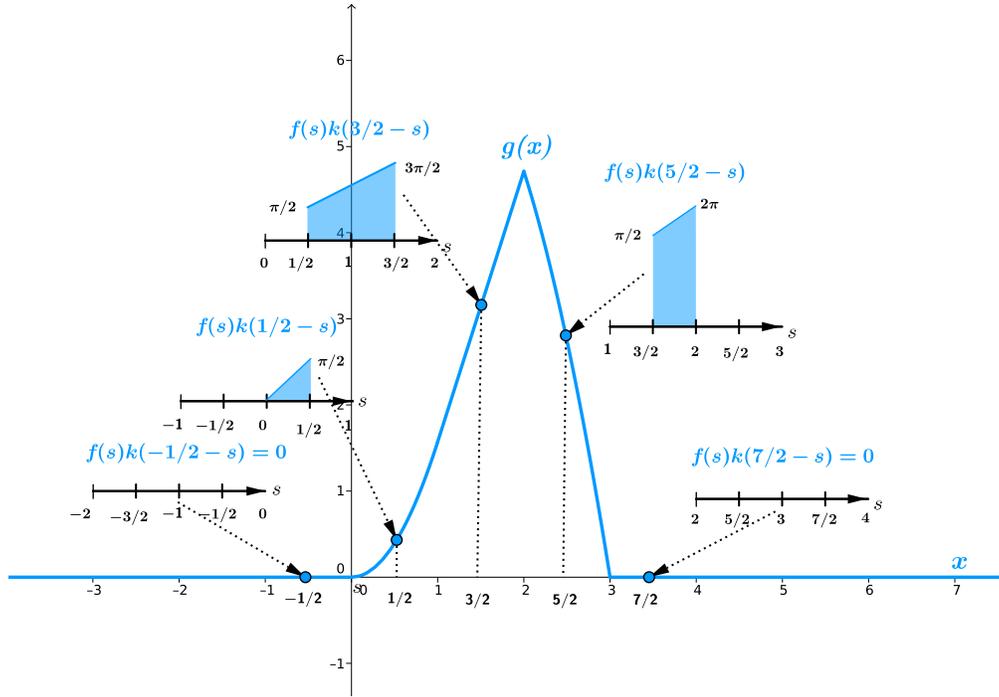
$$f(s)k(x - s) = \begin{cases} \pi s, & \text{si } s \in I_x, \\ 0, & \text{en otro caso.} \end{cases}$$

De acuerdo al valor de  $x$ , tenemos que

$$I_x = \begin{cases} \emptyset, & \text{si } x < 0 \text{ o } x > 3, \\ [0, x], & \text{si } 0 \leq x \leq 1, \\ [x - 1, x], & \text{si } 1 \leq x \leq 2, \\ [x - 1, 2], & \text{si } 2 \leq x \leq 3. \end{cases}$$

En consecuencia,

$$g(x) = \begin{cases} 0 & \text{si } x \leq 0 \text{ o } x \geq 3, \\ \frac{\pi}{2}x^2 & \text{si } 0 \leq x \leq 1, \\ \frac{\pi}{2}(2x - 1) & \text{si } 1 \leq x \leq 2, \\ \frac{\pi}{2}(3 - x^2 + 2x) & \text{si } 2 \leq x \leq 3. \end{cases}$$



**Figura 4.19:** La función  $g$  es la convolución de las funciones de soporte compacto  $f$  y  $k$ . El valor de  $g(x)$  es el área bajo la curva de la transformación  $s \rightarrow f(s)g(x - s)$ .

Mencionamos algunas propiedades de la convolución:

\* **Linealidad.** Dadas funciones integrables  $k, f_1, f_2 : \mathbb{R}^2 \rightarrow \mathbb{R}$  y la constante  $\alpha \in \mathbb{R}$ , tenemos que

$$k * (\alpha f_1 + f_2) = \alpha(k * f_1) + k * f_2$$

\* **Conmutatividad.** Dadas funciones integrables  $k, q : \mathbb{R}^2 \rightarrow \mathbb{R}$ , tenemos que

$$\int \int_{\mathbb{R}^2} k(x - s, y - t)q(s, t)dsdt = \int \int_{\mathbb{R}^2} q(x - s, y - t)k(s, t)dsdt \quad \forall x, y \in \mathbb{R},$$

esto es,  $k * q = q * k$ .

Ahora que tenemos nuestro modelo lineal de difuminación dado por la convolución del imagen ideal con la PSF, podemos dar una formulación matemática del deblurring. Recordamos que éste es el problema inverso de difuminación. Por lo que el deblurring donde la PSF es espacialmente invariante se trata de una deconvolución. Véase la Figura 4.20.

### Problema de deblurring para PSF espacialmente invariante

Dadas las funciones  $g \in L^2(\mathbb{R}^2)$  y  $k : \mathbb{R}^2 \rightarrow \mathbb{R}$ , hallar una función  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  que cumpla la ecuación integral

$$g(x, y) = \int \int_{\mathbb{R}^2} k(x - s, y - t) f(s, t) ds dt. \quad \forall x, y \in \mathbb{R}. \quad (4.3)$$

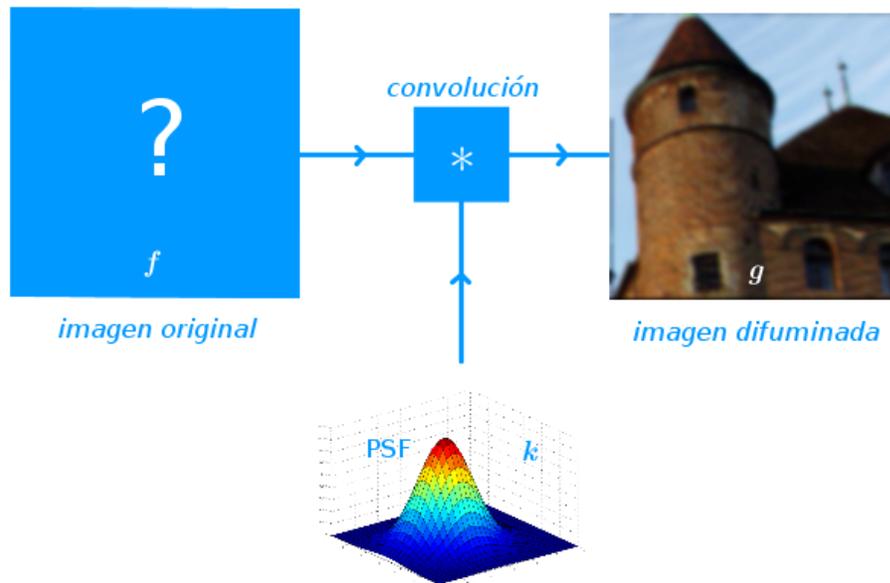


Figura 4.20: Deblurring como problema de deconvolución

#### Observaciones 4.4:

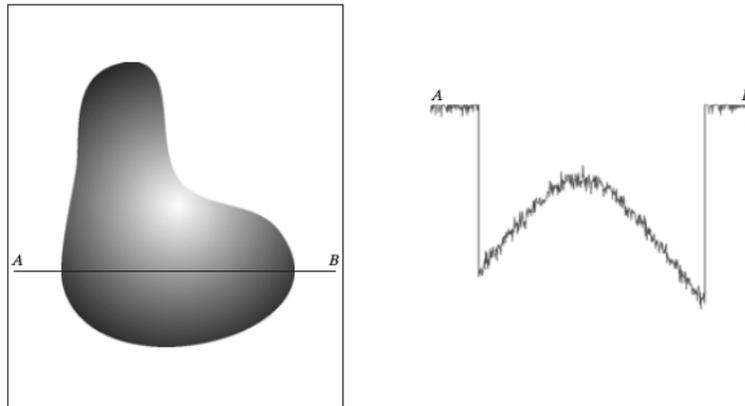
Debido a que los dispositivos ópticos-electrónicos solamente nos ofrecen muestras de la imagen difuminada, necesitamos una versión discreta del modelo de difuminación. Por ello, discretizamos la ecuación integral (4.3).

#### 4.1.2. Digitalización de Imágenes

Las cámaras y otros dispositivos para adquirir imágenes manejan una resolución limitada y algunas intensidades de colores. Por lo que tenemos una aproximación discreta de la imagen. La **digitalización** de una imagen consiste en transformar la función de variables y valores continuos asociada a la imagen en otra de variables y valores discretos. La digitalización se realiza mediante los procesos de **muestreo** y **cuantificación**. De este modo se adquieren imágenes digitales. Para comprender esto relacionamos las imágenes con señales.

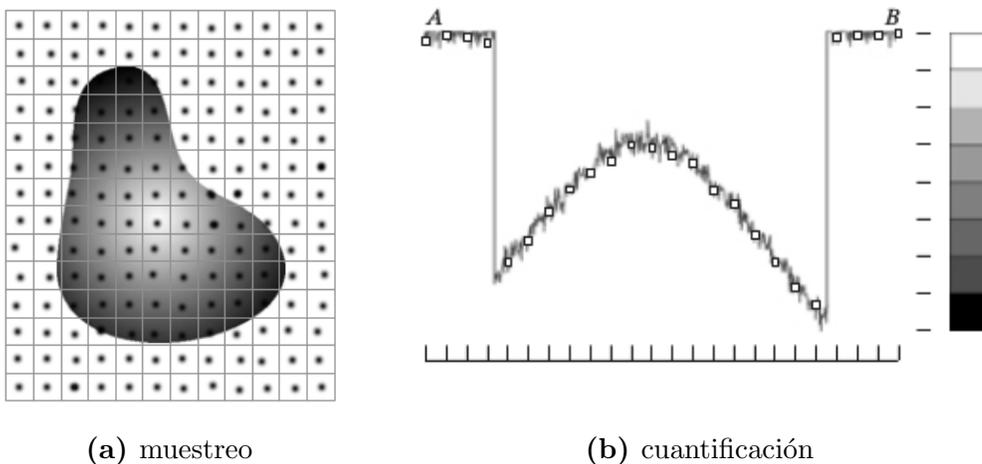
Cuando tomamos una foto, los sensores de una cámara digital devuelven señales con forma de onda que constituyen una sección transversal de la imagen. La intensidad de gris asociada a esa sección está dada por la amplitud de la señal. Véase la Figura 4.21. Cada señal es una función de variable y valores continuos.

Cuando hacemos muestreo, transformamos las variables continuas en discretas. Se crea una partición de cada sección transversal. Cuando las juntamos, obtenemos una malla sobre la imagen. Las muestras que se obtienen son los llamados píxeles. Véase Figura 4.22(a).



**Figura 4.21:** Señal asociada a un corte transversal de la imagen.

En la cuantificación, transformamos los valores continuos de la señal en discretos, de modo que a las amplitudes de la señal se les asignan tonos de grises indicados por una colección finita del intervalo  $[0, 1]$ . Véase Figura 4.22(b).



(a) muestreo

(b) cuantificación

**Figura 4.22:** Adquisición de una imagen digital por muestreo y cuantificación.

Así, a partir de la distribución continua de tonos de gris de una imagen en blanco

y negro, los procesos de muestreo y cuantificación nos dan píxeles coloreados con una colección finita de tonos de gris [40], [62].

### 4.1.3. Modelo Discreto para Difuminación

En el deblurring, suponemos que al difuminar una imagen  $\mathcal{F}$ , obtenemos la imagen  $\mathcal{G}$ . Conocemos a  $\mathcal{G}$ , y queremos recuperar a  $\mathcal{F}$ . En vista del muestreo y cuantificación, disponemos solamente de conjunto finito de valores de la función  $g$  asociada a la imagen  $\mathcal{G}$ . Así que debemos hallar la función  $f$  asociada a la imagen  $\mathcal{F}$  a partir de los píxeles de  $\mathcal{G}$ .

A fin de obtener un arreglo  $\mathbf{f}$  con los valores de los píxeles de  $\mathcal{F}$  a partir de un arreglo  $\mathbf{g}$  con los de  $\mathcal{G}$ , lo que hacemos es discretizar nuestro modelo de difuminación dado por la convolución (4.3) en un sistema de ecuaciones lineales  $A\mathbf{f} = \mathbf{g}$ . Para conocer la estructura que tiene la matriz  $A$ , usamos las siguientes hipótesis:

- \* Los píxeles de  $\mathcal{G}$  se encuentran en puntos de  $\mathbb{Z}^2$ .
- \* La función  $g$  tiene soporte acotado en  $R_g = [1, m] \times [1, n]$ .
- \* La PSF  $k$  tiene soporte acotado en  $R_k = [-l, l] \times [-r, r]$ , donde  $2l < m$  y  $2r < n$ .

Discretizamos la convolución (4.3) con el método de colocación en los puntos de  $\mathbb{Z}^2$ . Aproximamos la doble integral por sumas de Riemann sobre cuadrados unitarios. Esto nos da la *convolución discreta*

$$\underbrace{g(i, j)}_{g_{i,j}} = \sum_{p,q \in \mathbb{Z}} \underbrace{f(p, q)}_{f_{p,q}} \underbrace{k(i-p, j-q)}_{k_{i-p,j-q}} \quad \forall i, j \in \mathbb{Z}.$$

de las sucesiones  $\{f_{p,q}\}$  y  $\{k_{p,q}\}$ .

Debido a que  $k$  y  $g$  tienen soporte acotado en  $R_k$  y  $R_g$ , respectivamente, tenemos que la convolución discreta se reduce a

$$g_{i,j} = \sum_{p=i-l}^{i+l} \sum_{q=j-r}^{j+r} f_{p,q} k_{i-p,j-q} \quad \begin{array}{l} i = 1, \dots, m, \\ j = 1, \dots, n. \end{array} \quad (4.4)$$

En las sumas anteriores el valor de cada píxel en la imagen difuminada es una suma ponderada de los valores del píxel correspondiente en la imagen original y sus vecinos.

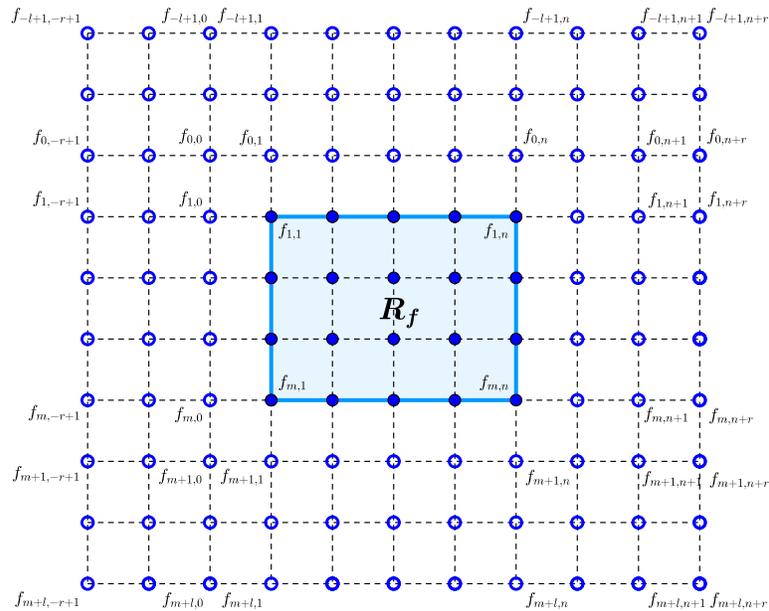
Queremos reescribir (4.4) como un sistema de ecuaciones lineales. Para ver esto, fijamos el primer subíndice y examinamos primero el caso unidimensional:

$$g_j = \sum_{q=j-r}^{j+r} k_{j-q} f_q, \quad j = 1, \dots, n. \quad (4.5)$$









**Figura 4.23:** Los  $(m + 2l)(n + 2r)$  valores de la función  $f$  sobre  $\mathbb{Z}^2$ . Marcamos los  $mn$  puntos dentro de  $R_f$  que corresponden a los píxeles de la imagen  $\mathcal{F}$ .

Tratamos la zona fuera del rectángulo que delimita la imagen  $\mathcal{F}$  de acuerdo al uso que le demos a la imagen. Por ejemplo, nos conviene tomar un fondo negro para imágenes de astros, repetimos la imagen si la captamos en una secuencia de video, o la reflejamos si notamos una simetría. Por esta razón manejamos condiciones de frontera en el rectángulo  $R_f$  donde se restringe la función  $f$ . Vamos a dar tres condiciones de frontera para los valores  $f_{p,q}$  fuera del rectángulo  $R_f$ .

- \* **Condiciones Cero.** La imagen  $\mathcal{F}$  fuera del rectángulo que la delimita tiene un fondo negro, esto quiere decir que la función  $f$  es cero fuera de  $R_f$ . Véase Figura 4.24. Por lo que

$$f_{p,q} = 0 \quad \text{si} \quad p \notin \{1, \dots, m\} \quad \text{o} \quad q \notin \{1, \dots, n\}.$$



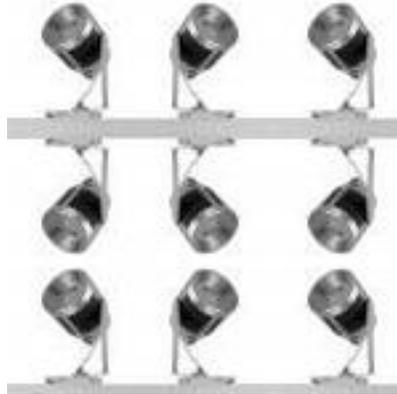
**Figura 4.24:** Imagen con condiciones de frontera cero.

\* **Condiciones Reflexivas.** La imagen  $\mathcal{F}$  se refleja en cada lado y esquina del rectángulo que la delimita. Véase Figura 4.25. En consecuencia,

$$f_{p,q} = f_{\varphi(p),\psi(q)},$$

donde

$$\varphi(p) = \begin{cases} -p + 1, & \text{si } p = -(l - 1), \dots, 0, \\ p, & \text{si } p = 1, \dots, m, \\ 2m - p + 1, & \text{si } p = m + 1, \dots, m + l, \end{cases} \quad \text{y} \quad \psi(q) = \begin{cases} -q + 1, & \text{si } q = -(r - 1), \dots, 0, \\ q, & \text{si } q = 1, \dots, n, \\ 2n - q + 1, & \text{si } q = n + 1, \dots, n + r. \end{cases}$$



**Figura 4.25:** Imagen con condiciones de frontera reflexivas.

\* **Condiciones Periódicas.** La imagen  $\mathcal{F}$  se repite fuera del rectángulo que la delimita. Véase Figura 4.26. Esto nos indica que  $f$  es periódica. Entonces

$$f_{p,q} = f_{p \bmod m, q \bmod n}, \quad \forall p, q \in \mathbb{Z}.$$



**Figura 4.26:** Imagen con condiciones de frontera periódicas.

Examinamos como repercuten las condiciones frontera en el sistema de ecuaciones lineales  $A\mathbf{f}_{\text{ext}} = \mathbf{g}$ . Vemos primero el caso 1D de la ecuación  $A_{\uparrow}\mathbf{f}_{\text{ext}\uparrow} = \mathbf{g}_{\uparrow}$ .

## Caso 1D

Sea

$$\mathbf{f}_{\dagger} = [f_1 \ \cdots \ f_m]^T.$$

Separamos la matriz  $A_{\dagger}$  de la siguiente manera:

- \* La matriz  $\underline{T}$  de tamaño  $n \times r$  con elementos

$$\underline{t}_{j,q} = \begin{cases} k_{j-q+r}, & \text{si } j \leq q, & j = 1, \dots, n \\ 0, & \text{en otro caso,} & q = 1, \dots, r, \end{cases}$$

formada por las primeras  $r$  columnas de  $A_{\dagger}$ .

- \* La matriz  $T$  de tamaño  $n \times n$  con elementos

$$t_{j,q} = \begin{cases} k_{j-q}, & \text{si } |j - q| \leq r, \\ 0, & \text{en otro caso,} \end{cases} \quad j, q = 1, \dots, r,$$

que obtenemos al quitar las primeras y últimas  $r$  columnas de  $A_{\dagger}$ .

- \* La matriz  $\overline{T}$  de tamaño  $n \times r$  con elementos

$$\overline{t}_{j,q} = \begin{cases} k_{j-q-n}, & \text{si } j - q \geq n - r, & j = 1, \dots, n \\ 0, & \text{en otro caso,} & q = 1, \dots, r. \end{cases}$$

formada por las últimas  $r$  columnas de  $A_{\dagger}$ .

Estas matrices nos permiten identificar la estructura que el sistema de ecuaciones lineales  $A_{\dagger} \mathbf{f}_{\text{ext}\dagger} = \mathbf{g}_{\dagger}$  adopta con las tres condiciones frontera [90].

- \* **Condiciones Cero.** En la ecuación  $A_{\dagger} \mathbf{f}_{\text{ext}\dagger} = \mathbf{g}_{\dagger}$ , las primeras y últimas  $r$  columnas de  $A_{\dagger}$  se multiplican por ceros. Así que podemos removerlas. De esa manera, obtenemos el sistema de ecuaciones lineales

$$T \mathbf{f}_{\dagger} = \mathbf{g}_{\dagger}. \quad (4.9)$$

La matriz  $T$  tiene diagonales constantes, pues  $t_{j,q}$  depende de la diferencia  $j - q$ . Las matrices que cumplen esto se llaman **matrices de Toeplitz**. Basta tener la primera columna y el primer renglón para generarlas.

- \* **Condiciones Periódicas.** En la ecuación  $A_{\dagger} \mathbf{f}_{\text{ext}\dagger} = \mathbf{g}_{\dagger}$ , las primeras  $r$  columnas de  $A_{\dagger}$  se multiplican por  $f_{n-r+1}, \dots, f_n$ , mientras que las últimas  $r$  columnas  $A_{\dagger}$  se multiplican por  $f_1, \dots, f_r$ , respectivamente. Por lo que sumamos las columnas de  $\underline{T}$

con las últimas  $r$  columnas de  $T$  y las columnas de  $\overline{\overline{T}}$  con las primeras  $r$  columnas de  $T$ . De ese modo, tenemos el sistema de ecuaciones lineales

$$C\mathbf{f}_{\uparrow} = \mathbf{g}_{\uparrow},$$

donde

$$C = [0_{n \times (n-r)} \quad \underline{\underline{T}}] + T + \left[ \overline{\overline{T}} \quad 0_{n \times (n-r)} \right].$$

Por ejemplo para  $n = 6$  y  $r = 2$ , tenemos que

$$C = \begin{bmatrix} k_0 & k_{-1} & k_{-2} & 0 & k_2 & k_1 \\ k_1 & k_0 & k_{-1} & k_{-2} & 0 & k_2 \\ k_2 & k_1 & k_0 & k_{-1} & k_{-2} & 0 \\ 0 & k_2 & k_1 & k_0 & k_{-1} & k_{-2} \\ k_{-2} & 0 & k_2 & k_1 & k_0 & k_{-1} \\ k_{-1} & k_{-2} & 0 & k_2 & k_1 & k_0 \end{bmatrix}.$$

La matriz  $C$  tiene una estructura especial. En su  $j$ -ésimo renglón, su primer renglón se recorre  $j - 1$  posiciones a la derecha y los  $j - 1$  elementos que sobran se agregan al inicio del renglón. Los elementos  $c_{j,q}$  cumplen que

$$c_{j,q} = c_{u,v} \quad \text{si} \quad q - j \equiv v - u \pmod{n}.$$

Las matrices que satisfacen esta relación son las *matrices circulares* [124].

\* **Condiciones Reflexivas.** En la ecuación  $A_{\uparrow}\mathbf{f}_{\text{ext}\uparrow} = \mathbf{g}_{\uparrow}$ , las primeras  $r$  columnas de  $A_{\uparrow}$  se multiplican por  $f_r, \dots, f_1$ , mientras que las últimas columnas de  $A_{\uparrow}$  se multiplican por  $f_m, \dots, f_{m-r+1}$ , respectivamente. Así que volteamos las columnas de  $\underline{\underline{T}}$  y las sumamos con las primeras  $r$  columnas de  $T$ . Análogamente, volteamos las columnas de  $\overline{\overline{T}}$  y las sumamos con las últimas  $r$  columnas de  $T$ . Para voltear las columnas usamos la matriz

$$J = \begin{bmatrix} 0 & & & 1 \\ & \ddots & & \\ & & \ddots & \\ 1 & & & 0 \end{bmatrix}_{m \times m}.$$

Así, a partir de la ecuación  $A_{\uparrow}\mathbf{f}_{\text{ext}\uparrow} = \mathbf{g}_{\uparrow}$ , tenemos al sistema de ecuaciones lineales

$$(T + H)\mathbf{f} = \mathbf{g}, \tag{4.10}$$

donde

$$H = [0_{n \times (n-r)} \quad \underline{\underline{T}}] J + \left[ \overline{\overline{T}} \quad 0_{n \times (n-r)} \right] J.$$

Por ejemplo, para  $n = 5$  y  $r = 3$ , tenemos que

$$H = \begin{bmatrix} k_1 & k_2 & k_3 & 0 & 0 \\ k_2 & k_3 & 0 & 0 & 0 \\ k_3 & 0 & 0 & 0 & k_{-3} \\ 0 & 0 & 0 & k_{-3} & k_{-2} \\ 0 & 0 & k_{-3} & k_{-2} & k_{-1} \end{bmatrix}.$$

La matriz  $H$  tiene antidiagonales constantes, ya que sus elementos  $h_{j,q}$  dependen de la suma  $j + q$ . Las matrices que cumplen esto se llaman *matrices de Hankel*. En consecuencia la matriz de coeficientes  $T + H$  es la suma de una matriz de Toeplitz con una matriz de Hankel.

Análogamente, si reemplazamos  $j, q, n, r$  por  $i, p, m, l$ , respectivamente. podemos reducir la ecuación  $\underset{\leftrightarrow}{A} \underset{\leftrightarrow}{\mathbf{f}}_{\text{ext}} = \underset{\leftrightarrow}{\mathbf{g}}$  a un sistema de ecuaciones lineales que tiene matriz de coeficientes con las mismas estructuras. A continuación, examinamos el caso bidimensional.

### Caso 2D

Ahora, incorporamos las condiciones de frontera a la ecuación  $A_{\text{ext}} \mathbf{f}_{\text{ext}} = \mathbf{g}$ . En vez del arreglo  $F_{\text{ext}}$ , buscamos la matriz

$$F = \begin{bmatrix} f_{1,1} & \cdots & f_{1,n} \\ \vdots & & \vdots \\ f_{m,1} & \cdots & f_{m,n} \end{bmatrix}$$

con los  $mn$  valores de los píxeles de la imagen restaurada. Denotamos por  $\mathbf{f}$  al vector con las columnas apiladas de  $F$ .

Reordenamos la estructura por bloques de la matriz  $A_{\text{ext}}$  de acuerdo con las condiciones de frontera. La Ecuación  $A_{\text{ext}} \mathbf{f}_{\text{ext}} = \mathbf{g}$  se reduce a un sistema de ecuaciones lineales  $A \mathbf{f} = \mathbf{g}$ , donde  $A$  es una matriz de orden  $nm$ .

Queremos ver la estructura de la matriz  $A$ . Lo que hacemos es expandir la forma que toman las matrices  $A_{\uparrow}$  y  $\underset{\leftrightarrow}{A}$  del caso 1D con cada condición de frontera, pues la estructura por bloques de  $A_{\text{ext}}$  se obtiene al fijar el primer subíndice en (4.4), mientras que la forma de sus bloques  $K^{(q)}$  aparece al fijar el segundo subíndice. Con las condiciones de frontera, la matriz  $A_{\text{ext}}$  se reordena en una matriz por bloques  $A$  de la misma forma en reordenamos que  $A_{\uparrow}$ . Asimismo, cada bloque  $K^{(q)}$  de  $A_{\text{ext}}$  se reduce a una matriz en banda  $A^{(q)}$  de tamaño  $m \times m$  de la misma manera en que reducimos  $\underset{\leftrightarrow}{A}$ .

Sabemos del caso 1D que  $A_{\uparrow}$  y  $\underset{\leftrightarrow}{A}$  se acomodan como matrices de Toeplitz si usamos condiciones cero, matrices circulares con condiciones periódicas, y en la suma de una matriz Toeplitz con una de Hankel si las condiciones son reflexivas. Luego,

- condiciones cero  $\implies$   $A$  es Toeplitz por bloques y  $A^{(q)}$  es matriz de Toeplitz,
- condiciones periódicas  $\implies$   $A$  es circular por bloques y  $A^{(q)}$  es matriz circular,
- condiciones reflexivas  $\implies$   $A$  es la suma de una matriz de Toeplitz por bloques con una matriz de Hankel por bloques y  $A^{(q)}$  es suma de una matriz de Toeplitz con una matriz de Hankel.

Así,  $A$  es una matriz en banda por bloques que tiene  $n \times n$  bloques y cada bloque es una matriz en banda de tamaño  $m \times m$ . Combinamos la estructura por bloques de  $A$  con la que subyace a su vez en sus bloques. De ese modo, podemos dar la estructura completa de  $A$  con cada condición de frontera.

\* **Condiciones Cero.**  $A$  es matriz de Toeplitz por bloques y cada bloque es matriz de Toeplitz (**BTTB**). Por ejemplo, para  $m = n = 5$  y  $l = r = 2$ , tenemos que

$$A = \begin{bmatrix} A^{(0)} & A^{(-1)} & A^{(-2)} & \mathbf{0} & \mathbf{0} \\ A^{(1)} & A^{(0)} & A^{(-1)} & A^{(-2)} & \mathbf{0} \\ A^{(2)} & A^{(1)} & A^{(0)} & A^{(-1)} & A^{(-2)} \\ \mathbf{0} & A^{(2)} & A^{(1)} & A^{(0)} & A^{(-1)} \\ \mathbf{0} & \mathbf{0} & A^{(2)} & A^{(1)} & A^{(0)} \end{bmatrix},$$

donde

$$A^{(q)} = \begin{bmatrix} k_{0,q} & k_{-1,q} & k_{-2,q} & \mathbf{0} & \mathbf{0} \\ k_{1,q} & k_{0,q} & k_{-1,q} & k_{-2,q} & \mathbf{0} \\ k_{2,q} & k_{1,q} & k_{0,q} & k_{-1,q} & k_{-2,q} \\ \mathbf{0} & k_{2,q} & k_{1,q} & k_{0,q} & k_{-1,q} \\ \mathbf{0} & \mathbf{0} & k_{2,q} & k_{1,q} & k_{0,q} \end{bmatrix}.$$

\* **Condiciones periódicas.**  $A$  es matriz circular por bloques y cada bloque es matriz circular (**BCCB**). Por ejemplo, para  $n = m = 5$ , y  $l = r = 2$ , tenemos que

$$A = \begin{bmatrix} A^{(0)} & A^{(-1)} & A^{(-2)} & A^{(2)} & A^{(1)} \\ A^{(1)} & A^{(0)} & A^{(-1)} & A^{(-2)} & A^{(2)} \\ A^{(2)} & A^{(1)} & A^{(0)} & A^{(-1)} & A^{(-2)} \\ A^{(-2)} & A^{(2)} & A^{(1)} & A^{(0)} & A^{(-1)} \\ A^{(-1)} & A^{(-2)} & A^{(2)} & A^{(1)} & A^{(0)} \end{bmatrix},$$

donde

$$A^{(q)} = \begin{bmatrix} k_{0,q} & k_{-1,q} & k_{-2,q} & k_{2,q} & k_{1,q} \\ k_{1,q} & k_{0,q} & k_{-1,q} & k_{-2,q} & k_{2,q} \\ k_{2,q} & k_{1,q} & k_{0,q} & k_{-1,q} & k_{-2,q} \\ k_{-2,q} & k_{2,q} & k_{1,q} & k_{0,q} & k_{-1,q} \\ k_{-1,q} & k_{-2,q} & k_{2,q} & k_{1,q} & k_{0,q} \end{bmatrix}.$$

\* **Condiciones reflexivas.** La matriz  $A$  es la suma de cuatro matrices de  $n \times n$  bloques:

$$\begin{aligned}
 & \begin{bmatrix} A^{(0)} & A^{(-1)} & \dots & A^{(-r)} & \mathbf{0} & \dots & \mathbf{0} \\ A^{(-1)} & A^{(0)} & \ddots & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & \mathbf{0} \\ A^{(r)} & & \ddots & \ddots & \ddots & & A^{(-r)} \\ \mathbf{0} & \ddots & & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & & \ddots & & A^{(-1)} \\ \mathbf{0} & \dots & \mathbf{0} & A^{(r)} & \dots & A^{(1)} & A^{(0)} \end{bmatrix} \\
 & \qquad \qquad \qquad A \\
 = & \begin{bmatrix} T^{(0)} & T^{(-1)} & \dots & T^{(-r)} & \mathbf{0} & \dots & \mathbf{0} \\ T^{(-1)} & T^{(0)} & \ddots & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & \mathbf{0} \\ T^{(r)} & & \ddots & \ddots & \ddots & & T^{(-r)} \\ \mathbf{0} & \ddots & & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & & \ddots & & T^{(-1)} \\ \mathbf{0} & \dots & \mathbf{0} & T^{(r)} & \dots & T^{(1)} & T^{(0)} \end{bmatrix} + \begin{bmatrix} H^{(0)} & H^{(-1)} & \dots & H^{(-r)} & \mathbf{0} & \dots & \mathbf{0} \\ H^{(-1)} & H^{(0)} & \ddots & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & \mathbf{0} \\ H^{(r)} & & \ddots & \ddots & \ddots & & H^{(-r)} \\ \mathbf{0} & \ddots & & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & & \ddots & & H^{(-1)} \\ \mathbf{0} & \dots & \mathbf{0} & H^{(r)} & \dots & H^{(1)} & H^{(0)} \end{bmatrix} \\
 & \qquad \qquad \qquad T \qquad \qquad \qquad T' \\
 + & \begin{bmatrix} T^{(1)} & T^{(2)} & \dots & T^{(r)} & \mathbf{0} & \dots & \dots & \dots & \mathbf{0} \\ T^{(2)} & & \ddots & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & & \vdots \\ T^{(r)} & & \ddots & \ddots & \ddots & & & & \vdots \\ \mathbf{0} & & & \ddots & \ddots & & & & \mathbf{0} \\ \vdots & & & \ddots & \ddots & & & & \vdots \\ \vdots & & & \ddots & \ddots & & & & T^{(-r)} \\ \vdots & & & \ddots & \ddots & & & & \vdots \\ \vdots & & & \ddots & \ddots & & & & T^{(-2)} \\ \mathbf{0} & \dots & \dots & \dots & \mathbf{0} & T^{(-r)} & \dots & T^{(-2)} & T^{(-1)} \end{bmatrix} + \begin{bmatrix} H^{(1)} & H^{(2)} & \dots & H^{(r)} & \mathbf{0} & \dots & \dots & \dots & \mathbf{0} \\ H^{(2)} & & \ddots & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & & \vdots \\ H^{(r)} & & \ddots & \ddots & \ddots & & & & \vdots \\ \mathbf{0} & & & \ddots & \ddots & & & & \mathbf{0} \\ \vdots & & & \ddots & \ddots & & & & \vdots \\ \vdots & & & \ddots & \ddots & & & & H^{(-r)} \\ \vdots & & & \ddots & \ddots & & & & \vdots \\ \vdots & & & \ddots & \ddots & & & & H^{(-2)} \\ \mathbf{0} & \dots & \dots & \dots & \mathbf{0} & H^{(-r)} & \dots & H^{(-2)} & H^{(-1)} \end{bmatrix} \\
 & \qquad \qquad \qquad H' \qquad \qquad \qquad H
 \end{aligned}$$

donde  $T^{(q)}$  y  $H^{(q)}$  son las matrices de Toeplitz y Hankel de orden  $m$  dadas por

$$T^{(q)} = \begin{bmatrix} k_{0,q} & k_{-1,q} & \dots & k_{-l,q} & \mathbf{0} & \dots & \mathbf{0} \\ k_{-1,q} & k_{0,q} & \ddots & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & \mathbf{0} \\ k_{l,q} & & \ddots & \ddots & \ddots & & k_{-l,q} \\ \mathbf{0} & \ddots & & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & & \ddots & & k_{-1,q} \\ \mathbf{0} & \dots & \mathbf{0} & k_{l,q} & \dots & k_1 & k_0 \end{bmatrix} \quad \text{y} \quad H^{(q)} = \begin{bmatrix} k_{1,q} & k_{2,q} & \dots & k_{l,q} & \mathbf{0} & \dots & \dots & \dots & \mathbf{0} \\ k_{2,q} & & \ddots & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & & \vdots \\ k_{l,q} & & \ddots & \ddots & \ddots & & & & \vdots \\ \mathbf{0} & & & \ddots & \ddots & & & & \mathbf{0} \\ \vdots & & & \ddots & \ddots & & & & k_{-l,q} \\ \vdots & & & \ddots & \ddots & & & & \vdots \\ \vdots & & & \ddots & \ddots & & & & k_{-2,q} \\ \mathbf{0} & \dots & \dots & \dots & \mathbf{0} & k_{-l,q} & \dots & k_{-2,q} & k_{-1,q} \end{bmatrix}$$

Las matrices por bloques tienen la siguiente estructura:

$T$  es matriz Toeplitz por bloques, cada bloque es matriz de Toeplitz (**BTTB**),  
 $T'$  es matriz Toeplitz por bloques, cada bloque es matriz de Hankel (**BTHB**),  
 $H$  es matriz Hankel por bloques, cada bloque es matriz de Hankel (**BHHB**),  
 $H'$  es matriz Hankel por bloques, cada bloque es matriz de Toeplitz (**BHTB**).

Así que la matriz  $A$  se separa como

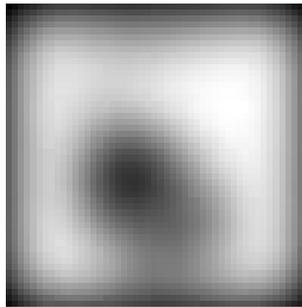
$$A = \text{BTTB} + \text{BTHB} + \text{BHHB} + \text{BHTB}.$$

**Ejemplo 4.3.** Difuminamos la imagen de  $50 \times 50$  píxeles de la Figura 4.27 mediante la PSF gaussiana. Delimitamos esta PSF al cuadrado  $[-20, 20] \times [-20, 20]$ .

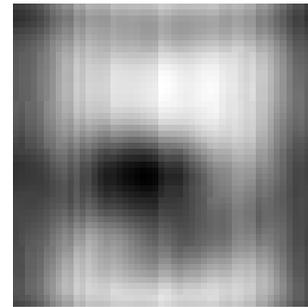
Usamos desviación estándar 5. La gráfica de la PSF es la campana de la Figura 4.30(a). Supongamos condiciones de frontera cero. Entonces obtenemos una matriz  $A$  de tamaño  $2500 \times 2500$  que es BTTB. En la Figura 4.30(b) mostramos una gráfica de sus elementos. Notamos que la intensidad del color es más fuerte en una banda alrededor de la diagonal principal.



**Figura 4.27:** Imagen de reflector

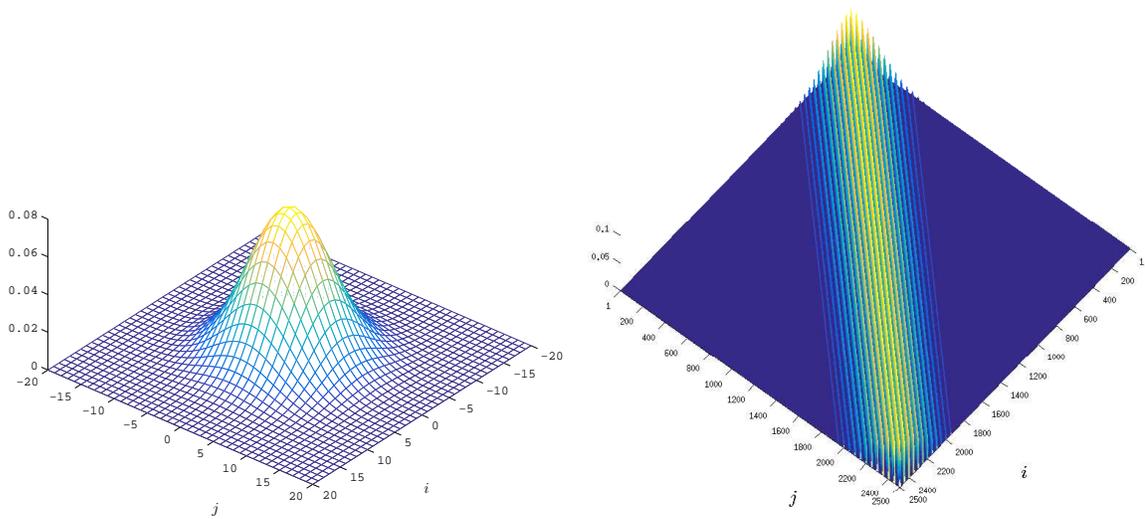


**Figura 4.28:** Imagen difuminada con PSF gaussiana de desviación estándar 5 con condiciones de frontera cero



**Figura 4.29:** Imagen difuminada con PSF gaussiana de desviación estándar 2 con condiciones de frontera periódicas

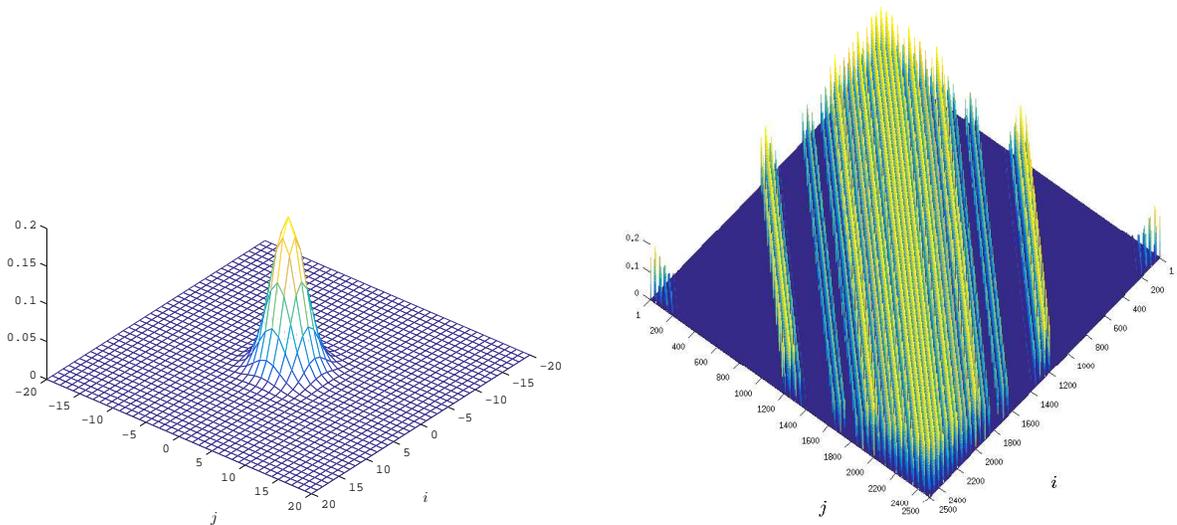
Ahora, usamos desviación estándar 2. En la Figura 4.31(a) notamos que la gráfica de la gaussiana se comprime. Supongamos condiciones de frontera periódicas. Entonces obtenemos una matriz  $A$  de tamaño  $2500 \times 2500$  que es BCCB. En la Figura 4.31(b) mostramos una gráfica de sus elementos. Notamos que la intensidad del color se distribuye en bandas alrededor de la diagonal principal así como en las esquinas de la matriz.



(a) PSF gaussiana de desviación estándar 5 sobre cuadrado  $[-20, 20] \times [-20, 20]$

(b) elementos de la matriz BTTB

**Figura 4.30:** PSF gaussiana + condiciones cero nos da una matriz BTTB.



(a) PSF gaussiana de desviación estándar 2 sobre cuadrado  $[-20, 20] \times [-20, 20]$

(b) elementos de matriz BCCB

**Figura 4.31:** PSF gaussiana + condiciones periódicas nos da una matriz BCCB.

#### 4.1.5. Reducción del Problema

Como las imágenes digitales pueden tener tamaños en píxeles de  $256 \times 256$ ,  $1024 \times 768$ ,  $2048 \times 1200$ , entre otros, entonces el problema discreto de deblurring  $A\mathbf{f} = \mathbf{g}$  involucra resolver un sistema de ecuaciones lineales de gran escala, su matriz de coeficientes

es cuadrada de orden  $256^2 = 62500$ ,  $(1024)(768) = 786432$  y  $(2048)(1200) = 2457600$ , respectivamente. Deseamos reducir las dimensiones de nuestro problema. La idea es que si podemos separar la PSF, separemos  $A$  en matrices de menor tamaño.

La PSF  $k : \mathbb{R}^2 \rightarrow \mathbb{R}$  es *separable* si existen funciones  $r, c : \mathbb{R} \rightarrow \mathbb{R}$  tales que

$$k(x - s, y - t) = r(x - y)c(s - t) \quad \forall x, y, s, t \in \mathbb{R}.$$

Vamos a trabajar con PSFs espacialmente invariantes y separables. Por ejemplo, la PSF gaussiana.

Entonces las evaluaciones de la PSF  $k$  en la malla uniforme sobre  $\mathbb{Z}^2$  se factorizan como

$$k_{i-p, j-q} = c_{i-p} r_{j-q}, \quad i, j, p, q \in \mathbb{Z}.$$

Sea  $C$  la matriz  $m \times m$  con elementos

$$c_{i,p} = c_{i-p}, \quad i, p = 1, \dots, m.$$

y sea  $R$  la matriz  $n \times n$  con elementos

$$r_{j,q} = r_{j-q}, \quad j, q = 1, \dots, n.$$

Entonces cada bloque  $A^{(q)}$  de  $A$  de tamaño  $m \times m$  distinto de  $\mathbf{0}_{m \times m}$  se puede escribir como

$$A^{(j-q)} = r_{j,q} C, \quad j, q = 1, \dots, n.$$

Identificamos una estructura conocida en la matriz  $A$ , llamada producto de Kronecker.

### Producto de Kronecker

El *producto de Kronecker* de  $X \in \mathbb{R}^{n \times p}$  con  $Y \in \mathbb{R}^{m \times q}$  es la matriz por bloques

$$X \otimes Y = \begin{bmatrix} x_{1,1}Y & \cdots & r_{1,p}Y \\ \vdots & & \vdots \\ x_{n,1}Y & \cdots & r_{n,p}Y \end{bmatrix}_{nm \times pq}.$$

Por ejemplo,

$$\begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 3 & 4 & 0 & 0 \\ 5 & 6 & 0 & 0 \\ 2 & 4 & -1 & -2 \\ 6 & 8 & -3 & -4 \\ 10 & 12 & -5 & -6 \end{bmatrix}.$$

Mencionamos algunas propiedades del producto de Kronecker:

\* **Producto mezclado.** Para todas las matrices  $X, Y, Z, W$  de productos compatibles  $XZ$  y  $YW$ , tenemos

$$(X \otimes Y)(Z \otimes W) = XZ \otimes YW.$$

\* **Distributividad de transposición.** Para cualesquiera matrices  $X$  e  $Y$ , tenemos

$$(X \otimes Y)^T = X^T \otimes Y^T.$$

\* **Distributividad de inversa.** Para todas las matrices invertibles  $X$  e  $Y$ , tenemos

$$(X \otimes Y)^{-1} = X^{-1} \otimes Y^{-1}.$$

\* **Vectorización.** Para cualesquiera matrices  $X, Y, Z$  tales que el producto  $XYZ$  es compatible, tenemos

$$\text{vec}(XYZ) = (Z^T \otimes X)\text{vec}(Y),$$

donde  $\text{vec}(X)$  denota al vector con las columnas apiladas de la matriz  $X$ .

#### Observaciones 4.7:

👉 El producto de Kronecker no es conmutativo.

👉  $X \otimes Y = XY^T$  cuando  $X$  e  $Y$  son vectores.

👉 La propiedad de vectorización nos permite pasar una multiplicación matriz-vector, donde la matriz es de gran tamaño, a una multiplicación de matrices de menor dimensión. De ese modo, reducimos las dimensiones de nuestro problema. Véase [75] para más propiedades del producto del Kronecker.

#### Condiciones de Frontera con PSF Separable

Regresemos al sistema de ecuaciones lineales  $A\mathbf{f} = \mathbf{g}$  con PSF separable. Tratamos las tres condiciones de frontera por separado:

\* Si usamos condiciones cero, entonces la matriz  $A$  que es BTTB se convierte en el producto de Kronecker de las matrices de Toeplitz  $R$  y  $C$ :

$$A = R \otimes C.$$

\* Para condiciones de frontera periódicas, la matriz  $A$  que es BCCB se escribe como el producto de Kronecker de las matrices de circulares  $R$  y  $C$ .

\* Para condiciones de frontera reflexivas,

$$A = \begin{array}{cccc} T & + & T' & + & H & + & H' \\ \text{BTTB} & & \text{BTHB} & & \text{BHHB} & & \text{BHTB} \end{array}$$

Podemos separar cada una de las cuatro matrices por bloques como

$$\begin{aligned} T &= R_T \otimes C_T, & H &= R_H \otimes C_H, \\ T' &= R_T \otimes C_H, & H' &= R_H \otimes C_T. \end{aligned}$$

donde

$R_T$  es una matriz de Toeplitz de orden  $m$ ,

$R_H$  es una matriz de Hankel de orden  $m$ ,

$C_T$  es una matriz de Toeplitz de orden  $n$ ,

$C_H$  es una matriz de Hankel de orden  $n$ ,

De ese modo, la propiedad del producto mezclado implica que  $A$  es el producto de Kronecker de matrices  $R$  y  $C$  que son la suma de una matriz de Toeplitz con una de Hankel, esto es,

$$R = R_T + R_H \quad \text{y} \quad C = C_T + C_H.$$

De aquí que

$$\text{BTTB} + \text{BTHB} + \text{BHTB} + \text{BHBB} = (\text{Toeplitz} + \text{Hankel}) \otimes (\text{Toeplitz} + \text{Hankel}).$$

En cualquier caso, la ecuación  $A\mathbf{f} = \mathbf{g}$  se reescribe como

$$(R \otimes C)\text{vec}(F) = \text{vec}(G). \quad (4.11)$$

La solución de esta ecuación nos da la matriz  $F$  de tamaño  $m \times n$  asociada a la imagen restaurada. El rango de  $R \otimes C$  es

$$r = \text{rango}(R)\text{rango}(C).$$

Por lo que distinguimos cuando los factores  $R$  y  $C$  son de rango completo o deficiente.

Si  $R$  y  $C$  son invertibles, la distributividad de la inversa bajo el producto de Kronecker implica que

$$\text{vec}(F) = (R^{-1} \otimes C^{-1})\text{vec}(G).$$

Luego, por la propiedad de vectorización del producto de Kronecker, tenemos que

$$F = C^{-1}GR^{-T}.$$

De este modo podemos restaurar la imagen difuminada.

#### Observaciones 4.8:

 Cuando la PSF es espacialmente invariante, pero no es separable, entonces  $A$  puede aproximarse por una suma de productos de Kronecker. Al respecto, Kamm propone en [66] que bajo condiciones de frontera cero, resolvamos

$$\text{mín} \|(R \otimes C) - A\|_F$$

sobre todos los productos de Kronecker  $R \otimes C$  de rango uno. Para obtener nuestra aproximación de rango uno, usamos la matriz

$$K = \begin{bmatrix} k_{-l,-r} & \cdots & k_{-l,r} \\ \vdots & & \vdots \\ k_{l,-r} & \cdots & k_{l,r} \end{bmatrix}$$

que tiene todos los valores de la PSF. Condensamos la información de esta matriz en los vectores singulares  $\mathbf{u}$  y  $\mathbf{v}$  asociados al valor singular más grande  $\sigma$  de  $K$ . Formamos la matrices de Toeplitz  $R$  y  $C$  que tengan por primer renglón

$$\left[ \sqrt{\sigma}u_{l+1} \quad \cdots \quad \sqrt{\sigma}u_1 \quad \mathbf{0}_{1 \times (m-l-1)} \right] \quad \text{y} \quad \left[ \sqrt{\sigma}v_{r+1} \quad \cdots \quad \sqrt{\sigma}v_1 \quad \mathbf{0}_{1 \times (n-r-1)} \right]$$

y por primera columna

$$\begin{bmatrix} \sqrt{\sigma}u_{l+1} \\ \vdots \\ \sqrt{\sigma}u_{2l+1} \\ \mathbf{0}_{(m-l-1) \times 1} \end{bmatrix} \quad \text{y} \quad \begin{bmatrix} \sqrt{\sigma}v_{r+1} \\ \vdots \\ \sqrt{\sigma}v_{2r+1} \\ \mathbf{0}_{(n-r-1) \times 1} \end{bmatrix},$$

respectivamente.

### Producto de Kronecker y SVD

Cuando  $R$  y  $C$  son de rango deficiente, resolvemos el problema de cuadrados mínimos

$$\min_{\mathbf{x} \in \mathbb{R}^{mn}} \|(R \otimes C)\mathbf{x} - \text{vec}(G)\|_2^2 \quad (4.12)$$

La solución de cuadrados mínimos de norma mínima nos da el arreglo  $\mathbf{f}_{\text{LS}}$  que reacomodamos como la matriz  $F_{\text{LS}}$  de tamaño  $m \times n$ . Sabemos como hallar esta solución a partir de la SVD

$$(R \otimes C) = U\Sigma V^T.$$

El inconveniente es que calcular la SVD para una matriz de gran tamaño es costoso. Nos preguntamos si es posible hallar una SVD del producto de Kronecker a partir de las SVDs de sus factores.

**Teorema 4.1.** Sean  $R \in \mathbb{R}^{n \times n}$  y  $C \in \mathbb{R}^{m \times m}$  con descomposiciones en valores singulares

$$R = U_R \Sigma_R V_R^T \quad \text{y} \quad C = U_C \Sigma_C V_C^T.$$

Entonces  $U_R \otimes U_C$  y  $V_R \otimes V_C$  son matrices ortogonales  $mn \times mn$  tales que

$$R \otimes C = (U_R \otimes U_C)(\Sigma_R \otimes \Sigma_C)(V_R \otimes V_C)^T.$$

*Demostración.* Primero probamos que  $U_R \otimes U_C$  es una matriz ortogonal. Sabemos de la SVD que  $U_R$  y  $U_C$  son matrices ortogonales de tamaños  $n \times n$  y  $m \times m$ , respectivamente. Por definición,

$$U_R \otimes U_C = \begin{bmatrix} u_{1,1}^{(R)} U_C & \cdots & u_{1,n}^{(R)} U_C \\ \vdots & & \vdots \\ u_{n,1}^{(R)} U_C & \cdots & u_{n,n}^{(R)} U_C \end{bmatrix},$$

mientras que

$$(U_R \otimes U_C)^T = U_R^T \otimes U_C^T = \begin{bmatrix} u_{1,1}^{(R)} U_C^T & \cdots & u_{n,1}^{(R)} U_C^T \\ \vdots & & \vdots \\ u_{1,n}^{(R)} U_C^T & \cdots & u_{n,n}^{(R)} U_C^T \end{bmatrix}.$$

Luego  $Q := (U_R \otimes U_C)^T (U_R \otimes U_C)$  es la matriz por bloques

$$Q = \begin{bmatrix} Q_{1,1} & \cdots & Q_{1,n} \\ \vdots & & \vdots \\ Q_{n,1} & \vdots & Q_{n,n} \end{bmatrix},$$

donde

$$Q_{i,j} = \begin{bmatrix} u_{1,i}^{(R)} U_C^T & \cdots & u_{n,i}^{(R)} U_C^T \end{bmatrix} \begin{bmatrix} u_{1,j}^{(R)} U_C \\ \vdots \\ u_{n,j}^{(R)} U_C \end{bmatrix}.$$

Notamos que

$$Q_{i,j} = (\text{columna } i \text{ de } U_R)^T (\text{columna } j \text{ de } U_R) U_C^T U_C.$$

Dado que  $U_C$  es ortogonal, tenemos que

$$Q_{ij} = (\text{columna } i \text{ de } U_R)^T (\text{columna } j \text{ de } U_R) I_n.$$

Luego, como  $U_R$  es ortogonal, se sigue que  $Q_{ij} = \delta_{ij} I_n$ . De aquí,  $Q = I_{n^2}$ . Esto muestra que  $U_a \otimes U_b$  es una matriz ortogonal. Análogamente, podemos verificar que  $V_a \otimes V_b$  es una matriz ortogonal.

Ahora, probemos la Factorización

$$R \otimes C = (U_R \otimes U_C)(\Sigma_R \otimes \Sigma_C)(V_R \otimes V_C)^T.$$

Sean

$$W_R = U_R \Sigma_R \quad \text{y} \quad W_C = U_C \Sigma_C.$$

Entonces las descomposiciones en valores singulares de  $X$  e  $Y$  implican que

$$R \otimes C = (W_R V_R^T) \otimes (W_C V_C^T).$$

Por la propiedad del producto mezclado tenemos que

$$R \otimes C = (W_R \otimes W_C)(V_R^T \otimes V_C^T).$$

Dado que la transposición se distribuye sobre el producto de Kronecker, se sigue que

$$R \otimes C = (W_R \otimes W_C)(V_R \otimes V_C)^T. \quad (4.13)$$

La propiedad del producto mezclado implica que

$$W_R \otimes W_C = (U_R \otimes U_C)(\Sigma_R \otimes \Sigma_C). \quad (4.14)$$

De aquí obtenemos el resultado deseado al sustituir (4.14) en la Factorización (4.13). ♣

**Observaciones 4.9:**

👉 El Teorema 4.1 es válido aún cuando las matrices  $R$  y  $C$  no son cuadradas.

👉 Para que la Factorización

$$R \otimes C = (U_R \otimes U_C)(\Sigma_R \otimes \Sigma_C)(V_R \otimes V_C)^T.$$

sea una SVD de  $R \otimes C$ , reordenamos  $\Sigma_R \otimes \Sigma_C$  con una matriz de permutación  $P$  de tamaño  $nm \times nm$  tal que los elementos sobre la diagonal principal de

$$\Sigma := P(\Sigma_R \otimes \Sigma_C)P^T$$

estén en orden descendente y definimos

$$U := (U_R \otimes U_C)P^T \quad \text{y} \quad V := (V_R \otimes V_C)P^T.$$

De ese modo,  $R \otimes C = U\Sigma V^T$  es SVD del producto de Kronecker. Los valores singulares  $\sigma_1, \dots, \sigma_{mn}$  de  $R \otimes C$  son productos de los valores singulares de  $R$  y  $C$ .

A consecuencia del Teorema 4.1, podemos usar la SVD de  $R$  y  $C$  para obtener  $F_{LS}$ :

1. Formamos el vector

$$\mathbf{h} = P\text{vec}(U_C^T G U_R).$$

con los coeficientes de  $\text{vec}(G)$  en la base de vectores singulares de derecha de  $R \otimes C$ .

2. Multiplicamos las primeras  $r$  componentes de  $\mathbf{h}$  por los recíprocos de los  $r$  valores singulares positivos de  $R \otimes C$  y reemplazamos las últimas  $mn - r$  componentes de  $\mathbf{h}$  por ceros. Obtenemos

$$\mathbf{w} = \text{diag} \left( \frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_r}, \overbrace{0, \dots, 0}^{mn-r} \right) \mathbf{h}.$$

3.  $\text{vec}(F_{LS})$  es la expansión de  $\mathbf{w}$  en la base de vectores singulares de derecha de  $R \otimes C$ , esto es,

$$F_{LS} = V_C W V_R^T,$$

donde  $W$  la matriz de tamaño  $m \times n$  tal que  $\text{vec}(W) = P^T \mathbf{w}$ .

**Observaciones 4.10:**

👉 En el caso de condiciones de frontera periódicas, las matrices  $R$  y  $C$  son circulares. Cuando el orden de estas matrices es par, podemos usar la transformada Discreta de Fourier (DFT) de las primeras columnas de  $R$  y  $C$  para hallar sus SVDs. Sea  $\zeta$  la la

primera columna de  $C$ . Entonces su DFT es  $\hat{\zeta} = F_{DT}\zeta$ , donde  $F_{DT}$  es la matriz de orden  $m$  con elementos

$$f_{p,q} = \exp(-2pq\pi i/m) \quad p, q = 1, \dots, m.$$

Puesto que las columnas de  $F_{DT}$  son ortogonales en  $\mathbb{C}^m$  y  $\overline{F_{DT}^T}F_{DT} = mI$ , podemos verificar [124] que

$$C = \frac{1}{\sqrt{m}}\overline{F_{DT}^T} \cdot \text{diag}(\hat{\zeta}) \cdot \frac{1}{\sqrt{m}}F_{DT}.$$

De aquí, la DFT  $\hat{\zeta}$  nos da el vector con los valores propios de  $C$ . Los valores y vectores propios son complejos porque los elementos de  $F_{DT}$  son complejos. Sea  $\theta_k$  el ángulo en radianes que forman las partes real e imaginaria de  $\hat{\zeta}_k$ . Formamos la matriz  $U$  de tamaño  $m \times m$  con elementos

$$u_{p,q} = \begin{cases} \frac{\text{sign}(\hat{\zeta}_1)}{\sqrt{m}}, & q = 1, \\ \sqrt{\frac{2}{m}} \cos\left(\frac{2\pi(p-1)(q-1)}{m} + \theta_q\right), & q \in \{2, \dots, \frac{m}{2}\}, \\ \frac{\text{sign}(\hat{\zeta}_{m/2+1})}{\sqrt{m}}(-1)^{p-1}, & q = \frac{m}{2} + 1, \\ \sqrt{\frac{2}{m}} \cos\left(\frac{2\pi(p-1)(q-1)}{m} + \theta_q\right), & q \in \{\frac{m}{2} + 2, \dots, m\}, \end{cases}$$

De la misma manera, formamos una matriz  $V$  del mismo tamaño, omitiendo los ángulos  $\theta_q$  y los signos de  $\hat{\zeta}_1$  y  $\hat{\zeta}_{m/2+1}$ . Así, obtenemos matrices ortogonales  $U$  y  $V$  tales que

$$C = U \text{diag}(|\hat{\zeta}_1|, \dots, |\hat{\zeta}_m|) V^T.$$

Basta reordenar los valores absolutos de los elementos de  $\hat{\zeta}$  en orden descendente y multiplicar  $U$  y  $V$  por la matriz de permutación correspondiente para obtener una SVD de  $C$ . Análogamente, obtenemos una SVD de la matriz circular  $R$  [100].

 Una vez que tengamos las muestras de la imagen ideal, interpolamos para reconstruirla. Esto es posible si la transformada de Fourier  $\hat{f}$  de  $f$  tiene soporte acotado:

$$\hat{f}(\xi_1, \xi_2) = 0 \quad \text{para} \quad |\xi_1| > \xi_{x0}, \quad |\xi_2| > \xi_{y0}$$

y se ha realizado un muestreo uniforme de  $f$  sobre una malla rectangular con espaciamiento  $\Delta x, \Delta y$  con una tasa de muestreo mayor que la tasa de Nyquist, esto es,

$$\frac{1}{\Delta x} = \xi_{xs} > 2\xi_{x0}, \quad \frac{1}{\Delta y} = \xi_{ys} > 2\xi_{y0}.$$

De ser así, el **Teorema del muestreo** [62] nos dice que podemos interpolar  $f$  en una base de funciones Sinc normalizadas a partir de los píxeles de  $\mathcal{F}$  como

$$f(x, y) = \sum_{p, q \in \mathbb{Z}} f(p\Delta x, q\Delta y) \text{Sinc}(x\xi_{xs} - p) \text{Sinc}(y\xi_{ys} - q) \quad \forall x, y \in \mathbb{R}$$

De aquí que estemos interesados en conocer solamente las evaluaciones de  $f$  en la malla uniforme.

### 4.1.6. Regularización en el Problema de Deblurring

En nuestro modelo de difuminación con PSF  $k((x, y), (s, t))$  espacialmente invariante y cuadrado integrable, tenemos que el operador de difuminación  $\mathcal{B} : L^2(\mathbb{R}^2) \rightarrow L^2(\mathbb{R}^2)$  está dado por

$$\mathcal{B}[f](x, y) = \int \int_{\mathbb{R}^2} k(x - s, y - t) f(s, t) ds dt$$

De acuerdo a lo visto en §1.3, el que la PSF sea cuadrado integrable da lugar a que  $\mathcal{B}$  sea compacto. Así que resolver la ecuación integral (4.3) es un problema mal planteado en el sentido de Hadamard. De aquí que el problema de deblurring antes formulado está mal planteado.

Cuando discretizamos la ecuación (4.3) e incorporamos las condiciones de frontera, obtenemos un sistema de ecuaciones lineales  $A\mathbf{f} = \mathbf{g}$ , donde  $G$  y  $F$  son matrices  $m \times n$  con los valores de los píxeles de la imagen difuminada y la restaurada, respectivamente, y  $A$  es una matriz de orden  $mn$  que tiene una estructura en banda por bloques.

#### Regularización de Imágenes Difuminadas sin Ruido

Bajo la hipótesis de que la PSF es espacialmente invariante, separable, y de soporte acotado, discretizamos la convolución como el sistema de ecuaciones lineales

$$\text{vec}(G) = (R \otimes C) \text{vec}(F), \quad (4.15)$$

donde  $R$  y  $C$  son matrices de orden  $m$  y  $n$  con estructura determinada por las condiciones de frontera. Cuando los valores singulares de  $R \otimes C$  decaen a cero sin saltos, tenemos un problema discreto mal planteado.

**Ejemplo 4.4.** Considere la imagen difuminada de  $256 \times 256$  píxeles de la Figura 4.32(a). Nuestro modelo discreto de difuminación está dado por la ecuación  $A\mathbf{f} = \mathbf{g}$ , donde  $\mathbf{f}$  y  $\mathbf{g}$  son vectores de  $256^2 (= 62500)$  componentes con los valores de los píxeles de la imagen restaurada  $\mathcal{F}$  y la imagen difuminada  $\mathcal{G}$ , respectivamente, y  $A$  es una matriz de orden  $256^2$  que tiene una estructura en banda por bloques.

Sabemos que la imagen se difumina mediante una PSF gaussiana de desviación estándar 5 con soporte en el rectángulo  $[-127, 127] \times [-127, 127]$  y que en la frontera se usan condiciones cero. Dado que la PSF gaussiana es separable, tenemos que la matriz  $A$  es el producto de Kronecker  $A = R \otimes C$ , donde  $R$  y  $C$  son matrices de orden 256. Además, las condiciones cero implican que  $R$  y  $C$  son matrices en banda de Toeplitz. Para restaurar la imagen usamos la solución de cuadrados mínimos de la Ecuación (4.15) reacomodada en el arreglo  $F_{LS}$ .

La imagen restaurada que mostamos en la 4.32(b) está dominada por el ruido introducido por errores de redondeo. Si examinamos el número de condición de  $R$  y  $C$ , encontramos que

$$\kappa_2(R) = \kappa_2(C) = 4.6521 \times 10^{18}.$$

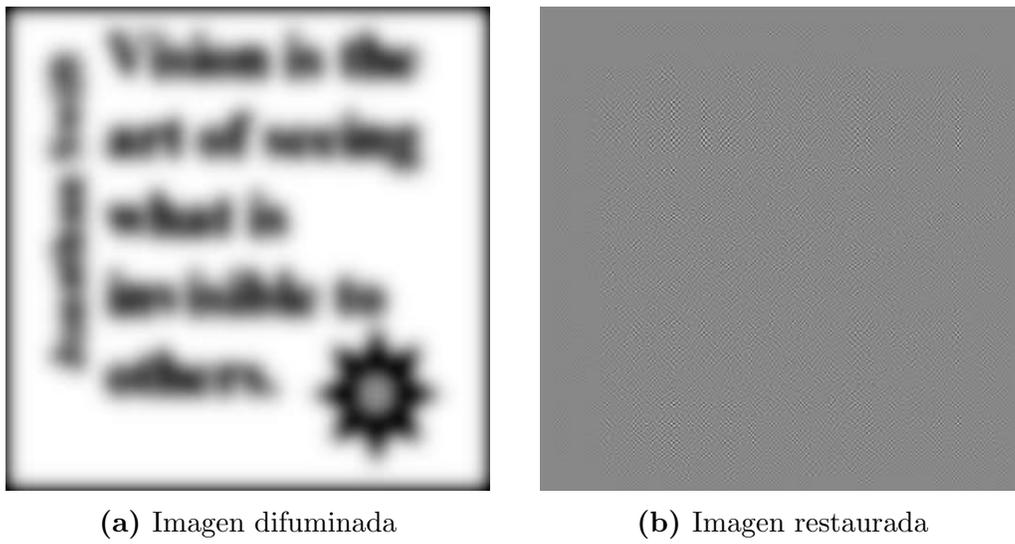


Figura 4.32: Restauración de imagen por solución de cuadrados mínimos

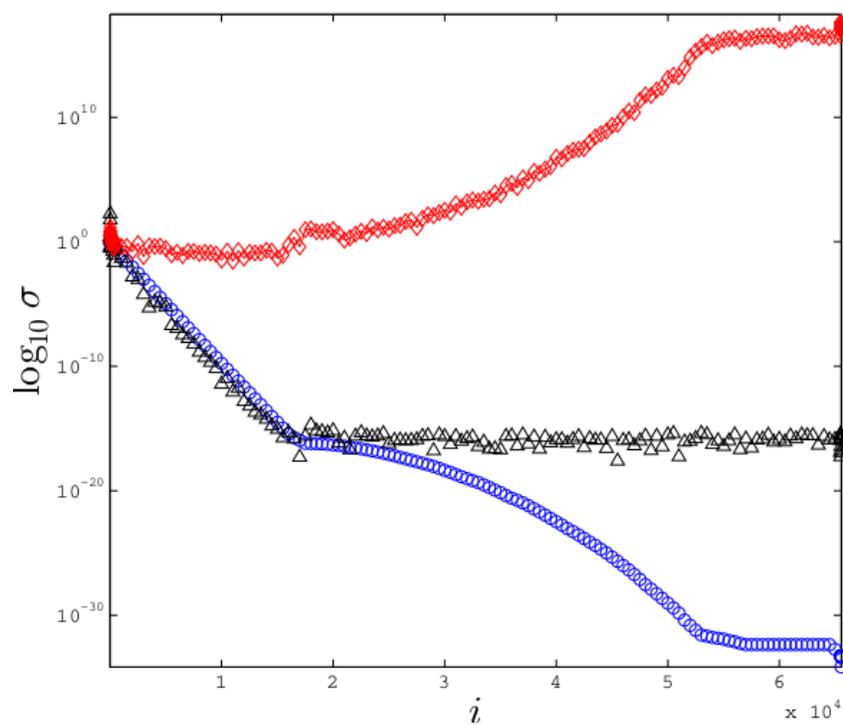


Figura 4.33: Gráfica de Picard del problema discreto mal planteado  $A\mathbf{f} = \mathbf{g}$ , valores singulares  $\sigma_i$  de  $A$  indicados por  $\circ$ , coeficientes  $|\mathbf{u}_i \mathbf{g}|$  de  $\mathbf{g}$  indicados por  $\triangle$ , coeficientes  $|\mathbf{u}_i \mathbf{g}|/\sigma_i$  de  $\mathbf{f}_{LS}$  indicados por  $\diamond$ .

Lo que pasa es que las matrices  $R$  y  $C$  son numéricamente singulares en la PC de 64 bits que usamos, más aún,  $A$  tiene rango deficiente, pues

$$\text{rank}(A) = \text{rank}(R \otimes C) = \text{rank}(R)\text{rank}(C) = 17956;$$

mientras que una matriz de rango completo de orden  $256^2$  debe tener rango 62500.

En la Figura 4.33 mostramos la gráfica de Picard del problema discreto mal planteado. Observamos que los valores singulares  $\sigma_i$  de  $A$  decaen rápidamente a cero sin saltos en escala logarítmica, mientras que los primeros 16500 coeficientes  $|\mathbf{u}^T \mathbf{g}|$  de  $\mathbf{g}$  en la base de vectores singulares de derecha decaen más rápido que  $\sigma_i$  y después se estabilizan. Luego, los coeficientes  $|\mathbf{u}_i^T \mathbf{g}|/\sigma_i$  de  $\mathbf{f}_{LS}$  en la misma base de vectores singulares crecen en promedio a partir del subíndice 16500.

Para obtener una imagen restaurada que no este dominada por el ruido, ocupamos métodos de regularización en el problema de deblurring. Vamos a regularizar mediante factores filtro  $\varphi_i$ . Sabemos que la solución regularizada es

$$\mathbf{f}_{\text{reg}} = V \text{diag} \left( \frac{\varphi_1}{\sigma_1}, \dots, \frac{\varphi_r}{\sigma_r}, \varphi_{r+1} \dots, \varphi_{mn} \right) U^T \text{vec}(G),$$

donde  $r = \text{rank}(A)$ . Aclaremos que  $\mathbf{f}_{\text{reg}}$  es el vector con las columnas apiladas de la matriz  $F_{\text{reg}}$  que tiene los valores de los píxeles de la imagen restaurada. Con las SVDs

$$R = U_R \Sigma_R V_R^T \quad \text{y} \quad C = U_C \Sigma_C V_C^T$$

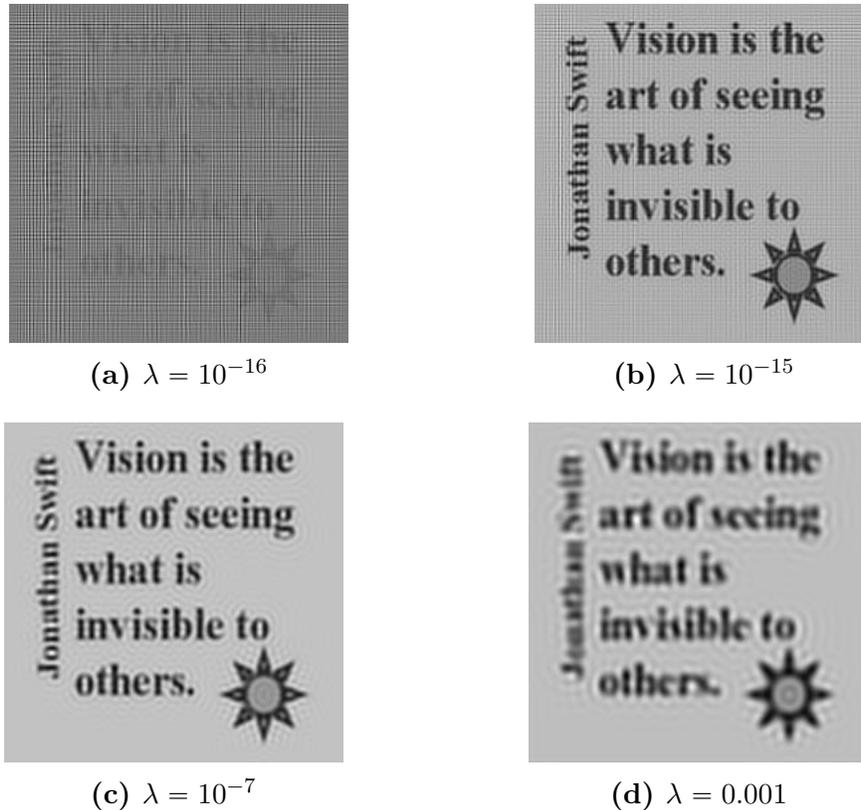
y el Teorema 4.1 formamos la solución regularizada de la misma manera que la solución de cuadrados mínimos  $\mathbf{f}_{LS}$ . La diferencia es que usamos  $\varphi_i/\sigma_i$  en lugar de  $1/\sigma_i$ :

$$\begin{aligned} \mathbf{h} &= P \text{vec}(U_C G U_R), \\ \mathbf{w} &= \text{diag} \left( \frac{\varphi_1}{\sigma_1}, \dots, \frac{\varphi_r}{\sigma_r}, \varphi_{r+1} \dots, \varphi_{mn} \right) \mathbf{h}, \\ W &\in \mathbb{R}^{m \times n} \quad \text{dada por } \text{vec}(W) = P^T \mathbf{w}, \\ \mathbf{f}_{\text{reg}} &= \text{vec}(V_C W V_R^T). \end{aligned}$$

En particular, si tomamos  $\varphi_i = \sigma_i^2/(\sigma_i^2 + \lambda^2)$ , usamos regularización de Tikhonov. En este caso, el parámetro de regularización es  $\lambda > 0$  y la matriz de tamaño  $m \times n$  con la solución regularizadora se denota por  $F_\lambda$ .

**Ejemplo 4.5.** En el Ejemplo 4.4 tratamos de restaurar una imagen difuminada por PSF gaussiana de desviación estándar 5, media cero y soporte en  $[-127, 127] \times [-127, 127]$ . El problema es resolver el sistema de ecuaciones lineales  $(R \otimes C)\mathbf{f} = \mathbf{g}$ , donde  $R$  y  $C$  son matrices en banda de Toeplitz de orden 256. Vimos que la solución de cuadrados mínimos de norma mínima  $F_{LS}$  nos da una imagen con más degradación, pues el problema está mal condicionado.

Mediante regularización de Tikhonov, reemplazamos el problema discreto mal planteado por otro mejor condicionado. Damos cuatro valores distintos al parámetro de regularización  $\lambda$  y generamos las imágenes restauradas con las soluciones regularizadoras recomodadas en matrices  $F_\lambda$ . Para  $\lambda = 10^{-16}$ , vemos que la imagen de la Figura 4.34(a) está dominada por ruido, mientras que para  $\lambda = 10^{-15}$ , podemos distinguir claramente las letras de la imagen de la Figura 4.34(b). Si aumentamos a  $\lambda = 10^{-7}$  suavizamos la imagen como se muestra en la Figura 4.34(c). Con  $\lambda = 10^{-3}$ , la imagen se difumina. Véase Figura 4.34(d).



**Figura 4.34:** Imágenes restauradas por regularización de Tikhonov. Indicamos el valor del parámetro de regularización  $\lambda$  de cada solución regularizadora.

### Regularización de Imágenes Difuminadas con Ruido

En ocasiones, las imágenes adquiridas por dispositivos electrónicos tienen ruido. Por simplicidad, trabajamos con ruido aditivo gaussiano de media cero no correlacionado. Incorporamos este ruido en nuestro modelo discreto de difuminación con PSF espacialmente invariante y separable como

$$\text{vec}(G + E) = (R \otimes C)\text{vec}(F), \quad (4.16)$$

donde  $E$  es una matriz aleatoria de  $\mathbb{R}^{m \times n}$  que tiene el ruido gaussiano. Así que nuestro modelo de difuminación es un modelo lineal general de regresión.

El mal condicionamiento de  $R$  y  $C$  ocasiona que el ruido del lado derecho se propague en la solución del problema. En consecuencia, la imagen recuperada a partir de una imagen difuminada con ruido gaussiano de desviación estándar pequeña está dominada por el ruido. De aquí que necesitamos regularizar.

Recordamos que en la regularización de Tikhonov, la solución regularizada  $\mathbf{f}_\lambda$  de la Ecuación (4.16) es un estimador lineal y sesgado de la solución de cuadrados mínimos de norma mínima del mismo sistema de ecuaciones sin ruido, y el parámetro  $\lambda$  da un balance entre el sesgo de este estimador y la traza de su matriz de covarianza.

Una elección adecuada del parámetro de regularización  $\lambda$  nos permite reducir la influencia del ruido en la imagen restaurada. Elegimos el valor de  $\lambda$  con los criterios de L-curva y GCV. Estos criterios requieren de los valores singulares de  $A$  y del vector  $\mathbf{h}$  con los coeficientes de  $\mathbf{g}$  en la base de vectores singulares de derecha. Con el producto de Kronecker  $A = R \otimes C$ , podemos obtener los valores singulares de  $A$  y el vector  $\mathbf{h}$ .

**Ejemplo 4.6.** Retomemos la imagen *alma mater* de  $1600 \times 1200$  píxeles del Ejemplo 3.5. Esta vez la imagen se difumina por PSF con soporte  $[-799, 799] \times [-599, 599]$  y radio de desenfoque 7, las condiciones de frontera son periódicas y se agregó ruido gaussiano idénticamente distribuido de desviación estándar 0.01. Véase Figura 4.35.

Nuestro modelo lineal está dado por la ecuación

$$\text{vec}(G + E) = \text{Avec}(F), \quad (4.17)$$

donde  $A$  es BCCB de  $1200 \times 1200$  bloques de tamaño  $1600 \times 1600$ ,  $G \in \mathbb{R}^{1600 \times 1200}$  tiene los valores de los píxeles de la imagen difuminada sin ruido,  $E$  es matriz aleatoria con el ruido gaussiano,  $F \in \mathbb{R}^{1600 \times 1200}$  con los valores de los píxeles de la imagen restaurada.

Como la PSF de desenfoque no es separable, aproximamos la matriz  $A$  con un producto de Kronecker  $R \otimes C$ . Lo que hacemos es condensar la información de la matriz

$$K = \begin{bmatrix} k_{-799,-599} & \cdots & k_{-799,599} \\ \vdots & & \vdots \\ k_{799,-599} & \cdots & k_{799,599} \end{bmatrix}$$

en los vectores singulares  $\mathbf{u}$  y  $\mathbf{v}$  de izquierda y derecha, asociados al valor singular más grande  $\sigma$  de  $K$ . Multiplicamos  $\mathbf{u}$  y  $\mathbf{v}$  por  $\sqrt{\sigma}$  y les agregamos ceros para que tengan 1600 y 1200 componentes, respectivamente. Los vectores que obtenemos se usan como los primeros renglones de matrices circulares  $R$  y  $C$  de orden 1200 y 1600, respectivamente. Así que nuestro problema discreto de deblurring es hallar  $F \in \mathbb{R}^{1600 \times 1200}$  que cumpla la Ecuación (4.16).

A pesar de que  $R$  y  $C$  son matrices de rango completo con  $\kappa_2(R) = 6.1597 \times 10^3$  y  $\kappa_2(C) = 3.1334 \times 10^3$ , el inconveniente es que el ruido en el lado derecho se propaga en la solución de modo que el ruido domina a la imagen restaurada. Véase la Figura 4.36.



**Figura 4.35:** Imagen difuminada del *alma mater* con radio de desenfoque 7.



**Figura 4.36:** Imagen restaurada por solución de cuadrados mínimos de la ecuación  $(R \otimes C)\text{vec}(F) = \text{vec}(G + E)$ .

Usamos regularización de Tikhonov para reducir el ruido. A partir de las SVDs

$$R = U_R \Sigma_R V_R^T \quad \text{y} \quad C = U_C \Sigma_C V_C^T.$$

obtenemos la solución regularizadora  $\mathbf{f}_\lambda$ . Para ello formamos la matriz diagonal

$$\text{diag}(\sigma_1, \dots, \sigma_{1600 \cdot 1200}) = P(\Sigma_R \otimes \Sigma_C)P^T$$

con los valores singulares de  $R \otimes C$  ordenados de manera decreciente por una matriz de permutación  $P$ , así como el vector de coeficientes

$$\mathbf{h} = P \text{vec}(U_C^T (G + E) U_R^T)$$

de  $\text{vec}(G + E)$  en la base de vectores singulares de derecha de  $R \otimes C$ . De ese modo,

$$\mathbf{f}_\lambda = \text{vec}(V_C W V_R^T),$$

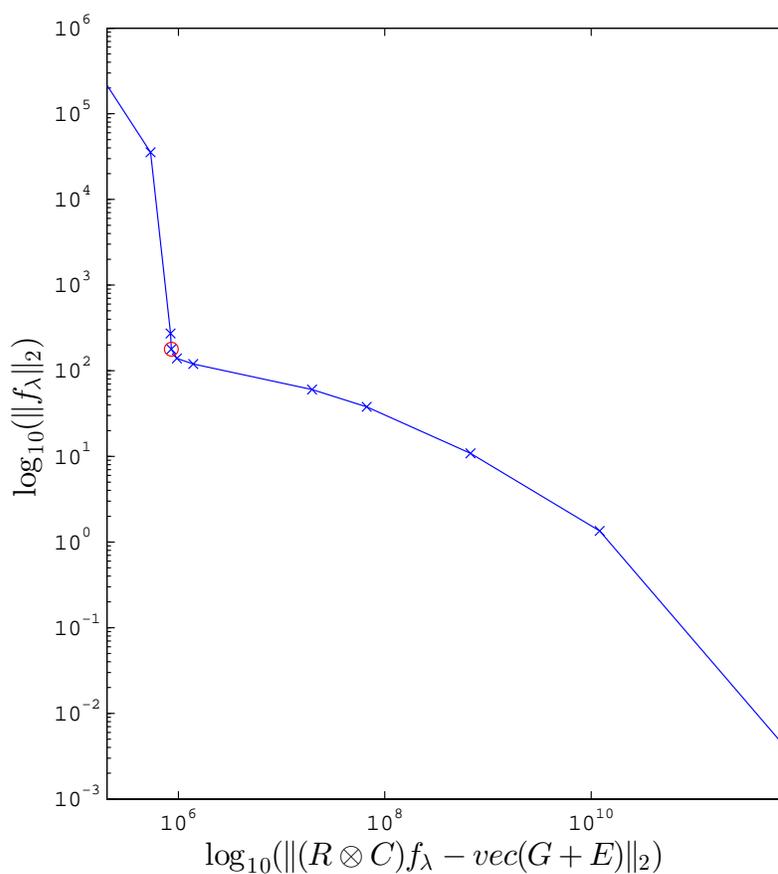
donde  $W$  es la matriz de tamaño  $1600 \times 1200$  dada por

$$\text{vec}(W) = P^T \text{diag} \left( \frac{\sigma_1}{\sigma_1^2 + \lambda^2}, \dots, \frac{\sigma_{1600 \cdot 1200}}{\sigma_{1600 \cdot 1200}^2 + \lambda^2} \right) \mathbf{h}.$$

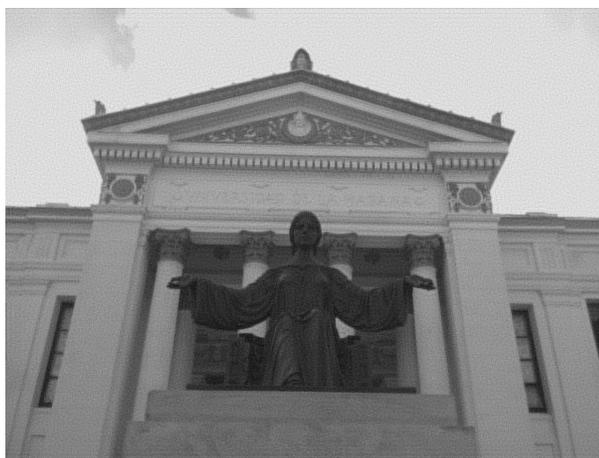
Elegimos el parámetro de regularización con el criterio de la  $L$ -curva. Recordamos que seleccionamos el valor de  $\lambda$  donde la curva

$$(\log_{10}(\|\mathbf{f}_\lambda\|_2), \log_{10}(\|(R \otimes C)\mathbf{f}_\lambda - \text{vec}(G + E)\|_2))$$

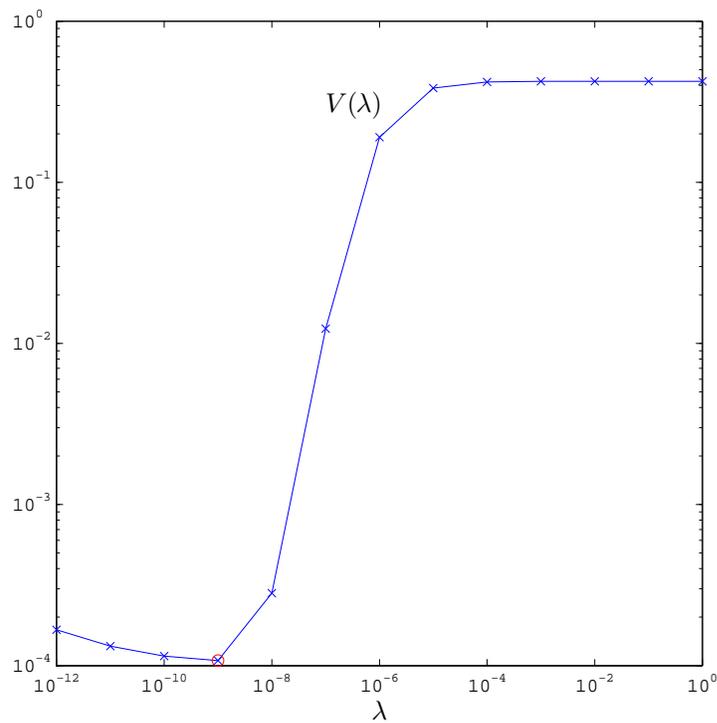
tiene la mayor curvatura. Véase Figura 4.37. Esta curva tiene forma de  $L$  y el punto de mayor curvatura está en la esquina de la  $L$ . En nuestro caso  $\lambda = 0.05$ . La imagen restaurada asociada se muestra en la Figura 4.38. Podemos apreciar detalles de la imagen que no se perciben con la solución de cuadrados mínimos  $\mathbf{f}_{LS}$ .



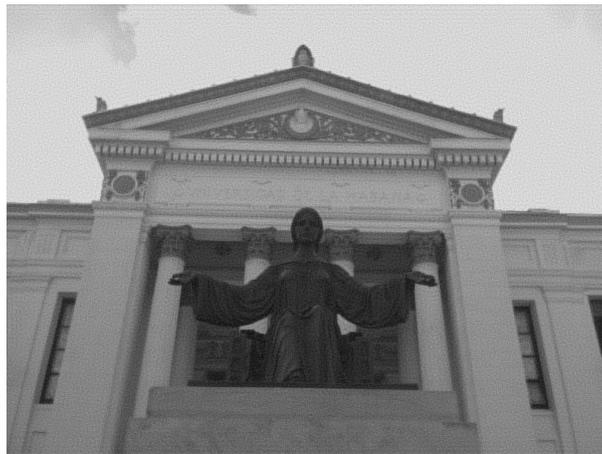
**Figura 4.37:** Gráfica en escala logarítmica base 10 de la L-curva dada por  $(\|f_\lambda\|_2, \|(R \otimes C)f_\lambda - \text{vec}(G + E)\|_2)$ . Marcamos la esquina con  $\circ$ . Este es el punto con  $\lambda = 0.05$



**Figura 4.38:** Imagen restaurada del *alma mater* por regularización de Tikhonov de la ecuación  $(R \otimes C)\text{vec}(F) = \text{vec}(G + E)$  con parametro de regularización  $\lambda = 0.05$



**Figura 4.39:** Gráfica en escala logarítmica base 10 del error predictivo  $V(\lambda)$ . Marcamos el mínimo  $\lambda_{GCV} = 10^{-9}$  de  $V$  con  $\circ$ .



**Figura 4.40:** Imagen restarada del *alma mater* por regularización de Tikhonov de la ecuación  $(R \otimes C)\text{vec}(F) = \text{vec}(G + E)$ . Usamos factores filtro  $\sigma_l/(\sigma_l^2 + 1600 \cdot 1200\lambda)$  en vez de  $\sigma_l/(\sigma_l^2 + \lambda^2)$  con  $\lambda = 10^{-9}$ .

Otra alternativa es usar GCV. En este caso usamos la solución regularizadora  $\mathbf{f}_\lambda$  con

$$\text{vec}(W) = P^T \text{diag} \left( \frac{\sigma_1}{\sigma_1^2 + 1600 \cdot 1200\lambda}, \dots, \frac{\sigma_{1600 \cdot 1200}}{\sigma_{1600 \cdot 1200}^2 + 1600 \cdot 1200\lambda} \right) \mathbf{h}.$$

Sean  $\mathbf{u}_1, \dots, \mathbf{u}_{1600 \cdot 1200}$  las columnas de  $(U_R \otimes U_C)P^T$ . El valor del parámetro que elegimos es el mínimo del error predictivo

$$V(\lambda) = 1600 \cdot 1200 \frac{\sum_{p=1}^{1600 \cdot 1200} \left( \frac{1600 \cdot 1200\lambda^2}{\sigma_p^2 + 1600 \cdot 1200\lambda} \right)^2 (\mathbf{u}_p^T \text{vec}(G + E))^2}{\left( \sum_{p=1}^{1600 \cdot 1200} \left( \frac{1600 \cdot 1200\lambda^2}{\sigma_p^2 + 1600 \cdot 1200\lambda} \right) \right)^2}$$

En la Figura 4.39 mostramos la gráfica de esta función. Nuestra aproximación del mínimo es  $\lambda = 10^{-9}$ . La imagen restaurada asociada se muestra en la Figura 4.40.

## 4.2. Super-resolución

En el problema de la Super-Resolución (SR), combinamos la información de varias imágenes de una misma escena para recuperar una imagen donde la densidad de píxeles sea más alta. La resolución de una imagen de acuerdo a la densidad de sus píxeles es la *resolución espacial*.

Para resolver el problema de SR, la idea que proponemos es verlo como el problema inverso del siguiente problema:

Dada una imagen  $\mathcal{F}$  de alta resolución (HR), obtener imágenes  $\mathcal{G}_1, \dots, \mathcal{G}_L$  de baja resolución (LR) de la misma escena.

Lo que hacemos es dar un modelo para generar imágenes de baja resolución a partir de una imagen ideal de alta resolución.

### 4.2.1. El modelo

Supongamos que la imagen  $\mathcal{G}_l$  de baja resolución se obtiene al aplicar una transformación  $\mathcal{H}_l$  a la imagen  $\mathcal{F}$  de alta resolución:

$$\mathcal{F} \xrightarrow{\mathcal{H}_l} \mathcal{G}_l$$

De acuerdo a [28] y [29], cada  $\mathcal{H}_l$  lleva a cabo tres transformaciones en el siguiente orden:

1. **Movimiento.** Mediante transformaciones geométricas  $\mathcal{W}_i$  movemos los puntos de la imagen. Por simplicidad, solo hacemos traslaciones con desplazamientos enteros. Con cada traslación generamos una imagen distinta.

2. **Difuminación.** Mediante un filtro  $\mathcal{B}$  difuminamos la imagen movida. Vamos a suponer que este filtro es lineal e invariante bajo traslaciones, más aún que difumina la imagen mediante una PSF espacialmente invariante y separable.
3. **Submuestreo.** Escalamos la imagen seleccionando muestras con un operador  $\mathcal{D}$ .

La estructura del espacio donde se define el operador  $\mathcal{H}_l$  es la de un espacio vectorial. Anteriormente, definimos el operador de difuminación  $\mathcal{B}$  sobre  $L^2(\mathbb{R}^2)$ . Así que vamos a definir la transformación  $\mathcal{H}_l$  sobre  $L^2(\mathbb{R}^2)$ .

El operador  $\mathcal{H}_l$  recibe a la función  $f : \mathbb{R}^2 \rightarrow [0, 1]$  asociada a la imagen  $\mathcal{F}$  y genera la función  $g_l : \mathbb{R}^2 \rightarrow [0, 1]$  asociada a la imagen  $\mathcal{G}_l$ :

$$g_l = \mathcal{H}_l(f).$$

Cuando movemos la imagen  $\mathcal{F}$ , aplicamos la traslación  $\mathcal{W}_l : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  dada por

$$\mathcal{W}_l \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x + x_l \\ y + y_l \end{pmatrix}.$$

La imagen móvida está dada por la función  $w_l = f \circ \mathcal{W}_l$ . Luego, aplicamos el filtro

$$B[w_l](x, y) = \int \int_{\mathbb{R}^2} k(x-s)k(y-t)w_l(s, t)dsdt.$$

De ese modo, la imagen difuminada está dada por la función  $b_l = \mathcal{B}[w_l]$ . Posteriormente, restringimos  $b_l$  al rectángulo  $[1, \tau m] \times [1, \tau n]$ , que separamos en rectángulos

$$R_{i,j} = [\tau(i-1), \tau i] \times [\tau(j-1), \tau j], \quad \begin{array}{l} i = 1, \dots, m, \\ j = 1, \dots, n, \end{array}$$

de tamaño  $\tau \times \tau$ , y tomamos muestras en cada  $R_{i,j}$  que representan promedios de  $b_l$ . Esto se hace mediante el operador  $\mathcal{D}$  sobre  $L^2(\mathbb{R}^2)$  dado por

$$\mathcal{D}[b_l](x, y) = \frac{1}{\tau^2} \int \int_{R_{i,j}} b_l(s, t)dsdt \quad \text{si } (x, y) \in R_{i,j}.$$

En consecuencia, el operador  $\mathcal{H}_l$  está dado por la composición

$$\mathcal{H}_l(f) = \mathcal{D}(\mathcal{B}(f \circ \mathcal{W}_l)).$$

Además de las transformaciones de movimiento, difuminación y submuestreo, nuestro modelo considera que cada imagen  $\mathcal{G}_l$  viene acompañada de ruido aditivo  $\epsilon_l$ , esto es,

$$g_l = \mathcal{H}_l(f) + \epsilon_l, \quad l = 1, \dots, L, \quad (4.18)$$

Por lo que

$$g_l = \mathcal{D}(\mathcal{B}(f \circ \mathcal{W}_l)) + \epsilon_l, \quad l = 1, \dots, L,$$

Esto nos dice que las imágenes LR generadas son versiones deformadas, degradadas, reescaladas y con ruido de la imagen HR como se muestra en la Figura 4.41.

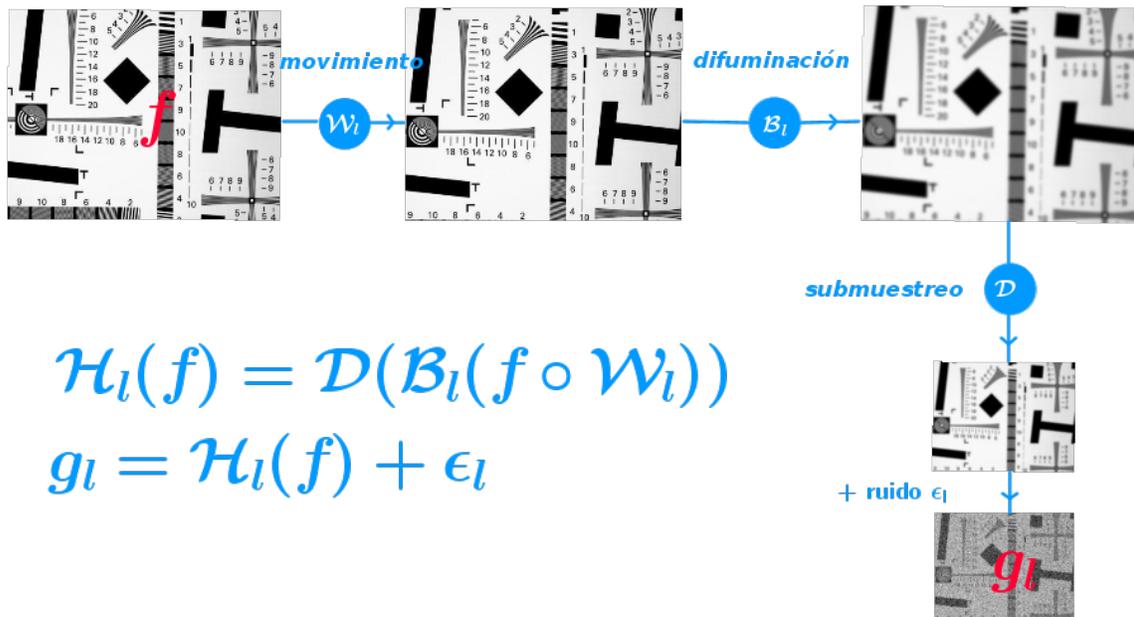
De aquí, el problema de generar imágenes LR a partir de una imagen HR es el siguiente

A partir de la función  $f : \mathbb{R}^2 \rightarrow [0, 1]$  asociada a la imagen  $\mathcal{F}$  y los operadores  $\mathcal{H}_l$  sobre  $L^2(\mathbb{R}^2)$  dados por

$$\mathcal{H}_l(f) = \mathcal{D}(\mathcal{B}(f \circ \mathcal{W}_l)), \quad l = 1, \dots, L,$$

obtener las funciones  $g_l : \mathbb{R}^2 \rightarrow \mathbb{R}$  asociadas a las imágenes  $\mathcal{G}_l$  tales que

$$g_l = \mathcal{H}_l(f) + \epsilon_l, \quad l = 1, \dots, L,$$



**Figura 4.41:** Modelo para obtener imágenes LR a partir de una imagen HR: La función  $f$  de la imagen HR se transforma por la composición de operadores  $\mathcal{H}_l$  en funciones  $g_l$  asociadas a las imágenes LR.

El problema de la SR es el problema inverso:

A partir de las funciones  $g_l : \mathbb{R}^2 \rightarrow \mathbb{R}$  asociadas a las imágenes  $\mathcal{G}_l$ , y los operadores  $\mathcal{H}_l$  sobre  $L^2(\mathbb{R}^2)$  dados por

$$\mathcal{H}_l(f) = \mathcal{D}(\mathcal{B}(f \circ \mathcal{W}_l)), \quad l = 1, \dots, L,$$

hallar una función  $f : \mathbb{R}^2 \rightarrow [0, 1]$  que cumpla las ecuaciones

$$g_l = \mathcal{H}_l(f) + \epsilon_l, \quad l = 1, \dots, L.$$

### 4.2.2. Discretización del modelo

Debido a que disponemos solamente de los píxeles de la imágenes LR, necesitamos discretizar nuestro modelo antes de resolver el problema de la SR.

A partir de una matriz  $F$  de tamaño  $\tau m \times \tau n$  con valores de la función  $f$ , generamos la matriz  $G_l$  de tamaño  $m \times n$  con los valores de  $g_l$  para  $l = 1, \dots, L$ . Lo que hacemos es discretizar las Ecuaciones (4.18) que usan los operadores  $\mathcal{H}_l$  en sistemas de ecuaciones lineales

$$\mathbf{g}_l = H_l \mathbf{f} + \boldsymbol{\epsilon}_l, \quad l = 1, \dots, L, \quad (4.19)$$

donde

$H_l$  es una matriz de tamaño  $mn \times \tau^2 mn$ ,

$\mathbf{g}_l = \text{vec}(G_l)$  es un vector de  $mn$  componentes,

$\boldsymbol{\epsilon}_l$  es un vector aleatorio de  $mn$  componentes,

$\mathbf{f} = \text{vec}(F)$  es un vector de  $\tau^2 mn$  componentes.

Vamos a obtener las matrices  $H_l$  a partir de la discretización de los tres transformaciones  $\mathcal{W}_l$ ,  $\mathcal{B}$  y  $\mathcal{D}$ . De acuerdo con [29], supongamos lo siguiente:

- \* Las muestras de la imágenes  $\mathcal{F}$  se ubican en una malla uniforme de  $\tau m \times \tau n$  píxeles del rectángulo  $[1, \tau m] \times [1, \tau n]$ .
- \* Las condiciones de frontera son periódicas.
- \* Los vectores aleatorios  $\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_L$  están distribuidos bajo una gaussiana de media cero y no están correlacionados.

#### Discretización de $\mathcal{W}_l$

La traslación  $\mathcal{W}_l$  desplaza los píxeles de la imagen  $\mathcal{F}$  con un desplazamiento vertical  $i_l \in \{1, \dots, \tau m\}$  y horizontal  $j_l \in \{1, \dots, \tau n\}$ . Cuando movemos la imagen, se crea una región donde los valores de los píxeles quedan libres. Como usamos condiciones de frontera periódicas, los píxeles desplazados se ponen en la región que se desocupa. Mediante la matriz de permutación circular

$$M_l = \begin{bmatrix} \mathbf{0}_{i_l \times (\tau m - i_l)} & I_{i_l \times i_l} \\ I_{(\tau m - i_l) \times (\tau m - i_l)} & \mathbf{0}_{(\tau m - i_l) \times i_l} \end{bmatrix}$$

hacemos una traslación hacia abajo como  $M_l F$  y hacia arriba como  $M_l^T F$ , mientras que la traslación horizontal se hace con la matriz de permutación circular

$$N_l = \begin{bmatrix} \mathbf{0}_{j_l \times (\tau n - j_l)} & I_{j_l \times j_l} \\ I_{(\tau n - j_l) \times (\tau n - j_l)} & \mathbf{0}_{(\tau n - j_l) \times j_l} \end{bmatrix}$$

como  $FN_l$  hacia la izquierda y  $FN_l^T$  hacia la derecha. Combinamos las traslaciones horizontales y verticales. Esto nos da cuatro matrices:

$$M_l FN_l, \quad M_l FN_l^T, \quad M_l^T FN_l, \quad M_l^T FN_l^T.$$

Por la propiedad de vectorización del producto de Kronecker, las columnas de la matriz  $M_l FN_l^T$  se apilan en el vector

$$\mathbf{w}_l = (N_l \otimes M_l) \mathbf{f}.$$

Lo mismo ocurre con las otras tres matrices, con la diferencia que tomamos  $N_l^T$  o  $M_l^T$ . Así que la traslación  $\mathcal{W}_l$  se discretiza en el producto de Kronecker

$$W_l = N_l \otimes M_l$$

de matrices de permutación circulares de orden  $\tau n$  y  $\tau m$ .

### Discretización de $\mathcal{B}$

Nuestro modelo discreto de difuminación por PSF espacialmente invariante y separable con condiciones de frontera periódicas nos dice que la evaluación de la función  $b_l = \mathcal{B}[w_l]$  en una malla uniforme del rectángulo  $[1, \tau m] \times [1, \tau n]$  nos da el vector

$$\mathbf{b}_l = B \mathbf{w}_l,$$

donde

$$B = R \otimes C$$

es el producto de Kronecker de matrices circulares  $R$  y  $C$  de orden  $\tau n$  y  $\tau m$  con elementos iguales a los valores de la PSF. En consecuencia,

$$\mathbf{b}_l = B W_l \mathbf{f}.$$

### Discretización de $\mathcal{D}$

En el submuestreo, tomamos muestras  $d_{i,j}^{(l)}$  de la transformada de  $b_l$  bajo  $\mathcal{D}$  en los puntos  $(x_i, y_j) \in R_{i,j}$  para  $i = 1, \dots, m$  y  $j = 1, \dots, n$ . Por lo que

$$d_{i,j}^{(l)} = \frac{1}{\tau^2} \iint_{R_{i,j}} b_l(s, t) ds dt.$$

Discretizamos la integral por sumas de Riemann. Así, obtenemos que

$$d_{i,j}^{(l)} = \frac{1}{\tau^2} \sum_{p=\tau(i-1)+1}^{\tau i} \sum_{q=\tau(j-1)+1}^{\tau j} b_l(p, q).$$

Reordenamos este promedio como

$$d_{i,j}^{(l)} = \frac{1}{\tau^2} \sum_{p,q=1}^{\tau} b_l(\tau(i-1) + p, \tau(j-1) + q),$$

y reacomodamos el vector  $\mathbf{b}_l$  como la matriz por bloques

$$B_l = \begin{bmatrix} B_{1,1}^{(l)} & \cdots & B_{1,n}^{(l)} \\ \vdots & & \vdots \\ B_{m,1}^{(l)} & \cdots & B_{m,n}^{(l)} \end{bmatrix},$$

donde

$$B_{i,j}^{(l)} = \begin{bmatrix} b_l(\tau(i-1) + 1, \tau(j-1) + 1) & \cdots & b_l(\tau(i-1) + 1, \tau j) \\ \vdots & & \vdots \\ b_l(\tau i, \tau(j-1) + 1) & \cdots & b_l(\tau i + 1, \tau j) \end{bmatrix}.$$

De ese modo, tenemos que

$$d_{i,j}^{(l)} = \frac{1}{\tau^2} \mathbf{1}_{1 \times \tau} B_{i,j}^{(l)} \mathbf{1}_{\tau \times 1}.$$

Sea

$$D_l = \begin{bmatrix} d_{1,1}^{(l)} & \cdots & d_{1,n}^{(l)} \\ \vdots & & \vdots \\ d_{m,1}^{(l)} & \cdots & d_{m,n}^{(l)} \end{bmatrix}.$$

Entonces

$$D_l = \frac{1}{\tau^2} (I_m \otimes \mathbf{1}_{1 \times \tau}) B_l (I_n \otimes \mathbf{1}_{\tau \times 1}).$$

Si apilamos las columnas de la matriz  $D_l$  en un vector  $\mathbf{d}_l$ , la propiedad de vectorización del producto de Kronecker implica que

$$\mathbf{d}_l = D \mathbf{b}_l,$$

donde

$$D = \frac{1}{\tau^2} (I_n \otimes \mathbf{1}_{1 \times \tau}) \otimes (I_m \otimes \mathbf{1}_{1 \times \tau})$$

es una matriz  $nm \times \tau^2 nm$ . Luego,

$$\mathbf{d}_l = DBW_l \mathbf{f}.$$

### Modelo Discreto para Generar Imágenes de Baja Resolución

De acuerdo a nuestro modelo,

$$\mathbf{g}_l = \mathbf{d}_l + \boldsymbol{\epsilon}_l,$$

esto es,

$$\mathbf{g}_l = DBW_l \mathbf{f} + \boldsymbol{\epsilon}_l.$$

Así, con las matrices

$$H_l := DBW_l, \quad l = 1, \dots, L$$

de tamaño  $nm \times \tau^2 nm$  formamos los Sistemas de ecuaciones lineales (4.19). De hecho, por la propiedad del producto mezclado del producto de Kronecker, se sigue

$$H_l = \frac{1}{\tau^2} ((I_n \otimes 1_{1 \times \tau}) R N_l) \otimes ((I_m \otimes 1_{1 \times \tau}) C M_l), \quad l = 1, \dots, L.$$

Podemos juntar los Sistemas de ecuaciones lineales (4.19) en uno solo. Sean

$$H = \begin{bmatrix} H_1 \\ \vdots \\ H_L \end{bmatrix}_{Lmn \times \tau^2 mn}, \quad \mathbf{g} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_L \end{bmatrix}_{Lmn \times 1} \quad \text{y} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_L \end{bmatrix}_{Lmn \times 1}.$$

Entonces nuestro modelo discreto para generar imágenes LR está dado por la ecuación

$$\mathbf{g} = H \mathbf{f} + \boldsymbol{\epsilon}. \quad (4.20)$$

**Ejemplo 4.7.** A partir de la imagen `tubos.jpg` de  $500 \times 400$  píxeles de la Figura 4.42, generamos dos imágenes LR de  $250 \times 200$  píxeles,  $\mathcal{G}_1$  y  $\mathcal{G}_2$ , con condiciones de frontera periódicas. Véanse Figuras 4.43(a) y 4.43(b). Tomamos la imagen  $\mathcal{G}_1$  de la Figura 4.43(a) como referencia. La otra imagen,  $\mathcal{G}_2$ , se ha movido  $i_2 = 5$  píxeles abajo y  $j_2 = 5$  píxeles a la derecha. Así que las matrices de traslación son

$$W_1 = I_{500 \cdot 400 \times 500 \cdot 400} \quad \text{y} \quad W_2 = \begin{bmatrix} \mathbf{0}_{5 \times 395} & I_{5 \times 5} \\ I_{395 \times 395} & \mathbf{0}_{395 \times 5} \end{bmatrix} \otimes \begin{bmatrix} \mathbf{0}_{5 \times 495} & I_{5 \times 5} \\ I_{495 \times 495} & \mathbf{0}_{495 \times 5} \end{bmatrix}.$$

Las imágenes LR están difuminadas por la PSF con radio de desenfoque 2.5. Como esta PSF no es separable, aproximamos la matriz de difuminación por el producto de Kronecker de matrices circulares  $R$  y  $C$ . Sean  $\mathbf{u}$  y  $\mathbf{v}$  los vectores singulares de derecha e izquierda asociados al valor singular más grande  $\sigma$  de la matriz

$$K = \begin{bmatrix} k_{-249, -199} & \cdots & k_{-249, 199} \\ \vdots & & \vdots \\ k_{249, -199} & \cdots & k_{249, 199} \end{bmatrix}$$

que tiene los valores de la PSF. Los primeros renglones de  $R$  y  $C$  son  $[\sqrt{\sigma} \mathbf{v}^T \ 0]$  y  $[\sqrt{\sigma} \mathbf{u}^T \ 0]$ , respectivamente.



**Figura 4.42:** Imagen HR tubos .jpg

Las dimensiones de la imagen `tubos .jpg` se reducen a la mitad con una tasa de muestreo  $\tau = 2$ . Así que la matriz de submuestreo es

$$D = \frac{1}{4}(I_{200} \otimes (1 \ 1)) \otimes (I_{250} \otimes (1 \ 1)).$$

De este modo, a partir del vector  $\mathbf{f}$  con los valores de los píxeles de la imagen HR, obtenmos los vectores

$$\mathbf{g}_l = DBW_l \mathbf{f}, \quad l = 1, 2$$

con los valores de los píxeles de las imágenes LR  $\mathcal{G}_1$  y  $\mathcal{G}_2$ .



(a) Imagen  $\mathcal{G}_1$



(b) Imagen  $\mathcal{G}_2$

**Figura 4.43:** Imágenes de baja resolución  $\mathcal{G}_1$  y  $\mathcal{G}_2$  de `tubos .jpg` obtenidas con el modelo  $\mathbf{g}_l = DBW_l \mathbf{f}$ ,  $l = 1, 2$ .

**Observaciones 4.11:**

☞ Si  $L > \tau^2$  (más observaciones que el cuadrado de la tasa de muestreo), entonces el Sistema de Ecuaciones (4.20) está sobredeterminado; mientras que si  $L < \tau^2$ , está indeterminado.

☞ Bajo las hipótesis anteriores sobre los vectores aleatorios  $\epsilon_l$ , el modelo para generar imágenes LR puede verse como un modelo lineal general de regresión.

☞ Si usamos condiciones de frontera cero, la región que queda libre se rellena con un fondo negro. Ahí, los píxeles valen cero. En este caso,

$$M_l = \begin{bmatrix} \mathbf{0}_{i_l \times (\tau m - i_l)} & \mathbf{0}_{i_l \times i_l} \\ I_{(\tau m - i_l) \times (\tau m - i_l)} & \mathbf{0}_{(\tau m - i_l) \times i_l} \end{bmatrix} \quad \text{y} \quad N_l = \begin{bmatrix} \mathbf{0}_{j_l \times (\tau n - j_l)} & \mathbf{0}_{j_l \times j_l} \\ I_{(\tau n - j_l) \times (\tau n - j_l)} & \mathbf{0}_{(\tau n - j_l) \times j_l} \end{bmatrix},$$

mientras que  $R$  y  $C$  son matrices de Toeplitz.

☞ Si los desplazamientos no son enteros, los valores de los píxeles en la malla desplazada se obtienen mediante interpolación bilineal, bicúbica o de Lanczos [13]. La traslación  $\mathcal{W}_l$  se realiza sobre una versión interpolada de la imagen HR [29], [73].

☞ Otras maneras para formar la matriz  $H_l$  con los operadores  $\mathcal{D}$ ,  $\mathcal{B}$  y  $\mathcal{W}_l$  pueden consultarse en Capel [15].

**4.2.3. Método para resolver el problema de la SR**

Por simplicidad, vamos a considerar un caso particular del problema de la SR:

A partir de dos imágenes LR  $\mathcal{G}_1$  y  $\mathcal{G}_2$  de  $m \times n$  píxeles, reconstruir una imagen HR  $\mathcal{F}$  de  $\tau m \times \tau n$  píxeles.

**Motivación**

Para resolver el problema de la SR, tratamos de revertir las transformaciones  $\mathcal{D}$ ,  $\mathcal{B}$  y  $\mathcal{W}_l$  que generan las imágenes LR mediante la composición  $\mathcal{D}(\mathcal{B}(f \circ \mathcal{W}_l))$ . En [34] y [80] sugieren cambiar el orden de  $\mathcal{B}$  y  $\mathcal{W}_l$  en la composición. Esto es posible con nuestras hipótesis, pues como el producto de matrices circulares del mismo orden es conmutativo, la propiedad del producto mezclado del producto de Kronecker nos permite intercambiar las matrices  $B$  y  $W_l$ :

$$\begin{aligned} BW_l &= (R \otimes C)(N_l \otimes M_l) \\ &= (RN_l) \otimes (CM_l) \\ &= (N_l R) \otimes (M_l C) \\ &= (N_l \otimes M_l)(R \otimes C) \\ &= W_l B. \end{aligned}$$

Luego,

$$H_l = DW_l B.$$

Por eso manejamos la composición  $\mathcal{D}(B(f) \circ \mathcal{W}_l)$ . Así, de ser invertibles las tres transformaciones, la imagen HR se obtiene con la inversa de esta composición:

$$\mathcal{B}^{-1}(\mathcal{D}^{-1}(g_l) \circ \mathcal{W}_l^{-1}).$$

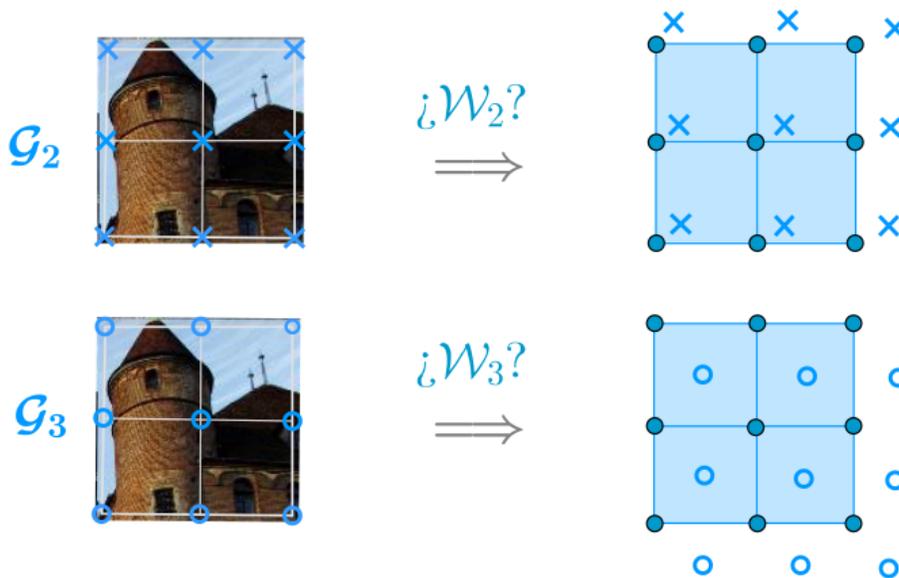
Esto nos dice que primero, debemos revertir el submuestreo, luego el movimiento, y posteriormente la difuminación. En vez de obtener inversas de las tres transformaciones, resolvemos un problema inverso por cada transformación aplicada.

### Upsampling

En el submuestreo, el problema es obtener muestras de la imagen HR  $\mathcal{F}$  para obtener una imagen LR. El problema inverso llamado *Upsampling* es obtener una imagen HR a partir de las muestras de la imagen LR  $\mathcal{G}_l$ . Lo que hacemos es formar una malla de tamaño  $\tau m \times \tau n$  a partir de los  $m \times n$  píxeles de  $\mathcal{G}_l$ , de modo que generamos una rejilla de tamaño  $\tau \times \tau$  por cada píxel de  $\mathcal{G}_l$ , donde todos los puntos tienen el mismo valor que el píxel. De esta manera, transformamos la matriz  $G_l$  de tamaño  $m \times n$  en el producto de Kronecker

$$S_l = G_l \otimes 1_{\tau \times \tau},$$

donde  $1_{\tau \times \tau}$  es la matriz de tamaño  $\tau \times \tau$  con todos sus elementos iguales a uno.



**Figura 4.44:** Las transformaciones geométricas  $\mathcal{W}_2$  y  $\mathcal{W}_3$  mueven la imagen  $\mathcal{G}_1$  para generar las imágenes  $\mathcal{G}_2$  y  $\mathcal{G}_3$ . En el registro, alineamos los puntos  $\times$  y  $\circ$  de las imágenes  $\mathcal{G}_2$  y  $\mathcal{G}_3$  con los puntos  $\bullet$  de la malla de  $\mathcal{G}_1$ , respectivamente. Si desconocemos  $\mathcal{W}_2$  y  $\mathcal{W}_3$ , hacemos estimaciones.

### Registro

Para mover la imagen HR  $\mathcal{F}$ , hacemos traslaciones  $\mathcal{W}_l$  con un desplazamiento horizontal  $j_l$  y uno vertical  $i_l$  mediante el producto  $M_l F N_l^T$ . Ahora, el problema es dados los desplazamientos  $i_l$  y  $j_l$ , alinear las imágenes  $\mathcal{G}_l$ , respecto a un mismo sistema de coordenadas. El proceso de alineación se llama **Registro**. Véase Figura 4.44. Tomemos como referencia las coordenadas de la imagen  $\mathcal{G}_1$ , de modo que  $M_l = I_{m \times m}$  y  $N_l = I_{n \times n}$ . Para alinear los píxeles de  $\mathcal{G}_2$ , usamos los mismos desplazamientos, pero en sentido contrario. Esto lo hacemos con el producto  $M_2^T F N_2$ . De aquí que formemos

$$Z_1 = S_1 \quad \text{y} \quad Z_2 = M_2^T S_2 N_2.$$

#### Observaciones 4.12:

 Si en el registro desconocemos las transformaciones  $\mathcal{W}_1, \dots, \mathcal{W}_L$ , debemos estimar el movimiento dado por  $\mathcal{W}_l$  a partir de los valores de la función  $g_l$ . Una manera es la siguiente: Sea  $\Omega$  la región donde se traslapan las imágenes  $\mathcal{G}_1, \dots, \mathcal{G}_L$ . buscamos la transformación  $\mathcal{W}_l$  que minimize el funcional

$$\phi(\mathcal{W}) = \int_{\Omega} [g_1(\mathcal{W}(x, y)) - g_l(x, y)]^2 dx dy$$

sobre todas las transformaciones de  $L^2(\Omega)$  [105]. En caso de que  $\mathcal{W}_l$  sea la transformación afín

$$\mathcal{W}_l \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos \theta_l & -\sin \theta_l \\ \sin \theta_l & \cos \theta_l \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} x_l \\ y_l \end{pmatrix},$$

basta estimar los parámetros  $\theta_l, x_l, y_l$  a partir de los valores de la función  $g_l$ . Al respecto, en [68] aproximan  $g_1(\mathcal{W}_l(x, y))$  por el polinomio de Taylor

$$p_l(x, y) = g_1(x, y) + \left( x_l - y\theta_l - x\frac{\theta_l^2}{2} \right) \frac{\partial g_1}{\partial x} + \left( y_l + x\theta_l - y\frac{\theta_l^2}{2} \right) \frac{\partial g_1}{\partial y}.$$

y buscan  $x_l, y_l, \theta_l$  que minimizen el error

$$E(x_l, y_l, \theta_l) = \sum_{(x, y) \in \Omega} [p_l(x, y) - g_l(x, y)]^2.$$

Esta estimación es válida si los valores de los parámetros son pequeños. Para un estudio más completo sobre registro de imágenes, consulte [12].

### Fusión

Cuando revertimos las transformaciones de submuestreo y movimiento, generamos  $L$  imágenes HR. A partir de éstas, queremos generar una sola imagen HR. El proceso que hace esta tarea se llama **Fusión**. La idea es combinar la información que aporta cada

imagen en una sola. La fusión que realizamos combina las descomposiciones en valores singulares de las matrices que tienen los valores de los píxeles [89].

Manejamos dos imágenes LR  $\mathcal{G}_1$  y  $\mathcal{G}_2$ . Tomamos la primera como imagen de referencia. De acuerdo a Devi, Madhu y Lal Kishore [26], fusionamos primero las imágenes antes de resolver el problema de deblurring. Entonces, después de reescalar  $\mathcal{G}_1$  y  $\mathcal{G}_2$  y trasladar la segunda, obtenemos las matrices  $Z_1$  y  $Z_2$  de tamaño  $\tau m \times \tau n$ , respectivamente. Sean  $\sigma_{\text{máx}}^1$  y  $\sigma_{\text{máx}}^2$  los valores singulares más grandes de  $Z_1$  y  $Z_2$ , y sean

$$Z_1 = U_1 \Sigma_1 V_1^T \quad \text{y} \quad Z_2 = U_2 \Sigma_2 V_2^T$$

sus respectivas SVD's. Definimos

$$\Sigma = \begin{cases} \Sigma_1, & \text{si } \sigma_{\text{máx}}^1 \geq \sigma_{\text{máx}}^2, \\ \Sigma_2, & \text{si } \sigma_{\text{máx}}^1 < \sigma_{\text{máx}}^2, \end{cases}$$

y formamos la matriz

$$Z = U_2 \Sigma V_2^T.$$

Así, comparando los valores singulares más grandes, decidimos si combinamos las imágenes al tomar  $\Sigma = \Sigma_1$  o ignoramos una contribución con  $\Sigma = \Sigma_2$ .

### Deblurring

El problema de difuminación por PSF espacialmente invariante es una convolución, ya que el operador  $\mathcal{B}$  de difuminación realiza la convolución de la función asociada a la imagen con la PSF. El problema inverso (el deblurring) es una deconvolución. Con PSF separable y condiciones de frontera periódicas, nuestro problema discreto de deconvolución es dada la matriz  $Z$  de tamaño  $\tau m \times \tau n$  de la imagen difuminada, hallar la matriz  $F$  del mismo tamaño que cumple la ecuación

$$\text{vec}(Z) = (R \otimes C) \text{vec}(F). \quad (4.21)$$

La matriz  $F$  tiene los valores de los píxeles de la imagen restaurada.

### Método para SR: Upsampling + Registro + Fusión + Deblurring

En resumen, el método que usamos para SR hace lo siguiente:

1. Realiza upsampling:

$$S_l = G_l \otimes 1_{\tau \times \tau}, \quad l = 1, 2.$$

2. Registra imágenes

$$Z_1 = S_1 \quad \text{y} \quad Z_2 = M_2^T S_2 N_2.$$

### 3. Fusiona imágenes

calcula SVD's  $Z_l = U_l \Sigma_l V_l^T$ ,  $l = 1, 2$ .  
 compara  $\sigma_{\text{máx}}^1$  y  $\sigma_{\text{máx}}^2$  para formar matriz  $Z$ .

### 4. Resuelve la Ecuación (4.21) del problema de Deblurring.

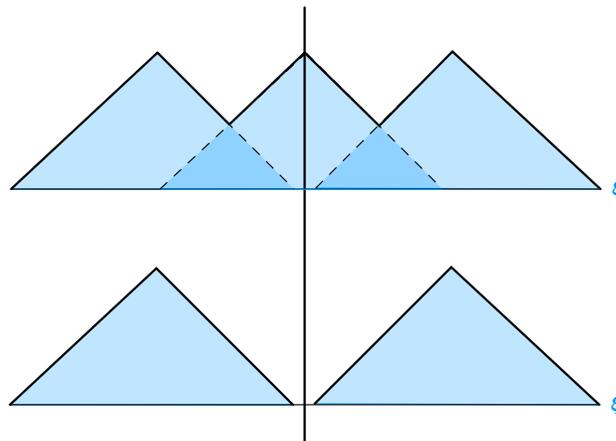
Al final obtenemos una matriz  $F$  de tamaño  $\tau m \times \tau n$  con los valores de los píxeles de la imagen HR. Véase Figura 4.46.

#### Observaciones 4.13:

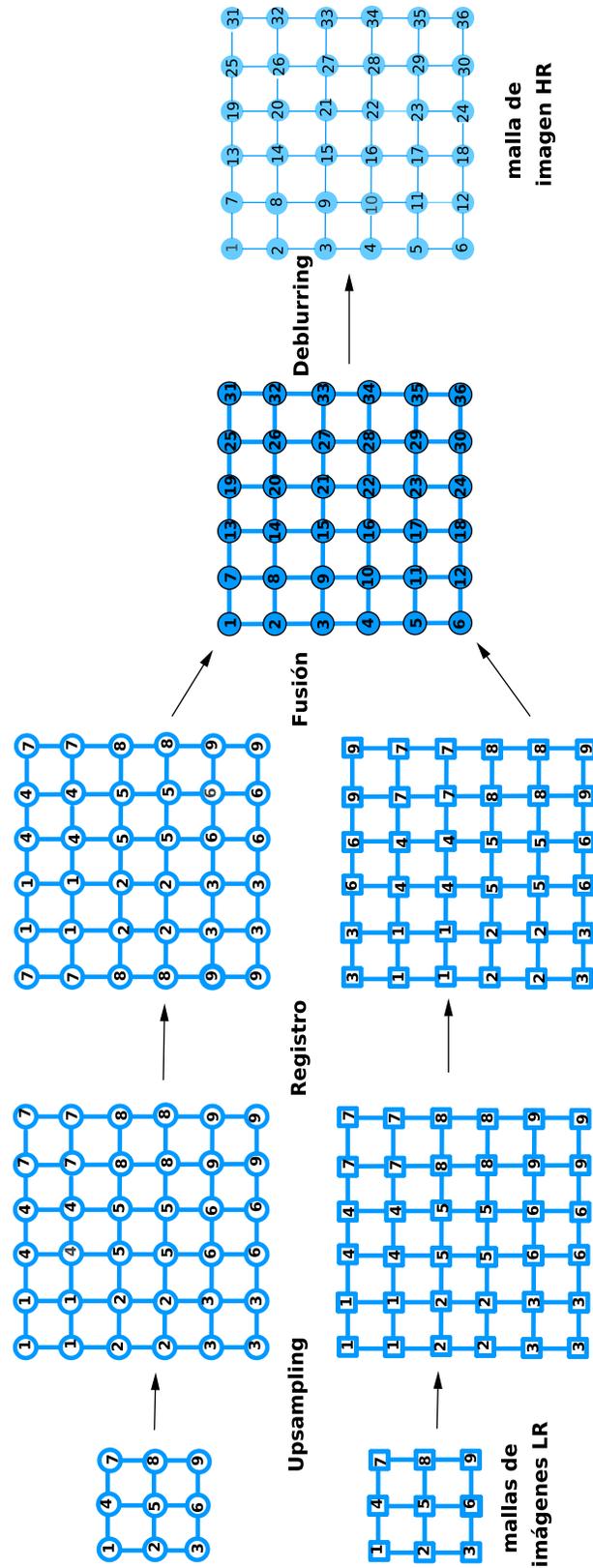
👉 Otra manera de resolver el problema de SR es usar la Transformada Discreta de Fourier (DFT) de la matriz  $F$  asociada a la imagen HR [69], [95]:

$$\hat{f}_{j_1, j_2} = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} f(p, q) e^{-i2\pi(pj_1 + qj_2)}$$

Con esta transformación pasamos del dominio espacial al dominio de frecuencias. La idea es descomponer en el dominio de las frecuencias las señales solapadas asociadas a las imágenes LR en partes de una señal no solapada de la imagen HR. Véase Figura 4.45. Luego de separar las señales, formamos un sistema de ecuaciones que relacione los coeficientes de la DFT de las imágenes LR con los de la DFT de  $F$ . De hecho, desde las primeras publicaciones sobre SR, como la de Tsai y Huang [116] de 1984, hasta en publicaciones más recientes como la de Vandewall [119] de 2006 se usa este enfoque.



**Figura 4.45:** Descomposición de señales solapadas en el dominio de frecuencia. La variable  $\xi$  indica que trabajamos en este dominio.



**Figura 4.46: Método para SR.** Aumentamos la resolución de las dos imágenes LR en el upsampling, repitiendo cada píxel cuatro veces. En el registro alineamos sus píxeles. Movemos los píxeles de una imagen una posición a la derecha, y una posición abajo los de la otra. Luego, fusionamos las imágenes en una sola de 36 píxeles. Al final, reducimos degradaciones en el deblurring.

#### 4.2.4. Regularización en SR

El inconveniente que tenemos con el método propuesto es que el problema de deblurring es un problema mal planteado en el sentido de Hadamard. Esto ocasiona que la SR sea a su vez un problema mal planteado. La dificultad se presenta al resolver el Sistema de Ecuaciones Lineales (4.21). A pesar de que las matrices  $R$  y  $C$  sean invertibles, su mal condicionamiento ocasiona que la imagen HR resturada este dominada por el ruido introducido por errores de redondeo en el cálculo de la matriz  $F$ .

Para reducir las degradaciones de la imagen HR que obtenemos con nuestro método para SR, regularizamos el problema de Deblurring. Esto lo podemos hacer con la SVD truncada de  $R \otimes C$  o con regularización de Tikhonov.

**Ejemplo 4.8.** Retomemos las imágenes LR  $\mathcal{G}_1$  y  $\mathcal{G}_2$  de  $250 \times 200$  píxeles que mostramos en la Figura 4.43 del Ejemplo 4.7. La primera es la imagen de referencia y la otra se obtiene al desplazar  $i_2 = 5$  píxeles abajo y  $j_2 = 5$  píxeles a la derecha. Ambas se obtienen al reescalar la imagen `tubos.jpg` de la Figura 4.42 con un factor de muestreo  $\tau = 2$  y están difuminadas por la PSF con radio de desenfoque 2.5 y soporte  $[-249, 249] \times [-199, 199]$ . Usamos condiciones de frontera periódicas.



(a) Imagen  $\mathcal{G}_1$



(b) Imagen  $\mathcal{G}_2$

Ponemos los valores de los píxeles de las imágenes LR en matrices  $G_1, G_2 \in \mathbb{R}^{250 \times 200}$  y los valores de la PSF en una matriz  $K$  de tamaño  $499 \times 399$ . Sean  $\mathbf{u}$  y  $\mathbf{v}$  los vectores singulares de derecha e izquierda asociados al valor singular  $\sigma$  más grande de  $K$ . Aproximamos la matriz BCCB del modelo discreto de difuminación por el producto de Kronecker  $R \otimes C$  de matrices circulares  $R$  y  $C$  de tamaño  $200 \times 200$  y  $250 \times 250$  con primeros renglones iguales a  $[\sqrt{\sigma} \mathbf{v}^T \ 0]$  y  $[\sqrt{\sigma} \mathbf{u}^T \ 0]$ , respectivamente.

Usamos el método propuesto para SR.

1. Upsampling: 
$$S_l = G_l \otimes I_{2 \times 2}, \quad l = 1, 2.$$

2. Registro:

$$Z_1 = S_1 \quad \text{y} \quad Z_2 = \begin{bmatrix} \mathbf{0}_{5 \times 495} & I_{5 \times 5} \\ I_{495 \times 495} & \mathbf{0}_{495 \times 5} \end{bmatrix}^T S_2 \begin{bmatrix} \mathbf{0}_{5 \times 395} & I_{5 \times 5} \\ I_{395 \times 395} & \mathbf{0}_{395 \times 5} \end{bmatrix}$$

3. Fusión:

$$Z_l = U_l \Sigma_l V_l^T, \quad l = 1, 2.$$

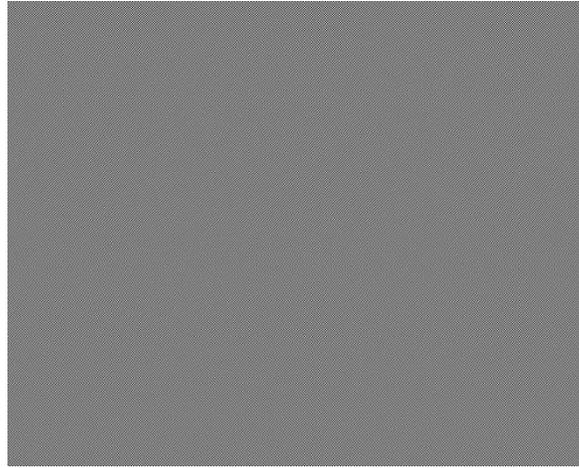
$$\Sigma = \begin{cases} \Sigma_1, & \text{si } \sigma_{\text{máx}}^1 \geq \sigma_{\text{máx}}^2, \\ \Sigma_2, & \text{si } \sigma_{\text{máx}}^1 < \sigma_{\text{máx}}^2, \end{cases}$$

$$Z = U_2 \Sigma V_2^T$$

4. Deblurring: En nuestro caso, las matrices  $R$  y  $C$  son invertibles, por lo que

$$\text{vec}(F) = (R^{-1} \otimes C^{-1}) \text{vec}(Z).$$

$$F = C^{-1} Z R^{-1T}.$$



**Figura 4.48:** Imagen HR restaurada por upsampling, registro, fusión y deblurring de las dos imágenes LR de la Figura 4.43. Usamos las inversas de  $R$  y  $C$  en el deblurring.

En principio, esperamos que la matriz  $F$  de tamaño  $500 \times 400$  nos de una imagen HR con menos degradaciones que las imágenes LR, pues  $\mathcal{G}_1$  y  $\mathcal{G}_2$  no tienen ruido aditivo, además  $\kappa_2(R) = 3.795 \times 10^3$  y  $\kappa_2(C) = 3.1334 \times 10^3$ . Sin embargo, el upsampling, el registro y la fusión introducen perturbaciones pequeñas en las matrices  $G_1$  y  $G_2$  de las imágenes LR. Estas perturbaciones presentes en la matriz  $Z$  se amplifican en la solución del problema de deblurring, dando lugar la imagen dominada por ruido de la Figura 4.48.

Regularizamos el problema de deblurring dado por la Ecuación (4.21). Primero usamos SVD truncada con nivel de truncamiento  $\rho$ . Recordamos que los valores singulares del producto de Kronecker  $R \otimes C$  son los valores singulares de  $R$  por los de  $C$  en orden decreciente. Por eso  $\rho \leq \text{rank}(R)\text{rank}(C) = 200000$ . En la Figura 4.49(a) vemos la imagen restaurada para  $\rho = 2000$ . Ésta presenta degradaciones notables. Con  $\rho = 25000$ , reducimos las degradaciones como puede verse en la Figura 4.49(b). De aumentar el nivel de truncamiento, el ruido domina la imagen como se muestra en la Figura 4.49(c) con  $\rho = 75000$ .



(a) Imagen HR con  $\rho = 2000$ .

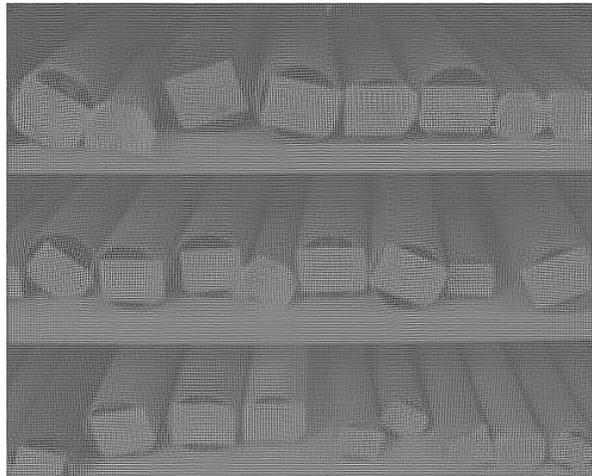


(b) Imagen HR con  $\rho = 25000$ .



(c) Imagen HR con  $\rho = 75000$ .

**Figura 4.49:** Imágenes HR obtenidas por upsampling, registro, fusión y deblurring de las dos imágenes LR de la Figura 4.43. Usamos SVD truncada con nivel de truncamiento  $\rho$  en el deblurring.



(a) Imagen HR con  $\lambda = 0.01$ .



(b) Imagen HR con  $\lambda = 0.1$ .



(c) Imagen HR con  $\lambda = 0.4$ .

**Figura 4.50:** Imágenes HR obtenidas por upsampling, registro, fusión y deblurring de las dos imágenes LR de la Figura 4.43. Usamos regularización de Tikhonov con parámetro  $\lambda$  en el deblurring.

Ahora, empleamos regularización de Tikhonov en el problema mal planteado con tres valores distintos del parámetro de regularización  $\lambda$ . Para  $\lambda = 0.01$ , el ruido aún domina la imagen como puede verse en la Figura 4.50(a). Si incrementamos el valor del parámetro a  $\lambda = 0.01$ , disminuimos la influencia del ruido en la imagen restaurada como se muestra en la Figura 4.50(b). Esta imagen aún tiene degradaciones. Para  $\lambda = 0.4$ , el filtro que desenfoca la imagen tiene mayor peso que el ruido. Por eso la imagen de la Figura 4.50(c) se ve más clara que las otras.

#### Observaciones 4.14:

 En [91], Nguyen, Milanfar y Golub resuelven un caso particular del problema de SR de otra manera. Ellos usan regularización de Tikhonov en la ecuación  $\mathbf{g} = H\mathbf{f} + \boldsymbol{\epsilon}$  y resuelven las ecuaciones normales regularizadas mediante CG con ayuda de preconditionadores circulares por bloques. Eligen el parámetro de regularización con GCV.



## CONCLUSIONES

---

En el trabajo realizado, los problemas mal planteados que abordamos están dados por la ecuación integral de Fredholm de primera clase. Su discretización da lugar a un sistema de ecuaciones lineales mal condicionado o incluso mal planteado. Remarcamos los siguientes puntos:

- \* La SVD juega un papel importante en la explicación del mal condicionamiento y en el diseño de métodos de regularización
- \* Revisamos diferentes métodos de regularización, haciendo hincapié en sus aspectos numéricos y estadísticos.
- \* Queremos abordar el el CG-TRS en problemas de gran escala. Buscar un preconditionador adecuado en este escenario requiere un tratamiento cuidadoso.
- \* La deconvolución es central en problemas asociados con la restauración de imágenes. Nosotros la abordamos en el dominio espacial. Nos interesa tratarla más adelante en el dominio de las frecuencias.
- \* Cuando la PSF es separable en el deblurring, aprovechamos la estructura de las matrices para reducir sus dimensiones mediante el producto de Kronecker. De ese modo, podemos emplear métodos de regularización por factores filtro y criterios para elegir el parámetro sin formar matrices por bloques.
- \* Abordar tanto el registro como el deblurring en la super-resolución hace complicado el problema. Estamos empezando a desarrollar métodos. Requerimos más experiencia en el área.
- \* Usamos la biblioteca REGUTOOLS en los métodos de regularización por factores filtro, la rutina `dgqt` en el TRS, `HNO` en deblurring. Para los ejemplos de super-resolución, programamos en `Matlab`.



## Proyecciones Ortogonales

**Definición.** Sea  $\mathcal{V}$  un espacio vectorial real de dimensión finita  $m$ . Equipamos a este espacio con un producto interno  $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$ .

\* Decimos que dos vectores  $q_1, q_2 \in \mathcal{V}$  son ortogonales si

$$\langle q_1, q_2 \rangle = 0.$$

Si además,

$$\|q_1\|_2 = \|q_2\|_2 = 1,$$

decimos que  $q_1$  y  $q_2$  son **vectores ortonormales**.

\* Sea  $\mathcal{S}$  un subespacio de  $\mathcal{V}$  de dimensión  $n \leq m$ . Una **base ortonormal**  $\mathcal{B}$  para  $\mathcal{S}$  es un conjunto de  $n$  vectores  $q_1, \dots, q_n \in \mathcal{S}$  tales que

$$\langle q_i, q_j \rangle = \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{en otro caso.} \end{cases}$$

\* El **complemento ortogonal** de  $\mathcal{S}$  es el subespacio

$$\mathcal{S}^\perp = \{q \in \mathcal{V} : \langle x, q \rangle = 0 \ \forall x \in \mathcal{S}\}$$

\* La **proyección ortogonal** sobre  $\mathcal{S}$  es la transformación lineal  $p : \mathcal{V} \rightarrow \mathcal{S}$  dada por

$$p(v) = \langle v, q_1 \rangle q_1 + \dots + \langle v, q_n \rangle q_n \quad \forall v \in \mathcal{V}.$$

### Observaciones:

 Todo vector  $v \in \mathcal{S}$  se expande en la base ortonormal  $\mathcal{B}$  como

$$v = \langle v, q_1 \rangle q_1 + \dots + \langle v, q_n \rangle q_n.$$

 La proyección ortogonal de  $\mathbb{R}^m$  sobre la recta generada por  $\mathbf{q} \in \mathbb{R}^m \setminus \{\mathbf{0}\}$  está dada por

$$p(\mathbf{x}) = \frac{\mathbf{q}\mathbf{q}^T}{\mathbf{q}^T\mathbf{q}} \mathbf{x} \quad \forall \mathbf{q} \in \mathbb{R}^m,$$

☞ La proyección ortogonal de  $\mathbb{R}^m$  sobre el subespacio generado por vectores ortonormales  $\mathbf{q}_1, \dots, \mathbf{q}_n \in \mathbb{R}^m$  se expresa como

$$p(\mathbf{x}) = [\mathbf{q}_1 \cdots \mathbf{q}_n][\mathbf{q}_1 \cdots \mathbf{q}_n]^T \mathbf{x} \quad \forall \mathbf{x} \in \mathbb{R}^m.$$

☞ Si  $\mathbf{q}_1, \dots, \mathbf{q}_m$  forman una base ortonormal de  $\mathbb{R}^m$ , entonces

$$[\mathbf{q}_1 \cdots \mathbf{q}_m][\mathbf{q}_1 \cdots \mathbf{q}_m]^T = I.$$

Una propiedad importante de la proyección ortogonal es que nos da la distancia mínima entre un vector dado y los vectores del subespacio  $\mathcal{S}$  con la norma inducida por el producto interno:

**Teorema** ([79]). *Sea  $\mathcal{V}$  un espacio con producto interno, sea  $\mathcal{S}$  un subespacio de  $\mathcal{V}$  y sea  $\mathbf{b}$  un vector de  $\mathcal{V}$ . Entonces la proyección ortogonal de  $\mathbf{b}$  sobre  $\mathcal{S}$  es el único vector  $p(\mathbf{b}) \in \mathcal{S}$  tal que*

$$\|p(\mathbf{b}) - \mathbf{b}\| = \min_{q \in \mathcal{S}} \|q - \mathbf{b}\|.$$

### Observaciones:

☞ Sea  $\mathbf{p}$  la proyección ortogonal de  $\mathbf{b} \in \mathbb{R}^m$  sobre  $\text{Col}(A)$ . Entonces el teorema anterior nos dice que  $\|\mathbf{b} - \mathbf{p}\|_2$  minimiza el tamaño del residuo  $\|\mathbf{b} - A\mathbf{x}\|_2$ . Esto justifica el papel de la proyección ortogonal sobre  $\text{Col}(A)$  en el problema lineal de cuadrados mínimos.

Toda matriz ortogonal  $Q \in \mathbb{R}^{m \times m}$  tiene las siguientes propiedades geométricas:

\*  $Q$  preserva el producto interno, es decir,

$$(Qu)^T(Qv) = u^T v \quad \forall u, v \in \mathbb{R}^m.$$

\*  $Q$  preserva la distancia entre vectores, esto es,

$$\|Qu - Qv\|_2 = \|u - v\|_2 \quad \forall u, v \in \mathbb{R}^m.$$

Además, las matrices ortogonales tienen las siguientes propiedades algebraicas:

\*  $Q$  es ortogonal  $\implies Q^T$  es ortogonal.

\*  $Q_{m \times m}$  y  $P_{m \times m}$  son ortogonales  $\implies QP$  es ortogonal.

\*  $Q_{m \times m}$  es ortogonal  $\implies \text{rank}(A) = \text{rank}(QA)$  para toda matriz  $A_{m \times n}$ .

## Descomposición en Valores Singulares

Sea  $A$  una matriz real de tamaño  $m \times n$  de rango  $r$ . Queremos dar bases ortonormales para las columnas de  $A$ .

Sean  $\lambda_1, \dots, \lambda_n$  los valores propios de  $A^T A$ . Como esta matriz es simétrica, sus valores propios son reales, más aún para esta matriz en particular son no negativos. Así que podemos ordenarlos de manera no creciente como  $\lambda_1 \geq \dots \geq \lambda_n$ . El **Teorema Espectral** nos dice que existe una base ortonormal de  $\mathbb{R}^n$  formada por vectores propios  $\mathbf{v}_1, \dots, \mathbf{v}_n$  de la matriz  $A^T A$ , donde cada  $\mathbf{v}_i$  se asocia al valor propio  $\lambda_i$  como

$$A^T A \mathbf{v}_i = \lambda_i \mathbf{v}_i \quad i = 1, \dots, n.$$

Estas ecuaciones implican que los vectores  $A \mathbf{v}_1, \dots, A \mathbf{v}_n$  son ortogonales y que sus respectivos tamaños en norma euclidea son

$$\sigma_1 = \sqrt{\lambda_1}, \dots, \sigma_n = \sqrt{\lambda_n}.$$

Puesto que el rango de  $A$  es  $r$ , tenemos que

$$\sigma_{r+1} = \dots = \sigma_n = 0,$$

Por lo que

$$A \mathbf{v}_{r+1} = \dots = A \mathbf{v}_n = \mathbf{0}.$$

Así que al expandir cualquier vector de  $\text{Ker}(A)$  en la base de vectores propios, encontramos que los vectores  $\mathbf{v}_{r+1}, \dots, \mathbf{v}_n$  forman una base ortonormal de  $\text{Ker}(A)$ .

Dado que  $A$  tiene rango  $r$  y los vectores  $A \mathbf{v}_1, \dots, A \mathbf{v}_r$  son ortogonales, tenemos que los vectores

$$\mathbf{u}_i = \frac{A \mathbf{v}_i}{\sigma_i}, \quad i = 1, \dots, r$$

forman una base ortonormal para  $\text{Col}(A)$ .

Podemos completar  $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$  a una base ortonormal  $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$  de  $\mathbb{R}^m$  con ayuda del proceso de Gram-Schmidt. Como los primeros  $r$  vectores de esta base generan  $\text{Col}(A)$  y son ortogonales a los otros  $m - r$  vectores, se sigue que  $\mathbf{u}_{r+1}, \dots, \mathbf{u}_m$  son ortogonales a todo vector de  $\text{Col}(A)$ . Por consiguiente,  $A^T \mathbf{u}_{r+1}, \dots, A^T \mathbf{u}_m$  son ortogonales a cualquier vector de  $\mathbb{R}^n$ . Luego,

$$A^T \mathbf{u}_{r+1} = \dots = A^T \mathbf{u}_m = \mathbf{0}.$$

Esto implica que los vectores  $\mathbf{u}_{r+1}, \dots, \mathbf{u}_m$  forman una base ortonormal para  $\text{Ker}(A^T)$ .

Los vectores  $\mathbf{u}_1, \dots, \mathbf{u}_r$  cumplen las ecuaciones

$$A \mathbf{v}_i = \sigma_i \mathbf{u}_i, \quad i = 1, \dots, r.$$

Multiplicamos ambos lados por  $A^T$ , y como los  $\mathbf{v}_i$ 's son vectores propios de  $A^T A$ , obtenemos

$$A^T \mathbf{u}_i = \sigma_i \mathbf{v}_i, \quad i = 1, \dots, r.$$

Si expandimos cualquier vector de  $\mathcal{R}^m$  en la base  $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$  y multiplamos por  $A^T$ , entonces las dos relaciones anteriores implican que los vectores  $\mathbf{v}_1, \dots, \mathbf{v}_r$  forman una base ortonormal para  $\text{Col}(A^T)$ .

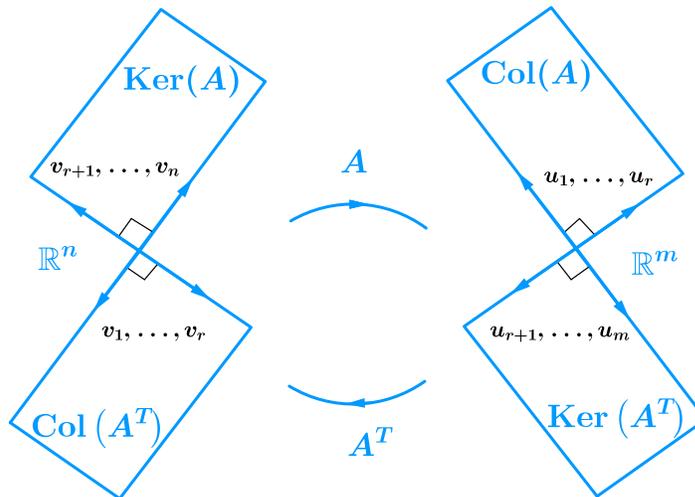
Podemos expresar en forma matricial las relaciones entre las bases ortonormales:

$$A \begin{matrix} \mathbf{v}_1 & \cdots & \mathbf{v}_r & \mathbf{v}_{r+1} & \cdots & \mathbf{v}_n \\ \mathbf{V} \end{matrix} = \begin{matrix} \mathbf{u}_1 & \cdots & \mathbf{u}_r & \mathbf{u}_{r+1} & \cdots & \mathbf{u}_m \\ \mathbf{U} \end{matrix} \begin{matrix} \left[ \begin{array}{cc|c} \sigma_1 & 0 & \mathbf{0} \\ & \ddots & \\ 0 & \sigma_r & \mathbf{0} \\ \hline \mathbf{0} & & \mathbf{0} \end{array} \right]_{m \times n} \\ \mathbf{\Sigma} \end{matrix}$$

En resumen,

- \*  $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$  es base ortonormal de  $\text{Col}(A)$ ,
- \*  $\{\mathbf{u}_{r+1}, \dots, \mathbf{u}_m\}$  es base ortonormal de  $\text{Ker}(A^T)$ ,
- \*  $\{\mathbf{v}_1, \dots, \mathbf{v}_r\}$  es base ortonormal de  $\text{Col}(A^T)$ ,
- \*  $\{\mathbf{v}_{r+1}, \dots, \mathbf{v}_n\}$  es base ortonormal de  $\text{Ker}(A)$ .

Las columnas de las matrices  $U$  y  $V$  son bases ortonormales de  $\mathbb{R}^m$  y  $\mathbb{R}^n$ , respectivamente. Por lo que  $U$  y  $V$  son matrices ortogonales. En consecuencia,  $A$  puede factorizarse como  $A = U\Sigma V^T$ . Enunciemos formalmente lo que hemos conseguido.



**Figura 51:** Relación entre los núcleos y espacios columna de  $A$  y  $A^T$

**Teorema.** Para cada matriz  $A \in \mathbb{R}^{m \times n}$  de rango  $r$ , existen matrices ortogonales  $U \in \mathbb{R}^{m \times m}$  y  $V \in \mathbb{R}^{n \times n}$ , y una matriz

$$\Sigma = \left[ \begin{array}{cc|c} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \\ \hline & \mathbf{0} & \mathbf{0} \end{array} \right]_{m \times n} \quad \text{con } \sigma_1 \geq \dots \geq \sigma_r > 0$$

tal que

$$A = U\Sigma V^T. \quad (22)$$

Los escalares  $\sigma_1, \dots, \sigma_r$  se llaman **valores singulares** de  $A$ . Cuando  $r$  es menor que  $p = \min\{m, n\}$ , decimos que  $A$  tiene  $p - r$  valores singulares iguales a cero. Las columnas de las matrices  $U$  y  $V$  se llaman **vectores singulares** de izquierda y derecha, respectivamente. La Factorización (22) se conoce como **Descomposición en Valores Singulares (SVD)**.

### Observaciones:

☞ Los valores singulares de  $A$  son las raíces cuadradas de los valores propios positivos de  $A^T A$ .

☞ El número de valores singulares positivos es igual al rango de  $A$ .

☞ Los vectores singulares de derecha forman una base ortonormal de vectores propios de  $A^T A$  para  $\mathbb{R}^n$ , de éstos, los primeros  $r$  forman una base ortonormal de  $\text{Col}(A^T)$  y los otros  $n - r$  nos dan una base ortonormal de  $\text{Ker}(A)$ .

☞ Los vectores singulares de izquierda forman una base ortonormal de vectores propios de  $AA^T$  para  $\mathbb{R}^m$ , de éstos, los primeros  $r$  forman una base ortonormal de  $\text{Col}(A)$ , mientras que el resto forma una base ortonormal de  $\text{Ker}(A^T)$ .

☞ Si consideramos solamente los valores singulares positivos de  $A$  y sus vectores singulares asociados, entonces podemos escribir la SVD en forma compacta como

$$A = [\mathbf{u}_1 \ \dots \ \mathbf{u}_r] \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{bmatrix} [\mathbf{v}_1 \ \dots \ \mathbf{v}_r]^T.$$

☞ La matriz  $A$  es la suma de  $r$  matrices de rango uno:

$$A = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

Hasta ahora hemos tratado el aspecto teórico de la SVD, Desde el punto de vista práctico, nos interesa calcular eficientemente esta factorización matricial. Una manera de obtener una SVD de  $A$  consiste de dos pasos:

1. Buscamos matrices ortogonales  $U_A \in \mathbb{R}^{m \times m}$  y  $V_A \in \mathbb{R}^{n \times n}$  para reducir la matriz  $A$  a una matriz bidiagonal :

$$B = \begin{bmatrix} d_1 & c_1 & & & & \\ & d_2 & c_2 & & & \\ & & \ddots & \ddots & & \\ & & & d_{n-1} & c_{n-1} & \\ & & & & & d_n \end{bmatrix}.$$

de modo que

$$A = U_A \begin{bmatrix} B \\ \mathbf{0} \end{bmatrix}_{m \times n} V_A^T,$$

Este proceso se llama *bidiagonalización*. En [36], Golub y Kahan nos dicen que con un producto de reflectores de Householder es posible obtener las matrices ortogonales  $U_A$  y  $V_A$ .

2. Calculamos una SVD de la bidiagonal:  $B = U_B \Sigma V_B^T$  mediante un proceso iterativo. La idea es construir una sucesión  $\{B_n\}$  de matrices bidiagonales que converga a la diagonal de valores singulares de  $A$ . Los mismos autores de [36] proponen usar rotaciones de Givens para generar la sucesión  $\{B_n\}$ .

### Observaciones:

 Los valores singulares de  $A$ ,  $B$  así como los de cada  $B_n$  son los mismos porque multiplicamos por matrices ortogonales.

Cuando juntamos las matrices obtenidas, conseguimos la SVD

$$A = (U_A U_B) \Sigma (V_A V_B)^T.$$

Una explicación más detallada sobre como emplear las matrices de Householder y de Givens para introducir ceros en matrices y sobre la convergencia de la sucesión  $\{B_n\}$  pueden consultarse en [74]. En [24] se muestran otras formas de calcular la SVD.

## Pseudo-inversa

Para problemas de cuadrados mínimos de rango completo, la solución de cuadrados mínimos de la ecuación  $A\mathbf{x} = \mathbf{b}$ , está dada por  $\mathbf{x} = (A^T A)^{-1} A^T \mathbf{b}$ . En el caso de problemas de cuadrados mínimos de rango deficiente nos interesa tener una expresión para la solución

$\mathbf{x}_{LS}$  de cuadrados mínimos de norma mínima. Para ello, damos una generalización de la inversa de una matriz cuando la matriz no es cuadrada o de rango completo.

Dada una matriz  $A \in \mathbb{R}^{m \times n}$  de rango  $r$  con SVD  $A = U\Sigma V^T$ , donde  $\sigma_1 \geq \dots \geq \sigma_r$  son sus valores singulares (positivos), definimos la matriz  $A^\dagger = V\Sigma^\dagger U^T$ , donde

$$\Sigma^\dagger = \left[ \begin{array}{cc|c} \frac{1}{\sigma_1} & 0 & \\ & \ddots & \mathbf{0} \\ 0 & & \frac{1}{\sigma_r} \\ \hline \mathbf{0} & & \mathbf{0} \end{array} \right]_{m \times n}.$$

La matriz  $A^\dagger$  se conoce como pseudo-inversa de  $A$ .

**Observaciones:**

 Si  $A$  es invertible,  $A^\dagger = A^{-1}$

 Si  $A$  es de rango completo por columnas y no es cuadrada, entonces  $A^\dagger = (A^T A)^{-1} A^T$ .

 La solución de cuadrados mínimos de norma euclidiana mínima para el sistema de ecuaciones lineales  $A\mathbf{x} = \mathbf{b}$  es  $\mathbf{x}_{LS} = A^\dagger \mathbf{b}$ .

La pseudo-inversa tiene las siguientes caracterizaciones que pueden encontrarse en la literatura [37]:

\*  $A^\dagger$  es la única solución de norma mínima Frobenius al problema

$$\min_{X \in \mathbb{R}^{m \times n}} \|AX - I_{m \times m}\|_F$$

\*  $A^\dagger$  es la única matriz  $X \in \mathbb{R}^{m \times n}$  que satisface las condiciones de Moore-Penrose

$$\begin{aligned} AXA &= A, & (AX)^T &= AX, \\ XAX &= X, & (XA)^T &= XA. \end{aligned}$$

**Observaciones:**

 Las condiciones de Moore-Penrose implican que  $AA^\dagger$  y  $A^\dagger A$  son proyecciones ortogonales sobre las imágenes de  $A$  y  $A^T$ , respectivamente.

## Descomposición Generalizada en Valores Singulares

Una manera de simplificar el análisis del método de regularización de Tikhonov con suavizamiento es modificar las bases que nos da la SVD para incluir tanto a la matriz de coeficientes como la matriz con la restricción de suavizamiento. La Descomposición Generalizada en Valores Singulares (GSVD) nos provee una representación de la solución regularizada que nos permite ver a la regularización de Tikhonov con suavizamiento como un método de regularización por factores filtro.

**Teorema (GSVD [37]).** Sean  $A \in \mathbb{R}^{m_1 \times n_1}$  y  $B \in \mathbb{R}^{m_2 \times n_1}$  con  $m_1 \geq n_1$  y

$$r = \text{rango} \left( \begin{bmatrix} A \\ B \end{bmatrix} \right).$$

Entonces existen matrices ortogonales  $U_1 \in \mathbb{R}^{m_1 \times m_1}$  y  $U_2 \in \mathbb{R}^{m_2 \times m_2}$  y una matriz invertible  $X \in \mathbb{R}^{n_1 \times n_1}$  tales que

$$U_1^T A X = D_A := \begin{bmatrix} I_{p \times p} & \mathbf{0}_{p \times (r-p)} & \mathbf{0}_{p \times (n_1-r)} \\ \mathbf{0}_{(r-p) \times p} & \text{diag}(\alpha_1, \dots, \alpha_{r-p}) & \mathbf{0}_{(r-p) \times (n_1-r)} \\ \mathbf{0}_{(m_1-r) \times p} & \mathbf{0}_{(m_1-r) \times (r-p)} & \mathbf{0}_{(m_1-r) \times (n_1-r)} \end{bmatrix},$$

$$U_2^T B X = D_B := \begin{bmatrix} \mathbf{0}_{p \times p} & \mathbf{0}_{p \times (r-p)} & \mathbf{0}_{p \times (n_1-r)} \\ \mathbf{0}_{(r-p) \times p} & \text{diag}(\beta_1, \dots, \beta_{r-p}) & \mathbf{0}_{(r-p) \times (n_1-r)} \\ \mathbf{0}_{(m_2-r) \times p} & \mathbf{0}_{(m_2-r) \times (r-p)} & \mathbf{0}_{(m_2-r) \times (n_1-r)} \end{bmatrix},$$

donde  $p = \max\{r - m_2, 0\}$  y  $\alpha_i, \beta_i$ ,  $i = 1, \dots, r - p$  son constantes no negativas.

### Observaciones:

☞ Si  $B = I_{n \times n}$  y  $X = U_2$ , entonces  $U_1^T A X = D_A$  y  $U_2^T B X = D_B$  nos dan SVD de  $A$ .

Para ver como usamos la GSVD en regularización de Tikhonov, consideramos el problema discreto mal planteado  $A\mathbf{x} = \mathbf{b}$ . Regularizamos el problema con el método de Tikhonov con suavizamiento, esto es, resolver

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda^2 \|B\mathbf{x}\|_2^2,$$

donde la matriz  $B$  es una discretización del tamaño de la primera o segunda derivada de una función cuadrado integrable.

El mínimo está dado por la solución de las ecuaciones normales regularizadas

$$(A^T A + \lambda^2 B^T B)\mathbf{x} = A^T \cdot \mathbf{b}$$

Cuando  $B$  es invertible, las ecuaciones  $U_1^T AX = D_A$  y  $U_2^T BX = D_B$  transforman las ecuaciones normales regularizadas en

$$\text{diag}(\alpha_1^2 + \lambda^2 \beta_1^2, \dots, \alpha_n^2 + \lambda^2 \beta_n^2) \mathbf{y} = D_A^T \hat{\mathbf{b}},$$

donde  $\mathbf{x} = X\mathbf{y}$  y  $\hat{\mathbf{b}} = U_1^T \mathbf{b}$ . En consecuencia, la solución regularizada se puede expresar en la base formada por las columnas  $\mathbf{x}_1, \dots, \mathbf{x}_n$  de  $X$  como

$$\mathbf{x}_\lambda = \sum_{k=1}^n \left( \frac{\alpha_k \hat{b}_k}{\alpha_k^2 + \lambda^2 \beta_k^2} \right) \mathbf{x}_k$$

### Observaciones:

 Con la representación anterior de la solución regularizada  $\mathbf{x}_\lambda$ , los términos

$$\frac{\alpha_k}{\alpha_k^2 + \lambda^2 \beta_k^2}$$

juegan un papel análogo al de los factores filtro de la Sección §3.2. El parámetro de regularización  $\lambda$  controla la influencia de la dirección  $\mathbf{x}_k$  en  $\mathbf{x}_\lambda$  junto con los valores de  $\alpha_k$  y  $\beta_k$ .

Consulte Golub [37] y Linz [76] para más detalles sobre la GSVD, y a Hansen [51] para ver otros aspectos que la relacionan con métodos de regularización.



- [1] R. C. Aster, B. Borchers, C. H. Thurber. *Parameter Estimation and Inverse Problems*. Elsevier Inc., 2005
- [2] G. Allaire, S.M. Kaber. *Numerical Linear Algebra*. Springer, 2008.
- [3] L. Allen, R. Angel, J. D. Mangus, G. A Rodney, R. R. Shannon, C. P. Spoelhof. *The Hubble Space Telescopy Optical System Failure Report*. NASA-TM-103443, November 1990.
- [4] R. L. Allen, D. W. Mills. *Signal Analysis: Time, Frequency, Scale, and Structure*. IEEE Press, 2004.
- [5] P. Barrera, V. Hernández, C. Durán. *El ABC de los Splines*. Sociedad Matemática Mexicana, 1996.
- [6] J. Barsley, S. Jefferies, J. Nagy, R. Plemmons. *Restoration of Images with an Unknown, Spatially-Varying Blur*. Optical Society of America, 2005.
- [7] J. V. Beck, B. Blackwell, C. R. St. Clair Jr. *Inverse Heat conduction: Ill-posed Problems*. John Wiley & Sons, 1985.
- [8] R. E. Bellman, R. S. Roth. *The Laplace Transform*. World Scientific Publishing Co Ltd, 1984.
- [9] M. W. Berry, S. T. Dumais, G. W. O'Brien. *Using Linear Algebra for Intelligent Information Retrieval*. SIAM Review Vol. 37, No. 4 (1995), pp. 573-595.
- [10] A. Björck. *Numerical Methods in Matrix Computations*. Springer, 2015.
- [11] M. Born, E. Wolf. *Principles of optics: Electromagnetic theory of propagation, interference and difraction of light*. Cambridge University Press, 1999.
- [12] L. G. Brown. *A survey of image registration techniques*. ACM Computing Surveys, Vol. 24 Issue 4 (1992), pp 325-376.
- [13] W. Burger, M. J. Burge. *Digital Image Processing: An Algorithmic Introduction using Java*. Springer, 2008.
- [14] D. Callaerts, B. De Moor, J.Vanderwalle, W. Sansen, G. Vantrappen, J. Janssens. *Comparison of SVD methods to extract the foetal electrocardiogram from cutaneous electrode signals*. Med. & Biol. Eng. & Comput. 28 (1990), pp. 217-224,
- [15] D. Capel. *Image Mosaicing and Super-Resolution*. Springer, 2004.

- 
- [16] A. Carasso. *Determining Surface Temperatures from Interior Observations*. SIAM J. on Appl. Math., Vol 42, No. 3, (1982) pp. 558-574.
- [17] F. Carmona. *Modelos Lineales*. Universidad de Barcelona, 2003.
- [18] S. Chatterjee, A .S. Hadi. *Regression Analysis by example*. Jonh Wiley & Sons, Inc. Fifth Edition, 2012.
- [19] N. Christophersen, O.C. Lingjaerde. *Regularization Principles: Solving ill-posed inverse problems*. 1998.
- [20] N. Clinthorne, T. Pan, P. Chiao, W. Rogers, J. Stamos. *Preconditioning methods for improved convergence rates in iterative reconstruction*. IEEE Trans. Med. Imag., Vol 12. No.1 (1993), pp. 78-83.
- [21] D. Colton, R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*. Springer, 2013.
- [22] A. R. Conn, N. I. M Gould, P. L. Toint. *Trust-Region Methods*. MPS-SIAM, 2000.
- [23] B. N. Datta. *Numerical Linear Algebra and Applications*. Brooks/Cole Publications, 1995.
- [24] J. W. Demmel. *Applied Numerical Linear Algebra*. SIAM, 1997.
- [25] J. L. Dennis, R.B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM, 1996.
- [26] A. G. Devi, T. Madhu, K. Lal Kishore. *An Improved Super Resolution Image Reconstruction using SVD based Fusion and Blind Deconvolution techniques*. International Journal of Signal Processing, Image Procesing and Pattern Recognition Vol. 7, No. 1 (2014), pp. 239-298.
- [27] P. H. C. Eilers, B. D. Marx. *Flexible Smoothing with B-splines and Penalties*. Statistical Science Vol. 11, No. 2 (1996), pp. 89-121.
- [28] M. Elad, A. Feuer. *Restoration of a Single Superresolution Image from Several Blurred, Noisy, and Undersampled Measured Images*. IEEE Transaction on Image Processing Vol. 6, No. 12 (1997), pp. 1646-1658.
- [29] M. Elad, Y. Hel-Or. *A Fast Super-Resolution Reconstruction Algorithm for Pure Translational Motion and Common Space-Invariant Blur*. IEEE Transaction on Image Processing Vol. 10, No. 8 (2001), pp. 1187-1193.
- [30] L. Eldén. *Algorithms for the regularization of ill-conditioned Least-Squares Problems*. BIT Vol. 17 (1977), pp. 134-145.

- 
- [31] L. Eldén. *The numerical solution of a non-characteristic Cauchy problem for a parabolic equation* en *Numerical Treatment of Inverse Problems in Differential and Integral Equations: Proceedings of an International Workshop, Germany, 1982*. Birkhäuser, 1983, pp. 246-268.
- [32] L. Eldén. *Matrix Methods in Data Mining and Pattern Recognition*. SIAM, 2007.
- [33] H.W. Engl, M. Hanke, A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 1996.
- [34] S. Farsiu, D. Robinson, M. Elad, P. Milanfar. *Fast and robust multi-frame super-resolution*. IEEE Transaction on Image Processing Vol. 13, No. 10 (2004), pp. 1327-1344.
- [35] W. Gander. *Least squares with a quadratic constraint*. Numer. Math. 36 (1981), pp. 291-307.
- [36] G. Golub, W. Kahan. *Calculating The Singular Values and Pseudo-Inverse of a matrix*. J. SIAM Numer. Anal. Ser. B, Vol. 2, No. 2 (1965), pp. 205- 224.
- [37] G. H. Golub, C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Fourth Edition, 2013.
- [38] G. H. Golub, U. Von Matt. *Quadratically constrained least squares and quadratic problems*. Numer. Math. 59 (1991), pp. 561-580.
- [39] G. H. Golub, M. Heath, G. Wahba. *Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter*. Technometrics, Vol. 21, No. 2 (1979), pp. 215-223.
- [40] R. C. Gonzalez, R. E. Woods. *Digital Image Processing*. Prentice Hall, Second Edition, 2002.
- [41] C.W. Groetsch. *Inverse Problems in the Mathematical Sciences*. Vieweg, 1993.
- [42] C.W. Groetsch. *Inverse Problems: Activities for Undergraduates*. The Mathematical Association of America, 1999.
- [43] V. Guerra, V. Hernández. *Numerical aspects in locating the corner of the L-curve*. Approximation, Optimization and Mathematical Economics, Physica-Verlag, Heidelberg, 2001, pp. 121-131.
- [44] J. Hadamard. *Sur les problèmes aux dérivées partielles et leur signification physique*. Bull. Univ. Princeton, 13 (1902) pp 49-52. (Œuvres 3, pp 1099-1105).
- [45] M. Hanke. *Conjugate gradient type methods for ill-posed problems*. Longman Scientific & Technical, 1995.

- 
- [46] M. Hanke, P. C. Hansen. *Regularization methods for large-scale problems*. Surv. of Math. Ind. 3 (1993), pp. 253-315.
- [47] P. C. Hansen. *The truncated SVD as method for regularization*. BIT 27 (1987), pp. 534-553.
- [48] P. C. Hansen. *Computation of the Singular Value Expansion*. Computing 40 (1988), pp. 185-199.
- [49] P. C. Hansen. *The Discrete Picard Condition for Discrete Ill-Posed Problems*. BIT 30 (1990), pp. 658-672.
- [50] P. C. Hansen. **REGULARIZATION TOOLS: A Matlab package for analysis and solution of discrete ill-posed problems**. Numerical Algorithms 6 (1994), pp. 1-35.
- [51] P. C. Hansen. *Rank-Deficient and discrete ill-posed problems*. SIAM, 1998.
- [52] P. C. Hansen *The L-curve and its use in the numerical treatment of inverse problems*. Invited chapter in Computational Inverse Problems in Electrocardiology. WIT Press (2001), pp. 119-142.
- [53] P. C. Hansen. *Deconvolution and regularization with Toeplitz matrices*. , Numerical Algorithms 29 (2002), pp. 323-378.
- [54] P. C. Hansen *Discrete Inverse Problems: Insight and Algorithms*. SIAM, 2010.
- [55] P.C. Hansen, J.G. Nagy, D.P. O’Leary. *Deblurring images: Matrices, spectra and filtering*. SIAM, 2006.
- [56] R. J. Hanson. *A numerical method for solving Fredholm integral equations of the first using singular values*. SIAM J. Numer. Anal., 8 (1971), pp. 616-622.
- [57] R. J. Hanson, J. L. Phillips. *An Adaptive numerical method for solving linear Fredholm integral equations of the first kind*. Numer. Math. 24 (1975), pp. 291-307.
- [58] C. H. Herrera. *Ridge Regression*. Master Thesis, Iowa State University, 1981.
- [59] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Second Edition, 2002.
- [60] A. Hoerl, R. Kennard. *Ridge Regression: Biased Estimation for Nonorthogonal Problems*. Technometrics, Vol. 12, No. 1 (1970), pp. 55-67.
- [61] V. A. Il’in. *Tikhonov’s work on methods of solving ill-posed problems*. Russ. Math. Surv. Vol 22 No. 2 (1967), pp. 142-149.
- [62] A. K. Jain. *Fundamentals of Digital Image Processing*. Prentice Hall, 1989.

- 
- [63] D. R. Jensen, D. E. Ramirez. Variations on Ridge Traces in Regression. *Communications in Statistics - Simulation and Computation*, Vol. 41 (2012), pp. 265-278.
- [64] I. Jovanović. *Inverse Problems in Acoustic Tomography: Theory and Applications* Ph.D. Thesis Ecole Polytechnique Federale de Lausanne, Suisse, 2008.
- [65] S.I. Kabanikhin. *Definitions and examples of inverse and ill-posed problems* Journal of Inv. Ill-Posed Problems, Vol 16 (2008), pp. 317-357.
- [66] J. Kamm. *Singular value decomposition based methods for signal and image restoration*. PhD thesis, Southern Methodist University, Dallas, TX, 1998.
- [67] J. Kamm, J.G. Nagy. *Kronecker products and SVD approximations for separable spatially variant blurs*. 1998.
- [68] D. Keren, S. Pegel, R. Brada. *Image Sequence Enhancement Using Sub-Pixel Displacements*. Computer Vision and Pattern Recognition, 1998.
- [69] S. P. Kim, W. Su. *Recursive High-Resolution Reconstruction of Blurred Multiframe Images*. IEEE Transactions on Image Processing, Vol. 2, No. 4 (1993), pp. 534-539.
- [70] G. D. Knott. *Interpolating Cubic Splines*. Birkhäuser, 2000.
- [71] R. Kress. *Numerical Analysis*. Springer, 1998.
- [72] E. Kreyszig. *Introductory Functional Analysis with Applications*. John Wiley & Sons, Inc., 1978.
- [73] R. L. Lagendijk, J. Biemond. *Iterative Identification and Restoration of Images*. New York: Kluwer, 1991.
- [74] C. L. Lawson, R. J. Hanson. *Solving Least Squares Problems*. SIAM, 1995.
- [75] A.J. Laub. *Matrix Analysis for Scientists and Engineers*. SIAM, 2005.
- [76] A.J. Linz. *A new numerical method for ill-posed problems*. Inverse Problems 10 (1994) pp. L1-L6.
- [77] D. W. Marquardt. *Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation*. Technometrics Vol. 12, No. 3 (1970), pp. 591-612.
- [78] May'za Vladimir, Shaposhnikova Tatyana. *Jacques Hadamard: A universal Mathematician*. AMS, 1998.
- [79] G. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2000.
- [80] P. Milanfar, *Super-Resolution Imaging*. Taylor & Francis Group LLC, 2011.

- 
- [81] A. Mohammand. *Inverse Problems in Vision and 3D Tomography*. ISTE Ltd, John Wiley & Sons, Inc. 2010.
- [82] J. Montalvo. *Solución de problemas de mínimos cuadrados a gran escala por métodos de gradientes conjugados*. Tesis de Licenciatura en Matemáticas Aplicadas, Universidad Autónoma de Coahuila, México, 2000.
- [83] H. Montegranario, J. Espinosa. *Variational Regularization of 3D Data: Experiments with MATLAB*. Springer, 2014.
- [84] J. J. Moré, D. C. Sorensen. *Computing a trust region step*. SIAM, J. Sci. Stat. Comput. Vol. 4, No. 3 (1983), pp. 553-572.
- [85] V. A. Morozov. *On the solution of functional equations by the method of regularization*. Soviet Math Dokl. 7 (1966), pp. 414-417.
- [86] R. H. Myers. *Classical and Modern Regression with Applications*. Duxbury Press, 1994.
- [87] D. Nagarajan, V. Nagarajan, P. Sunitha, V. Seethalckshmi. *Analysis of Birth weight using Singular Value Decomposition* International Journal of Computer Science and Information Security, Vol. 7 (2010), No. 1.
- [88] J. G. Nagy, D. P. O’Leary. *Restoring images degraded by spatially-variant blur*. SIAM, J. Sci. Comput. 19 (1998), pp. 1063-1082.
- [89] H. Nasir, V. Stankovic, S. Marshall. *Singular value decomposition based fusion for super-resolution image reconstruction*. IEEE International Conference on Signal and Image Processing Applications(2011), pp. 393-398.
- [90] M. K. Ng, R. H. Chan, W. Tang. *A Fast Algorithm for Deblurring Models with Neumann Boundary Conditions*. SIAM, J. Sci. Comput. Vol. 21, No. 3 (1999), pp. 851-866.
- [91] N. Nguyen, P. Milanfar, G. Golub. *A Computationally Efficient Superresolution Image Reconstruction Algorithm*. IEEE Transaction on Image Processing, Vol. 10, No. 4 (2001), pp. 573-583.
- [92] N. Nguyen, P. Milanfar, G. Golub. *Efficient Generalized Cross-Validation with Applications to Parametric Image Restoration and Resolution Enhancement*. IEEE Transaction on Image Processing, Vol. 10, No. 9 (2001), pp. 1299-1308.
- [93] D. P. O’Leary. *Scientific Computing with Case Studies*. SIAM, 2008.
- [94] F O’Sullivan. *A statistical Perspective of ill posed problems*. Stastical Science Vol. 1, No.4 (1986), pp. 502-527.

- 
- [95] S. C. Park, M. Y. Park, M. G. Kang. *Super-Resolution Image Reconstruction: A Technical Overview*. IEEE Signal Processing Magazine, 2003, pp. 21-36.
- [96] J. Pasupathy, R. A. Damodar. *The Gaussian Toeplitz Matrix* Linear Algebra and its Applications, Vol 171 (1992), pp. 133-147.
- [97] D. L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. J. Assoc. Comput. Mach., 9 (1962), pp 84-97.
- [98] J. O. Ramsay, G. Hooker, S. Graves. *Functional Data Analysis with R and MATLAB*. Springer, 2009.
- [99] M. Rojas. *Regularization of large scale ill-conditioned least squares problems*. 1996.
- [100] J. Romberg. *The SVD of a circulant matrix*, 2011.
- [101] B. W. Rust. *Truncating the singular value decomposition for ill-posed problems*. Report NISTIR 6131, Mathematical and Computational Sciences Division, NIST, 1998.
- [102] J. C. Santamarina, D. Fratta. *Discrete Signals and Inverse Problems: An introduction for Engineers and Scientists*. John Wiley & Sons Ltd, 2005.
- [103] S. A. Santos, D. C. Sorensen. *A new matrix-free algorithm for the large-scale trust region subproblem*. Technical Report 95-20, Department of Computational and Applied Mathematics, Rice university, 1995.
- [104] O.I. Sarajlic, A.B. Smirnova *Numerical representation of weirs using the concept of inverse problems* International Journal of Hydraulic Engineering 2013, 2(3) pp. 53-58.
- [105] O. Scherzer. *Mathematical Models for Registration and Applications to Medical Imaging*. Springer, 2006.
- [106] R. J. Schwarz, B. Friedland. *Linear Systems*. McGraw Hill, Inc, 1965.
- [107] J. L. Schiff. *The Laplace Transform: theory and applications*. Springer, 1999.
- [108] G. A. F. Seber, A. J. Lee. *Linear Regression Analysis*. John Wiley & Sons, Inc., Second Edition, 2003.
- [109] C. B. Shaw, Jr. *Improvement of the Resolution of an Instrument by Numerical Solution of an Integral Equation*. Journal of Mathematical Analysis and Applications 37, (1972) 83-112.
- [110] F. Smithies. *Integral Equations*. Cambridge University Press, 1958.

- [111] T. Steihaug. *The conjugate gradient method and trust regions in large scale optimization*. SIAM J. Numer. Anal., Vol. 20, No. 3 (1983), pp. 626-637.
- [112] G. W. Stewart. *Collinearity and Least Squares Regression*. Statistical Science Vol. 2. No.1 (1987), pp. 68-100.
- [113] A. N. Tikhonov. *The solution of ill-posed problems*. Doklady Akad. Nauk SSSR Vol. 151, No. 3 (1963).
- [114] A. N. Tikhonov. *Regularization of ill-posed problems*. Doklady Akad. Nauk SSSR Vol. 153, No. 1 (1963).
- [115] L.N. Trefethen, J .A C. Weideman. *The exponentially convergent trapezoidal rule*. SIAM Review, Vol. 56, No. 3 (2014), pp. 385-458.
- [116] R. Y. Tsai, T. S. Huang. *Multiframe image restoration and registration*. Advances in Computer Vision and Image Processing, Greenwich, CT: JAI Pres Inc. (1984), pp. 317-339
- [117] S. Twomey. *On the numerical solution od Fredholm integral equations of the first kind by the inversion of the Linear system produced by quadrature*. J. Assoc. Comput. Mach., 10 (1963), pp. 97-101.
- [118] U. S. Navy. *Procedures and Analysis for Staffing Standards Development: Data/Regression Analysis Handbook*. Navy Manpower and Material Analysis Center, San Diego, C. A., 1979.
- [119] P. Vandewalle, S. Süsstrunk, M. Vetterli. *A Frequency domain approach to registration of aliased images with application to super-resolution*. EURASIP Journal on Applied Signal Processing, Volume 2006, pp 1-14.
- [120] G. Wahba. *Numerical and Statistical Methods for Midly, Moderately and Severaly Ill-Posed Problems with Noisy Data*. Technical Report No. 595, Department of Statistics, University of Winsconsin (1980).
- [121] Y. Wang, A.G. Yagola, C. Yang. *Optimization and Regularization for Computational Inverse Problems and Applications* Springer, 2010.
- [122] D. Widder. *The Laplace Transform* Princeton University Press, 1946.
- [123] G.M. Wing, J.D. Zahrt. *A Primer on Integral Equations of the First Kind: The Problem of Deconvolution and Unfolding*. SIAM, 1991.
- [124] A. Wyn-jones. *Circulants*. New York, 2013.
- [125] L. A. Zadeh. *An Extended Definition of Linearity*. Proc. IRE, Vol. 50 (1962), p.200

- [126] A. H. Zemanian. *Distribution Theory and Transform Analysis: An Introduction to Generalized Functions and Applications*. Dover Publications, 1965.