UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

DISEÑO ÓPTIMO EN PROBLEMAS INVERSOS LINEALES

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN CIENCIAS

PRESENTA:
JOSÉ RODRIGO ROJO GARCÍA

DIRECTOR
DR. MARCOS AURELIO CAPISTRÁN OCAMPO
CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS A.C. (CIMAT)

CIUDAD DE MÉXICO NOVIEMBRE 2020.

Agradecimientos

Agradezco primeramente a mi asesor el Dr. Marcos Aurelio Capistrán Ocampo, por absolutamente todo lo que hizo por mi a lo largo de la maestría y sobre todo por la sabia dirección que tomo al dirigir esta tésis. También agradezco mucho la confianza que ha tenido para orientarme y aconsejarme sobre mis futuros estudios.

Al CIMAT por haberme brindado la oportunidad de cursar materias, darme un espacio de trabajo y permitirme usar el equipo de supercómputo. Todo esto ha sido una experiencia maravillosa que ha abierto mi panorama enormemente.

Al Dr. Pedro González Casanova Henríquez, pues siempre me ha dado buenos consejos y ánimos.

A mis sinodales, Dr. Eduardo Gutierrez Peña, Dra. Ursula Xiomara Iturrarán Viveros, Dr. Pedro Gonzalez Casanova Henríquez, Dr. Antonio Capella Kort y Dr. Marcos Aurelio Capistrán OCampo. Ya que me apoyaron con valiosas observaciones en este trabajo.

A mi familia y a mis amigos entrañables. Todos ellos me han apoyado de manera incondicional en estas aventuras escolares.

Siglas y Abreviaturas

BVP	Problema de Valores en la Frontera (por sus siglas en inglés *Boundary Value Problem*)
FEM	Método de Elemento Finito (por sus siglas en inglés *Finite Element Method*)
GMRES	Método del Residuo Mínimo Generalizado (en inglés *Generalized Minimal Residual*)
GPU	Cómputo paralelo vía Unidad de Procesamiento Gráfico (en inglés *Graphics Processing Unit*)
MCMC	Métodos Montecarlo basados en Cadenas de Markov (en inglés *Markov Chain Monte Carlo*)
ODE	Ecuación Diferencial Ordinaria (por sus siglas en inglés *Ordinary Differential Equation*)
OED	Diseño Óptimo Experimental (por sus siglas en inglés *Optimal Experimental Design*)
PDE	Ecuación Diferencial Parcial (por sus siglas en inglés *Partial Differential Equation*)
RWHM	Métodos de Caminatas Aleatorias Hasting-Metropolis (por sus siglas en inglés *Random Walk Hasting-Metropolis*)
SPD	Matriz Simétrica Positiva Definida (por sus siglas en inglés *Symmetric Positive Definite*)
SSPD	Matriz Simétrica Semipositiva Definida (por sus siglas en inglés *Symmetric Positive Semi-definite*)

Resumen

En este trabajo se implementaron diferentes técnicas relacionadas con problemas inversos bayesianos para la formulación de un diseño experimental óptimo (OED) en Ecuaciones Diferenciales Parciales. Dicho experimento tiene la característica de que el usuario puede elegir la cantidad mínima de nodos posibles para medir y de manera que dichos nodos aporten la mayor información posible a la hora de hacer inferencias de los parámetros.

Índice general

Índice de figuras

Capítulo 1

Introducción

Dentro de la ciencia, la ingeniería y las finanzas existen modelos matemáticos que dependen de parámetros, a las ecuaciones que gobiernan dichos modelos se les conoce como **problema directo** o **mapeo directo**. Sin embargo, la reproducibilidad de dichos modelos dependen del conocimiento de los parámetros involucrados; es entonces que la estimación de estos parámetros se convierte en el objeto de estudio el cual se denomina **problema inverso** (para varios ejemplos puede consultar [9], [18], [19]). A su vez, los problemas inversos necesitan de mediciones hechas con aparatos de alta precisión en laboratorios o en posiciones geográficas específicas.

En la mayoría de casos estas mediciones están hechas por personal calificado y/o por aparatos cuyo funcionamiento puede ser en sí costoso, por lo que la adquisición de datos puede ser inviable si se desean muchas mediciones. Es entonces cuando se requiere formular un modelo matemático para un experimento que contenga sólo las mediciones que sean indispensables; esto se conoce como diseño experimental óptimo.

Existen varios enfoques para la solución de problemas inversos; en nuestro caso nos enfocamos en el paradigma bayesiano, el cual consiste en hacer la inferencia del parámetro m asociándole una distribución de probabilidad, la cual llamaremos **distribución posterior**. Dicha distribución se construye a través de una distribución a priori de m por medio de información previa, y de mediciones englobadas en una función llamada **verosimilitud**. En particular nos enfocamos en modelos gaussianos y aproximaciones de rango bajo para el modelado de la distribución posterior.

Una vez aplicadas dichas herramientas para construir una distribución posterior en algunos problemas de PDEs, se procedió a su implementación en la formulación del diseño óptimo.

1.1. Trabajos previos

El enfoque usado en este trabajo es que cuando se trabajen problemas inversos donde el espacio parametral está en un espacio de Hilbert con dimensión infinita, entonces se use una distribución a priori formulada en dicho espacio y después se discretice ésta.

Un trabajo pionero es el hecho por [9] en probemas inversos bidimensionales de mecánica de fluidos. Después se generalizó esta idea para varios problemas en [36], así mismo en dicho trabajo se pudo probar existencia y estabilidad en la solución al problema inverso.

Para la discretización numérica de la distribución a priori tenemos el trabajo de [6] el cual a su vez está basado en la tesis doctoral [34]. En estos trabajos se puede construir el operador de covarianzas y la media de la distribución a priori gaussiana por medio del método de elemento finito.

En el caso de la distribución posterior basada en aproximaciones de rango bajo tenemos los trabajos de [11] y [35]. El primero de ellos se basa en hacer una proximación por medio de descomposición SVD truncada, mientras que el segundo usa eigenvalores generalizados además de dar una proximación que es óptima con la distancia de Hellinger, la de Förstner y la Divergencia de Kullback-Leibler respecto a la posterior de rango completo.

Respecto a la parte de diseño óptimo experimental basado en la minimización de la traza, uno de los primeros compendios del área se encuentra en [27]; sin embargo, dicho trabajo no está enfocado a problemas inversos como tal. Una referencia de diseño óptimo en problemas inversos usando un enfoque diferente es [24]. En nuestro caso, este trabajo usa como marco de referencia a [2], el cual se basa en la minimización de la traza. Cabe mencionar que un trabajo muy reciente basado en la minimización de la traza es el dado por [20]; sin embargo, dicha investigación necesita del desarrollo de más herramientas que no son estudiadas en esta tesis.

1.2. Objetivos

El objetivo de este trabajo es modificar la parte de la regularización en la metodología de diseño óptimo propuesta en [2], aquí la parte de la regularización se sustituye por una secuencia de regularizaciones LASSO. Una vez hecho esto, otro objetivo es analizar los resultados de manera sucesiva con respecto a datos sintéticos para tres estudios de caso.

1.3. Contribuciones

En el trabajo de [2] se usó una regularización LASSO en la que se propone de manera arbitraria un coeficiente de regularización alto. Esto seguido de una regularización basada en funciones convexas y suaves a trozos, lo anterior con la finalidad de discriminar los nodos que no se consideran informativos.

Nuestra propuesta de regularización LASSO sucesiva genera una discriminación de los nodos de manera suave. Así mismo esta metodología reduce la dimensión del problema, evitando un mal condicionamiento del problema y guardando la información secuencialmente para un análisis más detallado.

1.4. Limitaciones y trabajo futuro

Las limitaciones en este trabajo se resumen primeramente en la capacidad de cómputo que se tenía para los estudios de caso propuestos, ya que el cálculo del diseño óptimo es bastante costoso computacionalmente. Esto a pesar de que se usó un cluster con aproximadamente 50 nodos para los cálculos.

En segundo lugar, tenemos que esta metodología para diseñar un problema OED está restringida a problemas inversos cuyo mapeo directo es un operador lineal.

Algunas propuestas de problemas a futuro son las siguientes:

1. Problemas inversos no lineales.
2. Dominios con obstáculos.
3. Problemas con datos reales.
4. Paralelización masiva vía GPUs.
5. Reformulación de la metodología con otros métodos numéricos para PDEs.

1.5. Contenido

El capítulo 2 se divide en cuatro secciones, la sección 2.1 hace referencia al sustento teórico de las PDEs que se usaron en este trabajo. La sección 2.2 consiste en un resumen sobre el método de elemento finito para ecuaciones elípticas y parabólicas, es aquí donde es explora

de manera numérica la formulación del problema directo. En el caso de la sección 2.3 ésta está completamente enfocada al estudio de problemas inversos, donde se hace una breve introducción a éstos, a la estadística bayesiana y al paradigma bayesiano como tal. Así mismo se menciona teoría específica en la solución a este problema, como es el caso de la formulación en espacios de Hilbert con dimensión infinita, la construcción de la distribución a priori y de la distribución posterior. El capítulo se cierra con la sección 2.4, la cual aborda el tema de diseño óptimo y en la que podemos encontrar desde su concepción hasta el desarrollo de las técnicas usadas para su implementación.

En el caso del capítulo 3, la metodología se divide en tres secciones. La sección 3.1 aborda la forma en la que rescribe la discretización numérica hecha en la sección de FEM. Mientras que las secciones 3.2 y 3.3 específican todos los parámetros usados y la paquetería de cómputo respectivamente.

Los resultados son presentados en el capítulo 4, en donde se hizo una sección para cada estudio de caso. La conclusión de estos resultados está dada en el capítulo 5 .

Finalmente, se escribieron dos apéndices. El apéndice A es referente a análisis funcional y está enfocado en espacios de Hilbert y Espacios de Sobolev para complementar la teoría de PDEs tratada en la sección 2.1. Por otro lado, el apéndice B aborda el tema de probabilidad en espacios de Hilbert con dimensión infinita, aquí se complementa la teoría abordada en la subsección 2.3.4.

Capítulo 2

Marco teórico

2.1. Ecuaciones diferenciales parciales

En esta sección estudiaremos algunas de las propiedades teóricas de las ecuaciones usadas en los estudios de caso mencionados previamente. Analizaremos brevemente los casos en los que estos problemas obedecen a un problema bien planteado en el sentido de Hadamard.

2.1.1. Preliminares

En el ámbito de la modelación, generalmente surgen problemas de matemáticas cuya solución en principio puede ser desconocida. Lo que se esperaría de este tipo de problemas es dar una solución ya sea analítica o númerica con las herramientas tanto teóricas como computacionales que se tengan a la mano. Algunos de los problemas con los que nos podemos enfrentar son que la solución no exista, o que existan múltiples soluciones, o bien que dicha solución sea sensible a pequeños cambios en el estado inicial del problema; esto último imposibilita la reproducibilidad de un experimento.

Hadamard [25] se vió en la necesidad de replantear la forma en la que se debe estudiar un problema antes de preguntarse por una solución, por lo que dió la siguiente definición.

Definición 2.1.1. Un problema está *bien planteado* si cumple con las siguientes condiciones:

1. La solución existe.
2. La solución es única.

3. Es Lipschitz continua con respecto a los datos iniciales.

En nuestro caso nos interesan ecuaciones diferenciales parciales, en específico del tipo elíptico y parabólico.

2.1.2. Operadores elípticos

Las ecuaciones que nos interesan son PDEs las cuales están formuladas por operadores diferenciales que pueden estar definidas en términos de derivadas débiles. En nuetro caso nos interesan las que llamaremos elípticas y las cuales definiremos a continuación.

Definición 2.1.2. Sea $\Omega \subset \mathbb{R}^n$ acotado, abierto y conexo. Sean también las funciones medibles $a_{i,j}, b_i, c_i \in L^\infty(\Omega)$. Una ecuación diferencial parcial elíptica es aquella que es de la forma

$$-\sum_{i,j=1}^{n} \partial_{x_i}(a_{ij}(\mathbf{x})u_{x_j}) + \sum_{i=1}^{n} \partial_{x_i}(b_i(\mathbf{x})u) + \sum_{i=1}^{n} c_i(\mathbf{x})u_{x_i} + a_0(\mathbf{x})u = f(\mathbf{x}), \tag{2.1}$$

y además cumple con la propiedad

$$\sum_{i,j=1}^{n} a_{ij}(\mathbf{x})\xi_i\xi_j > 0, \qquad \forall \mathbf{x} \in \Omega, \xi \in \mathbb{R}^n, \xi \neq 0.$$

El término asociado a los coeficientes a_{ij} puede interpretarse físicamente como *difusión*, mientras que en el caso de los coeficientes b_i, c_i, éstos determinan *transporte* y finalmente el término asociado a a_0 es una componente de *reacción*. Al término f se le conoce como la *fuente* de la ecuación.

Esta misma ecuación se puede escribir en forma de divergencia como

$$-\nabla \cdot (\mathbf{A}(\mathbf{x})\nabla u) + \nabla \cdot (\mathbf{b}(\mathbf{x})u) + \mathbf{c}(\mathbf{x}) \cdot \nabla u + a_0(\mathbf{x})u = f(\mathbf{x}),$$

de la cual empezaremos a definir algunas condiciones de regularidad que son necesarias para plantear el problema.

Definición 2.1.3. Sea $\mathcal{L}$ el operador diferencial asociado a la ecuación 2.1 sobre un dominio Ω acotado, entonces este operador es **uniformemente elíptico** si existen constantes $M_1, M_2 > 0$ tales que

$$M_1|\xi|^2 \leq \mathbf{A}(\mathbf{x})\xi \cdot \xi \leq M_2|\xi|^2.$$

Notemos que en un operador uniformemente elíptico los eigenvalores de $\mathbf{A}$ son estrictamente positivos, por lo que en el caso de matrices diagonales basta con que sus elementos de la diagonal sean funciones estrictamente positivas y acotadas en Ω.

En nuestro caso nos enfocaremos en ecuaciones con término de transporte nulo y con condición de frontera nula

$$\begin{cases} -\nabla \cdot (\mathbf{A}(\mathbf{x})\nabla u) + a_0(\mathbf{x})u = f(\mathbf{x}), & \mathbf{x} \in \Omega \\ u = 0, & \mathbf{x} \in \partial\Omega \end{cases} \tag{2.2}$$

También buscaremos soluciones débiles, las cuales son aquellas que pertenecen al espacio de Sobolev $H_0^1(\Omega)$. Por lo que el problema 2.2 se puede reformular en este sentido como encontrar $u \in H_0^1(\Omega)$ en la forma variacional 2.3

$$\int_\Omega (\mathbf{A}\nabla u \cdot \nabla v + a_0 uv)\, d\mathbf{x} = \int_\Omega f v d\mathbf{x}, \qquad v \in H_0^1(\Omega). \tag{2.3}$$

A partir de ahora es necesario dar las hipótesis para formular un problema bien planteado, por lo que el siguiente teorema establece el problema.

Teorema 2.1.1. *Sea $\mathcal{L}$ el operador diferencial asociado al problema 2.2 tal que este es uniformemente elíptico y además se satisfacen las siguientes hipótesis:*

1. *Ω es un dominio Lipschitz.*
2. *$a_0(\mathbf{x}) \geq 0$, c.s. en Ω*
3. *$f \in L^2(\Omega)$.*

Entonces existe solución única en $H_0^1(\Omega)$ y existe una constante M_0 que se satisface la siguiente cota de estabilidad

$$\|u\|_{H_0^1(\Omega)} \leq \frac{1}{M_0}\|f\|_{L^2(\Omega)}$$

Demostración. Puede consultar la prueba en [33]. □

Adicionalmente es posible dar algunas condiciones de regularidad para la solución por medio del siguiente teorema.

Teorema 2.1.2. *Sea $\mathcal{L}$ el operador diferencial con las hipótesis del teorema anterior y además cumple que:*

1. *Ω es poligonal convexo o tiene frontera suave.*
2. *$a_{i,j}(\mathbf{x}), a_0(\mathbf{x}) \in C^1(\overline{\Omega})$.*

Entonces, $u \in H^2(\Omega) \cap H_0^1(\Omega)$.

Demostración. Puede consultar la prueba en [29]. □

2.1.3. Ecuación de calor o difusión

Consideremos el problema dado por la Ecuación de Calor también conocida como Ecuación de Difusión, la cual es de la forma

$$\begin{aligned} \frac{\partial u}{\partial t} - A\nabla^2 u = 0, \quad & (\mathbf{x}, t) \in \Omega \times (0, T) \\ u(\mathbf{x}, 0) = g(\mathbf{x}), \quad & \mathbf{x} \in \overline{\Omega} \\ u(\mathbf{x}, t) = 0, \quad & (\mathbf{x}, t) \in \partial\Omega \times [0, T]. \end{aligned} \tag{2.4}$$

En forma débil este problema puede escribirse como

$$\int_\Omega u_t(\mathbf{x}, t) v(\mathbf{x}) d\mathbf{x} + A \int_\Omega \nabla u(\mathbf{x}, t) \cdot \nabla v(\mathbf{x}) d\mathbf{x} = 0, \qquad \forall t \in [0, T],$$

con $u(\mathbf{x}, 0) = g(\mathbf{x})$ y donde $v \in H_0^1(\Omega)$.

Dicha ecuación representa un problema bien planteado y puede probarse por medio del siguiente teorema.

Teorema 2.1.3. *Sea Ω un dominio poligonal convexo o bien con frontera $\partial\Omega \in C^2$ y sea $g(\mathbf{x}) \in H_0^1(\Omega) \cap H^2(\Omega)$, entonces el problema 2.4 tiene solución única y se cumple lo siguiente:*

1. *$u \in L^2(0, T; H^2(\Omega)) \cap H^1(0, T; L^2(\Omega)) \cap C^0([0, T]; H_0^1(\Omega) \cap H^2(\Omega))$.*
2. *Existe $C(A) > 0$ independiente de t tal que*

$$\max_{t \in [0,T]} \|u(\mathbf{x}, t)\|_{H^1(\Omega)}^2 + \int_0^T \left(\left\| \frac{\partial u}{\partial t}(\mathbf{x}, t) \right\|_{L^2(\Omega)}^2 + \|u(\mathbf{x}, t)\|_{H^2(\Omega)}^2 \right) dt \leq C(A) \|g\|_{H^1(\Omega)}^2$$

Demostración. Puede consultar la prueba en [29]. □

Esta ecuación como lo dice su nombre, fue modelada para estudiar la transmisión del calor a través de la temperatura. Otra interpretación física es la disipación de alguna sustancia (por ejemplo un contaminante) a través de un medio; a este tipo de procesos se les conoce como difusivos.

Teniendo en cuenta lo anterior, la ecuación de calor tiene como una de sus características más representativas, el ser una función decreciente en función del tiempo.

2.1.4. Estudio de casos

En esta sección presentaremos los estudio de casos asociados a los problemas inversos que abordaremos más adelante, el objetivo es ver si en principio el problema directo (el cual es encontrar a u en función de sus parámetros) es un problema bien planteado. Cabe mencionar que a partir de este momento solo consideraremos casos en los que la matriz $\mathbf{A}(\mathbf{x}) = A(\mathbf{x})\mathbf{I}$, por lo que usaremos la notación $A(\mathbf{x})$ en lugar de $\mathbf{A}(\mathbf{x})$ para referirnos a la acción $A(\mathbf{x})\nabla u = \mathbf{A}(\mathbf{x})\nabla u$.

Caso 1

Sea $n = 2$ y el dominio $\Omega = [0, 1] \times [0, 1]$. El primer problema es la ecuación de Poisson

$$\begin{aligned} -\nabla \cdot (A(\mathbf{x})\nabla u) &= f, \quad \mathbf{x} \in \Omega \\ u &= 0, \quad \mathbf{x} \in \partial\Omega, \end{aligned} \tag{2.5}$$

donde el parámetro de difusión queda dado por

$$A(\mathbf{x}) = \exp\left[-\left((x - 1/2)^2 + (y - 1/2)^2\right)\right],$$

y el término fuente es

$$\begin{aligned} f(\mathbf{x}) = {} & 2\pi^2 A(\mathbf{x}) \sin(\pi x) \sin(\pi y) + 2\pi(x - 1/2)A(x) \sin(\pi y) \cos(\pi x) \\ & + 2\pi(y - 1/2)A(\mathbf{x}) \sin(\pi x) \cos(\pi y). \end{aligned} \tag{2.6}$$

Es fácil verificar que el parámetro $A(\mathbf{x})$ define un operador uniformemente elíptico y la fuente es una función infinitamente suave en un conjunto compacto, por lo que $f \in L^2(\Omega)$. Con esto estamos probando que este estudio de caso asociado al problema directo es un problema bien planteado y su solución por tanto debe ser única.

Finalmente notemos que la función $u(\mathbf{x}) = \sin(\pi x) \sin(\pi y)$ satisface ser solución al problema Dirichlet 2.5. Dicha función la podemos observar en la figura 2.1.1.

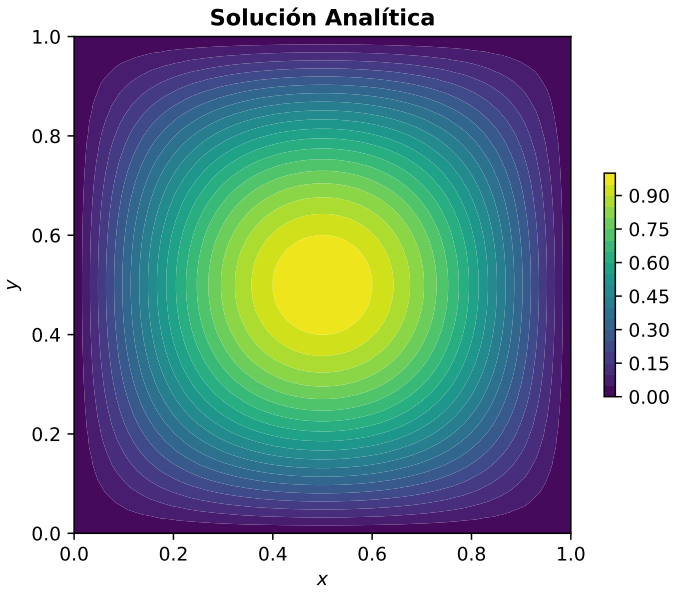

Figura 2.1.1: Solución analítica, estudio de caso 1.

Caso 2

Sea $n = 2$ con dominio $\Omega = [0, 1] \times [0, 1]$; este problema se conoce como ecuación de reacción-difusión estacionaria y es de la forma

$$\begin{aligned} -\nabla \cdot (A(\mathbf{x})\nabla u) + a_0(\mathbf{x})u &= f, \quad \mathbf{x} \in \Omega \\ u &= 0, \quad \mathbf{x} \in \partial\Omega, \end{aligned} \tag{2.7}$$

con los siguientes parámetros:

1. Difusión
$$A(\mathbf{x}) = \exp\left[-\left(x^2 + y^2\right)\right]$$

2. Reacción
$$a_0(\mathbf{x}) = 3.0.$$

3. Fuente
$$\begin{aligned} f(\mathbf{x}) = &\left[2\frac{1-2x}{1-x} + 2\frac{1-2y}{1-y}\right] A(\mathbf{x})v(\mathbf{x}) \\ &+ \left[\frac{1-p}{(1-x)^2} + \frac{1-p}{x^2} + \frac{2p}{x(1-x)}\right] A(\mathbf{x})v(\mathbf{x}) \\ &+ \left[\frac{1-p}{(1-y)^2} + \frac{1-p}{y^2} + \frac{2p}{y(1-y)}\right] A(\mathbf{x})v(\mathbf{x}) + a_0(\mathbf{x})v(\mathbf{x}), \end{aligned} \tag{2.8}$$

donde $p > 2$ y

$$v(\mathbf{x}) = x^p(1-x)^p y^p(1-y)^p. \tag{2.9}$$

Este problema define un problema elíptico con los operadores $A(\mathbf{x})$ y $a_0(\mathbf{x})$. Así mismo está claro que $f \in L^2(\Omega)$, todos estos parámetros satisfacen las hipótesis del teorema 2.1.2 por lo que el problema es bien planteado y su solución es: $u(\mathbf{x}) = v(\mathbf{x})$.

Usando $p = 10$ podemos observar la solución analítica en la figura 2.1.2.

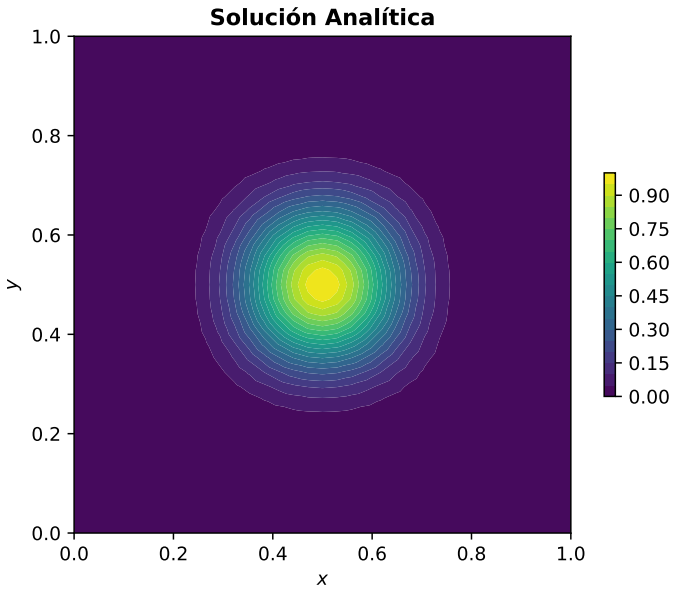

Figura 2.1.2: Solución analítica, estudio de caso 2.

Caso 3

Consideremos el dominio $\Omega = [0,1] \times [0,1]$ y el problema

$$\begin{aligned} \frac{\partial u}{\partial t} - A\nabla^2 u &= 0, \quad (\mathbf{x}, t) \in \Omega \times (0, T) \\ u(\mathbf{x}, 0) &= g(\mathbf{x}), \quad \mathbf{x} \in \overline{\Omega} \\ u(\mathbf{x}, t) &= 0, \quad (\mathbf{x}, t) \in \partial\Omega \times [0, T], \end{aligned} \tag{2.10}$$

con los parámetros $A = 0.1$, $T = 1$, y la condición inicial

$$g(\mathbf{x}) = 100 \sin(\pi x) \sin(\pi y).$$

Debido al teorema 2.1.3 se trata de un problema bien planteado y tiene como solución

$$u(\mathbf{x}, t) = 100 \sin(\pi x) \sin(\pi y) \exp(-2\pi^2 A t). \tag{2.11}$$

Dicha solución la podemos observar en la figura 2.1.3 para los tiempos $t = 0, 0.4$ y 0.8.

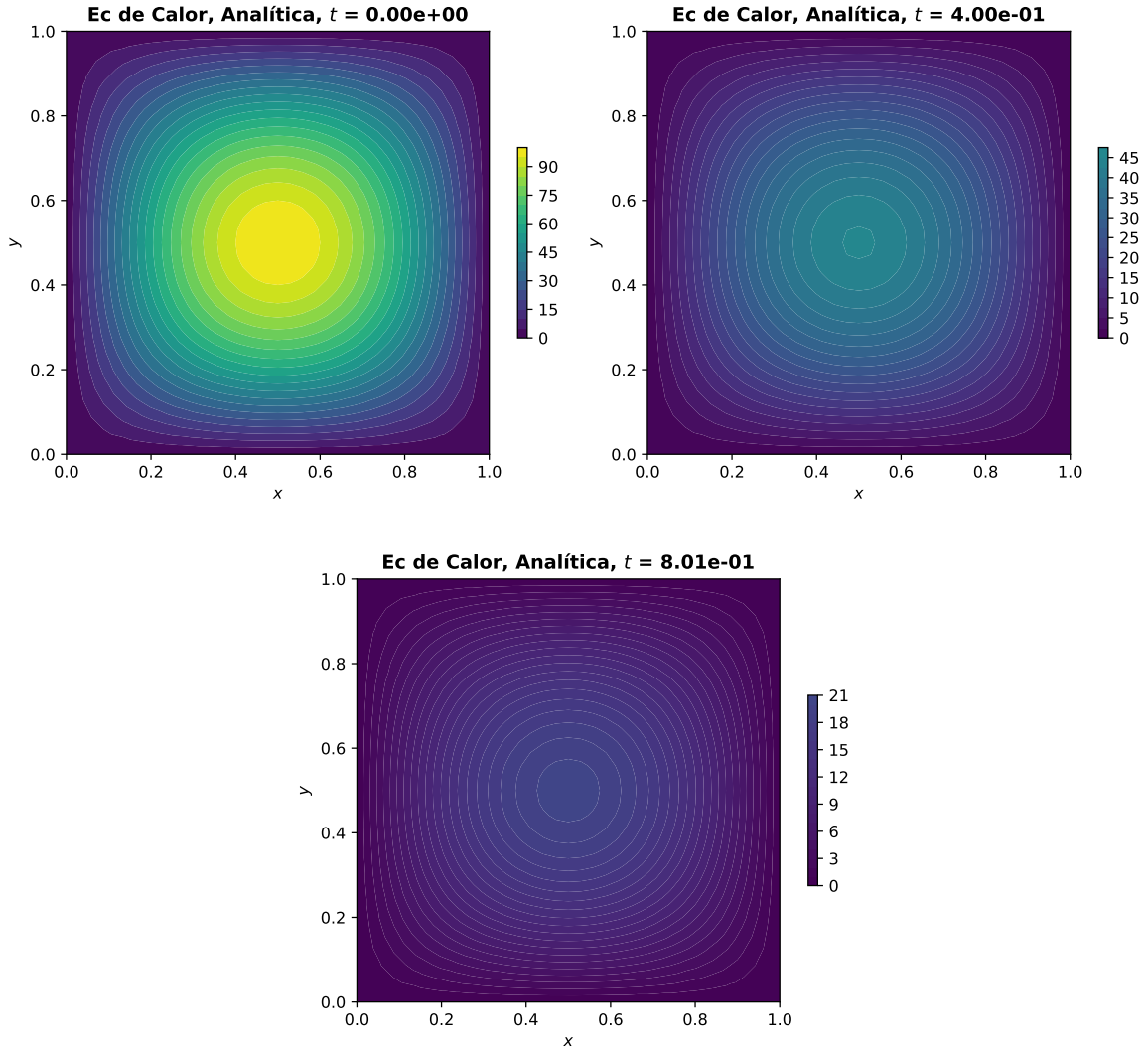

Figura 2.1.3: Solución analítica, estudio de caso 3.

.

2.2. Método de Elemento Finito

En este capítulo abordaremos el método de elemento finito, el cual funciona como método de discretización numérica para el mapeo directo 2.37. Estudiaremos su convergencia y algunas propiedades de la discretización en los estudios de caso mencionados. Gran parte de esta sección para los problemas elípticos está basada en [22], así mismo se recomienda consultar en la parte de referente a los problemas parabólicos a [23].

2.2.1. Justificación

En la solución de ecuaciones diferenciales parciales es común usar métodos numéricos, esto generalmente se debe a la imposibilidad de exhibir una solución en términos de expresiones algebraicas ya conocidas. Los métodos clásicos para resolver cada problema dependen del tipo de ecuación a estudiar, del dominio sobre el cual se quiere resolver y de la información disponible. Por ejemplo, los métodos de diferencias finitas otorgan excelentes resultados con ecuaciones elípticas y parabólicas, así mismo son baratos computacionalmente hablando, sin embargo, éstos pueden usarse solamente en nodos sobre una rejilla y no ofrecen buenos resultados en ecuaciones hiperbólicas no lineales.

Los métodos de elemento finito por su parte también ofrecen excelentes resultados en ecuaciones elípticas y parabólicas; sin embargo, requieren mayor costo computacional ya que necesitan crear mallas para su implementación. Su gran ventaja radica en que los nodos pueden escogerse de manera aleatoria salvo algunas excepciones. De igual manera, la amplia literatura que existe sobre FEM ha permitido desarrollar las herramientas de problemas inversos que después serán mencionadas.

2.2.2. Aproximación polinomial a trozos

Supongamos que tenemos un dominio $\Omega \subset \mathbb{R}^2$ con frontera Lipschitz y dentro de él un conjunto de nodos $\{\mathbf{x}_i\}_{i=1}^N$ sobre los cuales queremos aproximar una función $f(\mathbf{x})$. El objetivo es encontrar una función $S_f(\mathbf{x}) \approx f(\mathbf{x})$, los requisitos que pediremos de esta función es que sea continua y que se pueda expresar por medio de una base finita.

Una opción viable es usar polinomios en $\mathbb{R}^2$, el más sencillo de ellos es el de grado 1

$$p_1(\mathbf{x}) = a_0 + a_1 x + a_2 y,$$

pues tiene dimensión 3 y por tanto necesita a lo más 3 nodos no colineales para quedar determinado de manera única por sus coeficientes.

Para abordar este problema se puede descomponer este dominio en pequeños triángulos, sobre los cuales quedaría definida esta aproximación. Para que dicha aproximación funcione de manera adecuada, los nodos en nuestro dominio deben definir una triangulación, la cual definimos a continuación.

Definición 2.2.1. Denotemos por K_i al triángulo formado por tres nodos, diremos que los nodos $\{\mathbf{x}_i\}_{i=1}^N$ definen una triangulación en el dominio Ω si para N_K triángulos se satisface lo siguiente.

1. $\Omega = \bigcup_{i=1}^{N_K} K_i$.
2. La intersección de dos triangulos es una arista, un vértice o bien el conjunto vacío.

A dicha triangulación la denotaremos por $\mathcal{K}$.

Cabe mencionar que una triangulación aceptable es aquella que no tiene *triángulos degenerados*, los cuales se definen como aquellos en los que sus vértices tienden a ser colineales.

Dada una triangulación, el espacio sobre el cual definiremos nuestra base es el siguiente

$$V_h = \{v : v \in C^0(\Omega),\ v|_K \in P_1(K),\ \forall K \in \mathcal{K}\},$$

el cual se trata de funciones que son polinomiales de grado 1 en cada triángulo K. Así mismo guarda la propiedad de ser continua entre triángulos vecinos, es decir, si dos triángulos vecinos K_1 y K_2 con funciones v_1 y v_2 respectivamente están conectados por una arista o un nodo, entonces en dicha intersección ocurrira que $v_1 = v_2$.

La forma ideal de escoger una base $\{\phi_i\}_{i=1}^N$ para V_h es lo que popularmente se conoce en teoría de interpolación como una base de Lagrange, es decir, una de la forma

$$\phi_j(\mathbf{x}_i) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}, \quad i, j = 1, 2, \ldots, N.$$

De esta manera, toda función $v \in V_h$ se puede escribir simplemente como

$$v = \sum_{i=1}^{N} \alpha_i \phi_i; \qquad \alpha_i = v(\mathbf{x}_i). \tag{2.12}$$

Ahora, existen dos enfoques clásicos para construir la función S_f. El primero es por medio de una interpolación πf; la construcción de este interpolante es simple ya que se define como

$$\pi f = \sum_{i=1}^{N} f(\mathbf{x}_i)\phi_i.$$

El otro enfoque se basa en construir una función de aproximación, es decir, una función cuya distancia a la función f sea mínima en norma. La norma natural que correspondería a esta base $\{\phi_i\}_{i=1}^N$ es la correspondiente a $L^2(\Omega)$. Así mismo el teorema de proyecciones ortogonales A.1.1 nos asegura que la función de distancia mínima es la proyección ortogonal en este espacio. En el caso de $L^2(\Omega)$ recordemos que una proyección ortogonal sobre V_h queda dada por la siguiente definición.

Definición 2.2.2. Sea $f \in L^2(\Omega)$, definimos la proyección ortogonal de f sobre V_h como aquella función $P_h(f) \in V_h$ que cumple lo siguiente:

$$\langle f - P_h f, v\rangle_{L^2(\Omega)} = 0, \qquad \forall v \in V_h.$$

Dado que $P_h f \in V_h$, entonces esta función se puede expresar en términos de la base, es decir

$$P_h f = \sum_{i=1}^{N} \zeta_i \phi_i(\mathbf{x}),$$

por lo que nuestro objetivo se reduce a determinar los coeficientes de la base ζ_i.

Notemos ahora que cualquier elemento v también se puede escribir en términos de dichas base, por lo que la definición de proyección ortogonal es equivalente a

$$\langle f - P_h f, \phi_i\rangle_{L^2(\Omega)} = 0, \qquad \forall i = 1, 2, \cdots, N. \tag{2.13}$$

Reescribiendo a $P_h f$ en términos de la base y usando la equivalencia 2.13, encontramos que los coeficientes de la proyección ortogonal quedan determinados por el sistema lineal

$$\mathbf{M}\Upsilon = \mathbf{b},$$

donde

$$(\Upsilon)_i = \zeta_i, \qquad (\mathbf{M})_{ij} = \int_\Omega \phi_i \phi_j d\mathbf{x}, \qquad (\mathbf{b})_i = \int_\Omega f\phi_i d\mathbf{x}. \tag{2.14}$$

Debido a que desde que aparecieron los métodos de elemento finito, éstos han sido ampliamente usados en problemas de mecánica, existe una fuerte tradición de llamar a $\mathbf{M}$ y $\mathbf{b}$ como *matriz de masa* y *vector de carga* respectivamente.

En este sistema debemos resaltar algunas cosas, la primera es que la solución a este problema es única. Esto puede verse desde el Teorema de proyecciones ortogonales, donde la proyección está únicamente determinada. Por otro lado, dado que la base es polinomial a trozos, entonces las entradas de la matriz de masa pueden calcularse de manera analítica bajo ciertas manipulaciones algebraicas, situación que no sucede con las entradas del vector de carga. Para solucionar este último inconveniente, se opta por usar métodos de Newton-Cotes o alguna regla de cuadratura.

Refiriendonos a la solución numérica del sistema lineal y su estabilidad, podemos remarcar las siguientes propiedades.

Teorema 2.2.1. *La matriz de masa* M *satisface lo siguiente:*

1. *Es simétrica definida positiva.*
2. *Tiene número de condición acotado.*

Demostración. [22] □

2.2.3. Solución a ecuaciones elípticas y parabólicas

Empezaremos por formular un problema Dirichlet para la ecuación de Poisson y extender este resultado para el caso elíptico 2.2. Finalmnete se formulará un problema de evolución parabólico como uno estacionario elíptico por medio de métodos numéricos para EDO.

Ecuación de Poisson

El primer problema a resolver será el de la ecuación de Poisson con coeficiente de difusión variable

$$\begin{aligned} -\nabla \cdot (A(\mathbf{x})\nabla u) &= f, \quad \mathbf{x} \in \Omega \\ u &= 0, \quad \mathbf{x} \in \partial\Omega. \end{aligned} \tag{2.15}$$

Como mencionamos en la sección anterior 2.1, estamos interesados en buscar una solución débil. Si bien, dicha solución corresponde a $H_0^1(\Omega)$, al usar elemento finito tenemos que replantear el espacio sobre el cual quedaría dada nuestra solución.

Al aproximar una función por medio de una aproximación polinómica vimos que la solución era una proyección ortogonal sobre el espacio V_h. En este método, la solución numérica debe

mimetizar a la solución analítica, por lo que el espacio de solución debe ser compatible con $H_0^1(\Omega)$. Para verificar dicha compatibilidad se puede probar primero que $V_h \subset H^1(\Omega)$ [23].

Por otro lado, el problema Dirichlet exige usar una condición de frontera nula, lo cual obliga a los elementos de la base a satisfacer esta misma condición. Así el espacio solución conveniente es el definido como

$$V_h^0 := \{v \in V_h : v|_{\partial\Omega} = 0\},$$

y el cual verifica que $V_h^0 \subset H_0^1(\Omega)$ [23].

Una vez replanteado el espacio de interés, llamaremos u_h a la solución numérica, la cual mimetiza la forma variacional analítica de la siguiente manera. Elegimos a $v \in V_h^0$ y multiplicamos ambos lados de la ecuación por dicho término, una vez hecho esto integramos por partes sobre Ω y obtenemos la expresión

$$\int_\Omega A\nabla u_h \cdot \nabla v d\mathbf{x} - \int_{\partial\Omega} Av\nabla u_h \cdot \widehat{\eta} d\mathbf{x} = \int_\Omega fv d\mathbf{x}.$$

Utilizando el hecho de que v se anula sobre la frontera, entonces el problema se reducen a la ecuación

$$\int_\Omega A\nabla u_h \cdot \nabla v d\mathbf{x} = \int_\Omega fv d\mathbf{x}.$$

De igual manera que en el caso de aproximación polinómica, u_h se puede expresar por medio de la base de Lagrange $\{\phi_i\}_{i=1}^N$ asociada a V_h^0, es decir, es de la forma

$$u_h(\mathbf{x}) = \sum_{i=1}^{N} U_i \phi_i(\mathbf{x}).$$

Sustituyendo en términos de la base tenemos nuevamente el sistema lineal

$$\mathbf{KU} = \mathbf{b}, \tag{2.16}$$

donde

$$(\mathbf{U})_i = U_i, \qquad (\mathbf{K})_{ij} = \int_\Omega A(\mathbf{x})\nabla\phi_i \cdot \nabla\phi_j d\mathbf{x}; \tag{2.17}$$

la matriz $\mathbf{K}$ también aparece en un contexto de mecánica, por lo que regularmente es conocida como matriz de rigidez.

Al igual que en el caso de aproximación polinómica, las entradas de $\mathbf{b}$ se calculan de manera numérica con alguna cuadratura bidimensional o con métodos de Newton-Cotes. Para el caso de la matriz de rigidez, recordemos que en cada triángulo los elemento de la base se pueden expresar por medio de un polinomio lineal, es decir

$$\phi_i(\mathbf{x}) = a_0^i + a_1^i x + a_2^i y.$$

Los coeficientes de dicho polinomio varían dentro de cada triángulo, esto de manera que la función siga siendo una base de Lagrange y tenga un soporte compacto. De esta manera, los triángulos que están dentro del soporte tienen gradientes dados por la expresión constante

$$\nabla\phi_i(\mathbf{x}) \cdot \nabla\phi_j(\mathbf{x}) = a_1^i a_1^j + a_2^i a_2^j.$$

Así, la integral se reduce a una regla de cuadratura para el término $A(\mathbf{x})$ sobre cada triángulo K en el soporte de la función.

Nuevamente tenemos una solución basada en una proyección ortogonal, y por ende con solución única. Así mismo tenemos que la matriz de rigidez cumple con propiedades similares a la matriz de masa.

Teorema 2.2.2. *La matriz de rigidez* $\mathbf{K}$ *asociada al problema 2.15 es simétrica definida positiva.*

Demostración. Puede verificar la prueba en [23]. □

Ecuación reacción-difusión estacionaria

Consideremos ahora el caso concerniente a la ecuación de reacción difusión estacionaria

$$\begin{aligned} -\nabla \cdot (A(\mathbf{x})\nabla u) + a_0(\mathbf{x})u &= f, \quad \mathbf{x} \in \Omega \\ u &= 0, \quad \mathbf{x} \in \partial\Omega, \end{aligned} \tag{2.18}$$

con los parámetros $A(\mathbf{x})$ y $a_0(\mathbf{x})$ tales que satisfacen las hipótesis del teorema 2.1.1; más aún, pediremos que $a_0(\mathbf{x}) > C > 0$ para alguna constate C. Al igual que en el caso de la Ecuación de Poisson 2.15, se trata de un problema bien planteado con solución única y estable en $H_0^1(\Omega)$ por el teorema 2.1.1.

Al igual que en el caso anterior, basta con proponer una solución $u_h \in V_h^0$. Así, al multiplicar la PDE del problema 2.18 por $v \in V_h^0$, tenemos la forma

$$\int_\Omega A\nabla u_h \cdot \nabla v d\mathbf{x} + \int_\Omega a_0 u_h v d\mathbf{x} - \int_{\partial\Omega} Av\nabla u_h \cdot \widehat{\eta} d\mathbf{x} = \int_\Omega f v d\mathbf{x}. \tag{2.19}$$

Por lo que al aplicar condiciones de frontera sobre la función v en la expresión 2.19, ésta se reduce a

$$\int_\Omega A\nabla u_h \cdot \nabla v d\mathbf{x} + \int_\Omega a_0 u_h v d\mathbf{x} = \int_\Omega f v d\mathbf{x}. \tag{2.20}$$

Sabemos que $u_h = \sum_{i=1}^N U_i \phi_i(\mathbf{x})$, por lo que la ecuación 2.20 puede reformularse en términos de la base de V_h^0. Para esto notemos que el primer término del lado izquierdo corresponde a la matriz de rigidez 2.17 en la ecuación de Poisson. Mientras que el otro término es muy similar al dado por la matriz de masa 2.14 salvo el término $a_0(\mathbf{x})$.

La solución queda entonces establecida por el nuevo sistema lineal

$$(\mathbf{K} + \mathbf{M}_{a_0})\mathbf{U} = \mathbf{b}, \tag{2.21}$$

donde

$$(\mathbf{M}_{a_0})_{ij} = \int_\Omega a_0 \phi_i \phi_j d\mathbf{x}, \tag{2.22}$$

Teorema 2.2.3. *Sea $\mathbf{M}_{a_0}$ la matriz de masa asociada a la integral*

$$(\mathbf{M}_{a_0})_{ij} = \int_\Omega a_0 \phi_i \phi_j d\mathbf{x}, \tag{2.23}$$

entonces M_{a_0} es SPD.

Demostración. Sea $\mathbf{z} \in \mathbb{R}^N$ no nulo, entonces

$$\begin{aligned}
\mathbf{z}^T \mathbf{M}_{a_0} \mathbf{z} &= \sum_{j=1}^N \sum_{i=1}^N z_i (\mathbf{M}_{a_0})_{ij} z_j \\
&= \sum_{j=1}^N \sum_{i=1}^N z_i \left(\int_\Omega a_0 \phi_i \phi_j d\mathbf{x} \right) z_j \\
&> \sum_{j=1}^N \sum_{i=1}^N z_i \left(\int_\Omega C \phi_i \phi_j d\mathbf{x} \right) z_j \\
&= C \sum_{j=1}^N \sum_{i=1}^N z_i \left(\int_\Omega \phi_i \phi_j d\mathbf{x} \right) z_j \\
&= C\mathbf{z}^T \mathbf{M}\mathbf{z} > 0.
\end{aligned}$$

Dado que la matriz de masa $\mathbf{M}$ es SPD, se tiene la última desigualdad de lado derecho.

□

Usando el hecho de que la matriz de masa y de rigidez son simétricas definidas positivas, es posible aseverar la misma propiedad para la matriz involucrada en el sistema lineal 2.21

Corolario 2.2.4. *Sea la matriz de rigidez* $\mathbf{K}_{a_0} := \mathbf{K} + \mathbf{M}_{a_0}$*, entonces es SPD.*

Demostración. Sea el vector $\mathbf{z} \in \mathbb{R}^N$ no nulo, al usar el hecho de que $\mathbf{K}$ y $\mathbf{M}_{a_0}$ son SPD tenemos que

$$\begin{aligned}\mathbf{z}^T(\mathbf{K}_{a_0})\mathbf{z} &= \mathbf{z}^T(\mathbf{K} + \mathbf{M}_{a_0})\mathbf{z} \\ &= \mathbf{z}^T\mathbf{K}\mathbf{z} + \mathbf{z}^T\mathbf{M}_{a_0}\mathbf{z} > 0.\end{aligned}$$

□

Ecuación de difusión o calor

Veamos ahora el caso de la ecuación de calor con condición de frontera Dirichlet

$$\begin{aligned}\frac{\partial u}{\partial t} - A\nabla^2 u = 0, \quad (\mathbf{x}, t) \in \Omega \times (0, T) \\ u(\mathbf{x}, 0) = g(\mathbf{x}), \quad \mathbf{x} \in \overline{\Omega} \\ u(\mathbf{x}, t) = 0, \quad (\mathbf{x}, t) \in \partial\Omega \times [0, T],\end{aligned} \tag{2.24}$$

donde el parámetro de difusión es una constante $A > 0$. Dado que este problema está bien planteado, es posible hacer una semidiscretización basada en la discretización de la parte temporal. A esto comúnmente se le conoce como método de líneas.

Para poder aplicar el método de líneas, empezaremos por definir el operador $\mathfrak{L}$ asociado a la parte elíptica de la ecuación como

$$\mathfrak{L}u := A\nabla^2 u.$$

Con lo anterior, la ecuación se puede escribir como

$$\frac{\partial u}{\partial t} = \mathfrak{L}u, \tag{2.25}$$

donde la estructura de los operadores diferenciales es análoga a la de un problema de valor inicial para una ODE. A su vez, esto conlleva al uso de métodos numéricos para ODEs.

Una propuesta adecuada para este problema es el método llamado regla trapezoidal, a veces conocida en el contexto de diferencias finitas como Crank-Nicolson. Éste consiste en usar la regla de integración trapezoidal sobre el lado derecho de la expresión 2.25. Para ello consideremos la partición del tiempo $t^k = k\Delta t \in [0, T]$ con $k = 0, 1, 2, \ldots, N_t$.

Entonces el esquema queda dado por

$$u^k(\mathbf{x}) - u^{k-1}(\mathbf{x}) = \frac{\Delta t}{2}\left(\mathfrak{L}u^k(\mathbf{x}) + \mathfrak{L}u^{k-1}(\mathbf{x})\right). \tag{2.26}$$

La primera razón para elegir este método es que es consistente, lo que quiere decir que mimetiza el operador diferencial temporal de manera adecuada, además su error local de truncamiento es del orden

$$|u^k - u(t^k)| = \mathcal{O}(\Delta t^2), \tag{2.27}$$

el cual es superior a las dos versiones de Euler clásicas. Otro beneficio es que este método califica en la categoría de A-estable, es decir, que su región de estabilidad es todo el semiplano complejo izquierdo, o equivalentemente, que cualquier tamaño de paso Δt asegura estabilidad lineal. Estos resultados pueden consultarse en [15].

Cabe mencionar que la A-estabilidad depende de la estructura de la ODE; para asegurarla es necesario que todos los eigenvalores del operador $\mathfrak{L}$ sean negativos. Esto es fácil de probar ya que la discretización espacial de dicho operador depende de la matriz de rigidez. Para probar la afirmación anterior, definamos $u_h^k \in V_h^0$ la solución por elemento finito en cada paso de tiempo k.

El esquema trapezoidal para la solución u_h^k es entonces

$$u_h^k(\mathbf{x}) - u_h^{k-1}(\mathbf{x}) = \frac{A\Delta t}{2}\left(\nabla^2 u_h^k(\mathbf{x}) + \nabla^2 u_h^{k-1}(\mathbf{x})\right), \tag{2.28}$$

y a su vez esto puede reescribirse

$$-\frac{A\Delta t}{2}\nabla^2 u_h^k(\mathbf{x}) + u_h^k(\mathbf{x}) = \frac{A\Delta t}{2}\nabla^2 u_h^{k-1}(\mathbf{x}) + u_h^{k-1}(\mathbf{x}). \tag{2.29}$$

Notemos primero que esta formulación se ha reducido a la solución con elemento finito del problema elíptico 2.18. Multiplicando la igualdad 2.29 por la función arbitraria $\phi \in V_h^0$ e integramos por partes, tenemos que

$$\frac{A\Delta t}{2}\int_\Omega \nabla u_h^k \cdot \nabla v d\mathbf{x} + \int_\Omega u_h^k v d\mathbf{x} = -\frac{A\Delta t}{2}\int_\Omega \nabla u_h^{k-1} \cdot \nabla v d\mathbf{x} + \int_\Omega u_h^{k-1} v d\mathbf{x}. \tag{2.30}$$

Así la discretización total de la ecuación de calor con elemento finito, está dada por la expresión 2.31

$$\left(\mathbf{M} + \frac{A\Delta t}{2}\mathbf{K}\right)\mathbf{U}^k = \left(\mathbf{M} - \frac{A\Delta t}{2}\mathbf{K}\right)\mathbf{U}^{k-1}. \tag{2.31}$$

Es claro que la matriz asociada a este problema es un caso particular de la matriz involucrada en la ecuación de reación difusión estacionaria, por lo que podemos garantizar que tenemos una solución numérica bien definida para cada paso k del tiempo.

2.2.4. Solución numérica a sistemas lineales

Como ya hemos visto, un problema de proyección ortogonal como de solución a PDEs elípticas o parabólicas lineales lo podemos resumir a la solución de un sistema líneal algebraico. Tanto las matrices de masa como rigidez cumplen con la peculiaridad de ser simétricas definidas positivas, más aún, el soporte compacto de la base asegura que son matrices ralas.

Debido a las propiedades anteriores, de manera clásica muchos autores como por ejemplo [3], [23], [22] y [32] han sugerido que el sistema lineal involucrado se resuelva con métodos del Subespacio de Kylov como Gradiente Conjugado, GMRES, BIGSTAB, entre otros. También se ha recomendado que el almacenamiento de las matrices se haga de manera rala; existen diferentes formatos para ello como el CSR, CSC, MSR, etc, los cuales reducen sustancialmente la memoria involucrada como las operaciones asociadas a cada método. Para consultar más sobre estos formatos puede consultar por ejemplo a [3] y [32].

Si bien nunca se recomienda calcular matrices inversas en la solución a sistemas linales de este tipo, más adelante veremos que en el estudio de los problemas inversos generalmente necesitamos calcular operadores de la forma

$$\mathbf{R} = \mathbf{E}^{-1}\mathbf{F}, \tag{2.32}$$

con las matrices estudiadas previamente.

Una forma de hacer esto eficientemente es por medio de métodos directos (basados en factorización) con almacenamiento ralo; estos métodos se basan en hacer una descomposición LU o de Cholesky de $\mathbf{R}$ tomando en cuenta las pocas operaciones y almacenamiento de una matriz en formato ralo. Una vez hecho esto, se calcula la solución a los sistemas lineales

$$\mathbf{E}\mathbf{R}_i = \mathbf{F}_i, \qquad i = 1, 2, \ldots, N_K, \tag{2.33}$$

por medio de sustitución hacia atrás y hacia adelante. En este caso $\mathbf{F}_i$ es la i-ésima columna de $\mathbf{F}$, así mismo la solución $\mathbf{R}_i$ resulta ser la i-ésima columna de $\mathbf{R}$. Con esto, el proceso de mayor complejidad es la factorización y ésta sólo se calcula una vez.

La librería de *Scipy* en *python3* conocida como *Sparse linear algebra* contiene la función *spsolve*, con la cual se puede calcular el operador $\mathbf{R}$ en 2.32 de manera eficiente. Para más detalles puede consultar [Sci].

2.2.5. Estimación del error

Cerraremos esta sección relativa al método de elmento finito mencionando algunas estimaciones de error conocidas para las funciones lineales a trozos. Comenzaremos por remarcar que en este contexto se prefiere el uso de la norma de energía para funciones en H_0^1 definida como A.2

Una vez hecha la observación anterior, tenemos las siguientes cotas de error.

Teorema 2.2.5. *Sean los problemas elípticos 2.15 y 2.18 con las hipótesis ya mencionadas en cada caso, entonces se tienen las siguientes estimaciones en:*

1. *Norma de energía*
$$\|u - u_h\|_{H_0^1(\Omega)} \leq Ch\|u\|_{H^2(\Omega)}. \tag{2.34}$$
2. *Norma $L^2(\Omega)$*
$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^2\|u\|_{H^2(\Omega)}. \tag{2.35}$$
3. *Norma $L^\infty(\Omega)$*
$$\|u - u_h\|_{L^\infty(\Omega)} \leq Ch\|u\|_{H_0^2(\Omega)}. \tag{2.36}$$

Demostración. Las pruebas pueden encontrarse en [29]. □

2.3. Problemas inversos

En este capítulo desarrollaremos la parte correspondiente a la teoría de los problemas inversos, haremos una breve descripción del planteamiento bayesiano tanto en dimensión finita como infinita. Así mismo estudiaremos técnicas relativamente recientes para atacar este problema en dimensión infinita y su correcta mimetización a dimensión finita.

2.3.1. Definición de problema directo e inverso

Supongamos que tenemos un modelo matemático determinista o estocástico que depende de uno o varios parámetros, los cuales englobaremos en la variable m. Si dicho parámetro ya se supone conocido desde un principio, entonces tenemos lo que llamamos un **problema directo**, y queda dado por el mapeo

$$u = \mathcal{G}(m), \tag{2.37}$$

mientras que a la estimación de m, conocidas mediciones de u se le llama **problema inverso**.

A continuación mencionaremos algunos ejemplos de problemas inversos:

1. Ecuación de van der Pol.

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= \mu\left(1 - x^2\right) y - x \\ x(0) &= x_0, \quad y(0) = y_0. \end{aligned} \tag{2.38}$$

   Esta ecuación es de gran importancia en la modelación de circuitos eléctricos [38]; el problema inverso consiste en la estimación del parámetro $m := \mu$ a partir de las observaciones

$$u := \begin{pmatrix} x \\ y \end{pmatrix}.$$

2. Ecuaciones de la elastodinámica
   Consideremos el dominio acotado y Lipschitz $\Omega \times (0, T)$, sobre dicho dominio las ecuaciones

$$
\begin{aligned}
\frac{\partial v_x}{\partial t} &= b(x,z)\left(\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xz}}{\partial z}\right) \\
\frac{\partial v_z}{\partial t} &= b(x,z)\left(\frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{zz}}{\partial z}\right) \\
\frac{\partial \tau_{xx}}{\partial t} &= (\lambda+2\mu)\frac{\partial v_x}{\partial x} + \lambda\frac{\partial v_z}{\partial z} \\
\frac{\partial \tau_{zz}}{\partial t} &= (\lambda+2\mu)\frac{\partial v_z}{\partial z} + \lambda\frac{\partial v_x}{\partial x} \\
\frac{\partial \tau_{xz}}{\partial t} &= \mu\left(\frac{\partial v_x}{\partial z} + \frac{\partial v_z}{\partial x}\right),
\end{aligned}
\tag{2.39}
$$

con condición inicial

$$(v_x, v_z, \tau_{xx}, \tau_{zz}, \tau_{xz})\,(\mathbf{x}, 0) = u_0(\mathbf{x}), \qquad \mathbf{x} \in \overline{\Omega},$$

y cuya condición de frontera es

$$(v_x, v_z, \tau_{xx}, \tau_{zz}, \tau_{xz})\,(\mathbf{x}, t) = g(\mathbf{x}, t), \qquad (\mathbf{x}, t) \in \overline{\partial\Omega} \times [0, T].$$

Se trata de un sistema de ecuaciones hiperbólicas que rigen la propagación de las ondas P-SV, es usado ampliamente en la modelación sísmica [39]. El problema inverso para un medio homogéneo consiste en la estimación de los coeficientes λ y μ conocidos como constantes de Lamé.

Notemos que el problema involucra más de un coeficiente, por lo que el problema se considera de dimensión dos y el parámetro por estimar es

$$m := \begin{pmatrix} \lambda \\ \mu \end{pmatrix}$$

dadas las observaciones

$$u := \begin{pmatrix} v_x \\ v_z \\ \tau_{xx} \\ \tau_{zz} \\ \tau_{xz} \end{pmatrix}.$$

3. Ecuación de calor en medios heterogéneos.

   Asumamos que queremos modelar la propagación del calor sobre un medio heterogéneo, en el dominio Lipschitz y acotado $\Omega \times (0, T)$. Entonces como ya hemos visto previamente,

las ecuaciones que modelan este problema son

$$\begin{aligned} \frac{\partial u}{\partial t} - \nabla \cdot (A(\mathbf{x})\nabla u) &= f(x), \quad (\mathbf{x}, t) \in \Omega \times (0, T) \\ u(\mathbf{x}, 0) &= g(\mathbf{x}), \quad \mathbf{x} \in \overline{\Omega} \\ u(\mathbf{x}, t) &= p(\mathbf{x}, t), \quad (\mathbf{x}, t) \in \partial\Omega \times [0, T]. \end{aligned} \tag{2.40}$$

Como parámetro a modelar en un problema inverso podríamos considerar alguno de los siguientes casos:

a) El coeficiente de difusión $A(\mathbf{x})$.

b) El términio fuente $f(\mathbf{x})$.

c) La condición inicial $g(\mathbf{x})$.

d) La condición de frontera $p(\mathbf{x}, t)$.

Notemos que para cada uno de los casos, el parámetro m es una función y en general el espacio al que pertenece puede tener dimensión infinita. En estos casos, obtener una estimación de una función suele ser mucho más complicado de tratar que los problemas finito dimensionales; así mismo computacionalmente muchas veces incluyen un gran reto en tiempo de cómputo. Más adelante veremos una técnica especial para este tipo de problemas.

2.3.2. Problema inverso no determinista

Cabe mencionar que los problemas inversos generalmente son problemas mal planteados en el sentido de Hadamard 2.1.1. Existen ejemplos de problemas inversos en dimensión infinita bien planteados (generalmente asociados a PDEs); sin embargo, las hipótesis de regularidad son demasiado restrictivas (para varios ejemplos puede consultar Isakov [18]).

Otro inconveniente es que, en el mundo real las mediciones de la variable de observación u siempre contienen ruido. Dicho ruido puede estar asociado a los intrumentos de medición, variables ambientales e incluso manipulación humana de los instrumentos. Este ruido genera una incertidumbre tanto de las observaciones como del parámetro m.

Existen diferentes formas de modelar el ruido, la más sencilla es suponer que el ruido es una variable aleatoria η con una distribución de probabilidad. La primer suposición que haremos para nuestra modelación es que tenemos **ruido aditivo**, es decir, las observaciones son de la forma

$$\tilde{u} = \mathcal{G}(m) + \eta. \tag{2.41}$$

Así mismo su esperanza tiene que ser nula, es decir,

$$\mathbb{E}(\eta) = 0.$$

A pesar de que los problemas inversos son mal planteados y están contaminados con ruido, se han estudiado técnicas para dar una solución numérica a estos problemas.

Uno de los enfoques clásicos consiste en regularizar estos problemas con algoritmos de optimización clásica. Por ejemplo, si nuestro mapeo es lineal, entonces es posible usar: mínimos cuadrados, descomposición SVD truncada, regularización de Tikhonov, iteración de Landweber-Fridman, etc. Mientras que en en los no lineales tenemos aquellos basados en algoritmos iterativos como los métodos Newton y Quasi-Newton, entre otros (para más detalles puede consultar [19]).

Otro enfoque del cual hablaremos a profundidad aquí, es considerar a todas las variables involucradas como variables aleatorias. De esta manera el problema inverso puede atacarse desde diferentes enfoques de la estadística, en particular nos enfocaremos en el paradigma bayesiano, el cual aprovecha la información previa sobre el parámetro a estimar.

Finalmente, mencionaremos que las medidas de probabilidad sobre las cuales trabajaremos se harán sobre espacios de medida no discretos y definen distribuciones continuas.

2.3.3. Elementos de Estadística Bayesiana

Supongamos que tenemos la variable aleatoria X definida sobre la tripleta $(\mathcal{X}, \mathfrak{X}, \mathbb{P})$, donde $\mathcal{X}$ es un espacio de medida, $\mathfrak{X}$ una σ-álgebra y $\mathbb{P}$ medida de probabilidad, estos dos últimos definidos sobre $\mathcal{X}$. Si X depende de uno o varios parámetros θ, entonces el objetivo al que se reducen varias teorías estadísticas es hacer inferencias sobrel el valor desconocido de θ dada información de las realizaciones de X.

La estadística bayesiana parte de la premisa de que al existir incertidumbre en el parámetro, se puede postular una distribución inicial de manera subjetiva para éste. Por tanto, debe existir una tripleta $(\Theta, \mathfrak{O}, \mathbb{H})$ sobre la cual la distribución está definida. Dicha distribución debe contener la información que se tenga a la mano del parámetro por estimar y se le conoce como *distribución a priori*. Adicionalmente supondremos que existe una densidad asociada a dicha distribución y la denotaremos por $\pi_{pr}(\theta)$.

De esta manera se puede definir una variable aleatoria (x, θ) sobre la σ-álgebra $\mathfrak{X} \times \mathfrak{O}$, así mismo se puede definir la densidad conjunta $\pi(x, \theta)$. Ahora, por definición de densidad condicional, se tiene que la densidad conjunta puede expresarse como

$$\pi(x, \theta) = \pi(x|\theta)\pi_{pr}(\theta),$$

donde a $\pi(x|\theta)$ modela las realizaciones de X, dada información del parámetro θ.

En un modelo bayesiano siempre se tiene información previa pero no suficiente del parámetro, de igual manera se cuenta con realizaciones en función del parámetro, por lo que la distribución a priori y la verosimilitud son información inicial. El objetivo entonces se reduce a buscar una distribución condicional de la forma $\pi_{post}(\theta|x)$, a esta distribución se le conoce como *distribución posterior*.

Teorema 2.3.1. *La distribución posterior puede calcularse como*

$$\pi_{post}(\theta|x) = \frac{\pi(x|\theta)\pi_{pr}(\theta)}{\int_\Theta \pi(x|\theta')\pi_{pr}(\theta')d\theta'}.$$

Demostración. Por el Teorema de Bayes [36] las distribuciones conjuntas quedan relacionadas por la expresión

$$\pi_{post}(\theta|x) = \frac{\pi(x|\theta)\pi_{pr}(\theta)}{\pi(x)},$$

pero la marginalización de la verosimilitud nos lleva a la densidad inicial de x, es decir

$$\pi(x) = \int_\Theta \pi(x|\theta')\pi_{pr}(\theta')d\theta'.$$

□

Es común dentro de la estadística bayesiana usar la siguiente notación para manejar la distribución posterior

$$\pi_{post}(\theta|x) \propto \pi(x|\theta)\pi_{pr}(\theta),$$

esto en primer lugar por comodidad; por otro lado, la constante de normalización dada por la integral puede no ser sencilla de calcular, e incluso puede ser imposible dar una expresión analítica en términos de funciones elementales. A continuación daremos algunos ejemplos del cálculo de una distribución posterior.

Ejemplo 1

Supongamos que tenemos la variable aleatoria $X \sim \mathcal{N}(\theta, \sigma^2)$ con σ^2 parámetro conocido. Si

tenemos realizaciones $\{x_i\}_{i=1}^d$ *i.i.d.* de X, entonces la verosimilitud queda dada por la expresión

$$\begin{aligned}\pi(x|\theta) &\propto \prod_{i=1}^{d} \exp\left(-\frac{1}{2\sigma^2}\left(x_i-\theta\right)^2\right)\\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{d}\left(x_i-\theta\right)^2\right),\\ &\propto \exp\left(-\frac{d}{2\sigma^2}(\bar{x}-\theta)^2\right)\end{aligned}$$

donde $\bar{x} = \frac{1}{d}\sum_{i=1}^{d} x_i$.

Si proponemos la distribución a priori asociada a θ como normal $\mathcal{N}(\rho, \tau^2)$, donde ρ y τ^2 son parámetros conocidos, entonces la distribución posterior queda determinada como

$$\begin{aligned}\pi_{post}(\theta|x) &\propto \pi(x|\theta)\pi_{pr}(\theta)\\ &\propto \exp\left(-\frac{d}{2\sigma^2}(\bar{x}-\theta)^2\right)\exp\left(-\frac{1}{2\tau^2}(\theta-\rho)^2\right)\\ &\propto \exp\left(-\frac{d}{2\sigma^2}(\bar{x}-\theta)^2\right)\exp\left(-\frac{1}{2\tau^2}(\theta-\rho)^2\right)\\ &\propto \exp\left(-\frac{1}{2}\left(\frac{\sigma^2\tau^2}{d\tau^2+\sigma^2}\right)^{-1}\left[\frac{\sigma^2\tau^2}{d\tau^2+\sigma^2}\left(\frac{\rho}{\tau^2}+\frac{d\bar{x}}{\sigma^2}\right)-\theta\right]^2\right).\end{aligned}$$

Por la estructura de la expresión anterior, puede deducirse que la distribución posterior es gaussiana y entonces la distribución asociada a la posterior es de la forma

$$\theta|x \sim \mathcal{N}\left(\frac{\sigma^2\tau^2}{d\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{d\tau^2+\sigma^2}\left(\frac{\rho}{\tau^2}+\frac{d\bar{x}}{\sigma^2}\right)\right).$$

Notemos que en este caso no fue necesario hacer el cálculo de la constante de normalización, sin embargo, existen casos en los que la constante no puede deducirse de esta forma y tampoco es fácil de calcular.

Ejemplo 2

Sea $X \sim Gamma(\alpha, \beta)$ con ambos parámetros desconocidos, supongamos que se tienen realizaciones $\{x_i\}_{i=1}^d$. Si suponemos que los parámetros pueden modelarse por medio de una distribución normal no truncada $\theta := (\alpha, \beta) \sim \mathcal{N}(\mu, \sigma^2 I)\, \mathbb{1}_{\{\alpha\geq 0,\, \beta\geq 0\}}$ con parámetro $\mu := (\mu_1, \mu_2)$ (de componenetes positivas) y σ^2 conocidos, entonces la verosimilitud queda determinada por

la expresión

$$\begin{aligned}\pi(x|\theta) &= \prod_{i=1}^{d} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x_i^{\alpha-1} \exp\left(-\beta x_i\right) \mathbb{1}_{\{\prod_{i=1}^{d} x_i > 0\}} \\ &= \frac{\beta^{\alpha d}}{\Gamma^d(\alpha)} \left(\prod_{i=1}^{d} x_i\right)^{\alpha-1} \exp\left(-\beta \sum_{i=1}^{d} x_i\right) \mathbb{1}_{\{\prod_{i=1}^{d} x_i > 0\}},\end{aligned}$$

mientras que la distribución a priori es de la forma

$$\pi_{pr}(\theta) \propto \exp\left(-\frac{1}{2\sigma^2}\left[(\alpha-\mu_1)^2 + (\beta-\mu_2)^2\right]\right) \mathbb{1}_{\{\alpha \geq 0,\, \beta \geq 0\}}.$$

Con esto la distribución posterior se puede expresar como

$$\pi_{post}(\theta|x) \propto \frac{\beta^{\alpha d}}{\Gamma^d(\alpha)} \left(\prod_{i=1}^{d} x_i\right)^{\alpha-1} e^{\left(-\beta \sum_{i=1}^{d} x_i - \frac{1}{2\sigma^2}\left[(\alpha-\mu_1)^2 + (\beta-\mu_2)^2\right]\right)} \mathbb{1}_{\{\prod_{i=1}^{d} x_i > 0,\, \alpha \geq 0,\, \beta \geq 0\}}.$$

Queda claro que conocer la constante de normalización es bastante complicado integrando de manera convencional; tampoco se puede asociar la expresión a una distribución ya conocida. Una metodología bastante usada en estos casos consiste en obtener una muestra de dicha distribución; esto es posible con métodos MCMC. Dicha muestra puede quedar representada por un histograma, así mismo ésta contiene de manera implícita información sobre la media y la varianza de la distribución (para más detalles [19]).

De manera concreta, supongamos que los verdaderos parámetros son $\alpha = 7.2$ y $\beta = 1.3$, con los cuales se generan $d = 50$ realizaciones. Por otro lado, los parámetros de la distribución a priori son $\mu = (6.5, 2.0)$ y $\sigma = 1.2$. Usando el método RWHM perteneciente a la familia de métodos MCMC, tenemos los histogramas dados en la figura 2.3.1, los cuales tienen una muestra de tamaño 90000 cada uno.

Notemos que los histogramas se encuentran centrados en valores cercanos a los verdaderos parámetros; sin embargo, el grosor de cada uno varía. Lo anterior nos da una idea de la incertidumbre asociada a cada uno de los parámetros.

Si bien el obtener una muestra de la distribución posterior puede ser una técnica suficiente, esto suele ser computacionalmente costoso y factible solo cuando la dimensión del espacio parametral no es muy grande.

Ejemplo 3

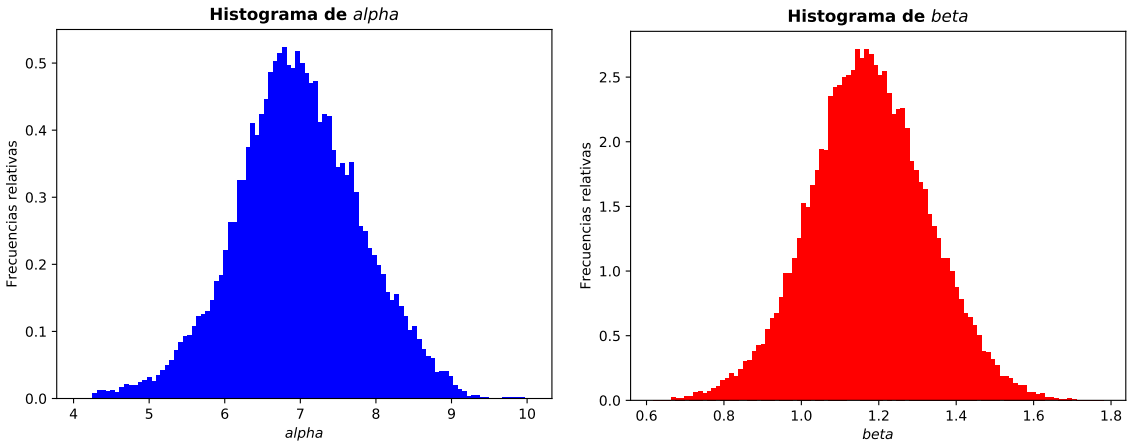

Figura 2.3.1: Histogramas de la distribución posterior

.

Consideremos ahora el modelo lineal de la forma

$$\mathbf{x} = \mathbf{G}\theta + \varepsilon, \tag{2.42}$$

donde $\mathbf{x} \in \mathbb{R}^n$, $\theta \in \mathbb{R}^m$ y $\mathbf{G} \in \mathbb{R}^{n\times m}$. Así mismo, tenemos ruido distribuido como $\varepsilon \sim \mathcal{N}(\mathbf{0}, \Gamma_{obs})$ con Γ_{obs} una matriz SPD.

Supongamos ahora que la distribución a priori del parámetro θ también es gaussiana y es de la forma $\mathcal{N}(\theta_0, \Gamma_{pr})$. El siguiente teorema nos garantiza que la distribución posterior es gaussiana y está bien definida con una fórmula explícita.

Teorema 2.3.2. *Sea el modelo lineal 2.42 con ruido $\varepsilon \sim \mathcal{N}(\mathbf{0}, \Gamma_{obs})$ y con distribución a priori $\theta \sim \mathcal{N}(\theta_0, \Gamma_{pr})$, donde Γ_{obs} y Γ_{pr} son matrices simétricas definidas positivas. Entonces la distribución posterior es gaussiana y es de la forma*

$$\pi_{post}(\theta|\mathbf{x}) = \mathcal{N}(\theta_{pos}, \Gamma_{pos}), \tag{2.43}$$

con parámetros

$$\Gamma_{post}^{-1} = \mathbf{G}^T\Gamma_{obs}^{-1}\mathbf{G} + \Gamma_{pr}^{-1}, \qquad \theta_{post} = \Gamma_{pos}(\mathbf{G}^T\Gamma_{obs}^{-1}\mathbf{x} + \Gamma_{pr}^{-1}\theta_0). \tag{2.44}$$

Demostración. Puede consultar la prueba en [36]. □

Para concluir este ejemplo, resaltaremos que a pesar de que tenemos una fórmula explícita para la distribución posterior, el cálculo de Γ_{post} y θ_{post} es bastante costoso cuando la dimensión

del espacio parametral m es muy grande, esto último debido a que se requiere invertir varias matrices. Por otro lado, el uso de MCMC también es inviable en dimensiones muy grandes; sin embargo, como veremos más adelante existen técnicas de aproximación de rango bajo que reducen de manera eficiente el costo computacional.

2.3.4. Planteamiento bayesiano en espacios de Hilbert con dimensión infinita

Recordemos que nuestro problema inverso de interés surge del problema directo 2.41. En este caso, nuestro parámetro de interés es m dada la información $\tilde{u}$.

Como ya mencionamos anteriormente, el paradigma bayesiano necesita información previa sobre nuestro parámetro m. Por ello, se puede usar información conocida o la opinión de un experto sobre el comportamiento de m; esto a su vez se traduce a definir una distribución con densidad $m \sim \pi_{pr}(m)$ la cual corresponde a la densidad de la distribución a priori de m.

Por otro lado, el parámetro es usado en modelos matemáticos definidos por el problema directo y con los cuales se obtiene información a partir de mediciones. De igual manera en el contexto bayesiano esta información queda reducida a $\tilde{u}|m \sim \pi(\tilde{u}|m)$. Esta densidad depende como tal de la distribución de η, por lo que si ϱ es la densidad de η, entonces

$$\pi(\tilde{u}|m) \propto \varrho(\tilde{u} - \mathcal{G}(m)). \tag{2.45}$$

Con esto, la estimación bayesiana del parámetro queda reducida una distribución posterior de la forma

$$\pi_{post}(m|\tilde{u}) \propto \pi_{pr}(m)\pi(\tilde{u}|m). \tag{2.46}$$

En dimensión finita la distribución posterior siempre queda bien definida por medio de la expresión 2.46. Algunos problemas inversos de este tipo son los ejemplos 2.38 y 2.39. Por otro lado, en el ejemplo 2.40 tenemos una serie de problemas inversos a plantear, todos en espacios de Hilbert de dimensión infinita.

Cuando la dimensión no es finita, podría no ser sencillo asociar una densidad de probabilidad para las distribuciones involucradas en el problema inverso. Por ello, es conveniente hablar de las medidas de probabilidad asociadas tanto a la a priori y la posterior, las cuales denotaremos respectivamente como $\mathbb{P}_{pr}$ y $\mathbb{P}_{post}$.

Sabemos que la medida de la posterior se define en términos tanto de la medida a priori como de la verosimilitud, en esencia están conectadas por medio del Teorema de Radon-Nikodym B.0.1, específicamente, la verosimilitud es la derivada de Radon-Nikodym $d\mathbb{P}_{post}/d\mathbb{P}_{pr}$.

Una vez aclarado lo anterior, es necesario decir que elegir un modelo para la medida a priori en dimensión infinita puede ser una tarea delicada, ya que podrían no cumplirse las hipótesis del Teorema de Radon-Nykodim B.0.1. Por ello, un paradigma para simplificar el problema de modelación, está sugerido por Stuart [36]. Por medio de este paradigma es posible:

1. Demostrar existencia a nuestro problema inverso bayesiano por medio de una distribución posterior.
2. Demostrar estabilidad en la solución respecto a los datos.
3. Dar estimaciones de error en la discretización numérica.

Existencia

Empezaremos por remarcar que el parámetro m y los datos $\tilde{u}$ viven en los espacios de Hilbert $(\mathcal{H}_M, \langle\cdot,\cdot\rangle_{\mathcal{H}_M})$ y $(\mathcal{H}_U, \langle\cdot,\cdot\rangle_{\mathcal{H}_U})$ respectivamente. En un problema inverso real, los datos $\tilde{u}$ no son más que q mediciones finitas contaminadas con ruido distribuido $\mathcal{N}(0, \mathbf{C}_{obs})$. Por lo que en este trabajo sólo nos enfocaremos en el caso donde $\mathcal{H}_U = \mathbb{R}^q$ con producto interior $\langle\cdot,\cdot\rangle_{\mathcal{H}_U} = \langle\cdot,\cdot\rangle_{\mathbf{C}_{obs}} = \langle \mathbf{C}_{obs}^{-1/2}\cdot, \mathbf{C}_{obs}^{-1/2}\cdot\rangle$; sin embargo, esta teoría se puede extender al caso donde $\mathcal{H}_U$ tiene dimensión infinita.

La primera suposición en este paradigma es que la distribución a priori es una medida gaussiana, es decir, es de la forma

$$\mathbb{P}_{pr} = \mathcal{N}(m_{pr}, \mathcal{C}_{pr}). \tag{2.47}$$

La segunda suposición es que la distribución posterior también debe ser Gaussiana y de la forma

$$\mathbb{P}_{post} = \mathcal{N}(m_{post}, \mathcal{C}_{post}). \tag{2.48}$$

Para que lo anterior sea posible, entonces la derivada de Radon-Nykodim tiene que estar bien definida. Recordemos que la verosimilitud está dada por la acción del operador $\mathcal{G}$ sobre m, por lo que empezaremos por definir las propiedades que debe cumplir este mapeo.

Es posible probar [36] que para que la posterior sea una medida Gaussiana, entonces existe la función potencial $\Phi : \mathcal{H}_M \times \mathcal{H}_U \to \mathbb{R}$ tal que

$$\frac{d\mathbb{P}_{post}}{d\mathbb{P}_{pr}} = \frac{1}{Z(\tilde{u})} \exp\left(-\Phi(m, \tilde{u})\right), \tag{2.49}$$

donde la constante de normalización $Z(\tilde{u})$ es

$$Z(\tilde{u}) = \int_{\mathcal{H}_M} \exp\left(-\Phi(m, \tilde{u})\right) d\mathbb{P}_{pr}(m). \tag{2.50}$$

En este sentido $d\mathbb{P}_{post}/d\mathbb{P}_{pr}$ está bien definida si dicha constante de normalización es acotada. Para probar esta afirmación empezaremos por decir que en nuestro contexto de problemas inversos, dicho potencial Φ es un operador de mínimos cuadrados, es decir,

$$\Phi(m, \tilde{u}) = \frac{1}{2}\|\tilde{u} - \mathcal{G}(m)\|^2_{\Gamma_{obs}}. \tag{2.51}$$

Por otro lado, daremos algunas hipótesis de regularidad sobre el operador $\mathcal{G}$ para que Φ esté bien definido.

Hipótesis 2.3.3. *El operador $\mathcal{G} : \mathcal{H}_M \to \mathcal{H}_U$ satisface las siguientes propiedades.*

1. *Dado $\varepsilon > 0$, existe $C = C(\varepsilon) \geq 0$ tal que, para todo $m \in \mathcal{H}_M$,*
$$\|\mathcal{G}(m)\|_{\mathcal{H}_U} \leq C \exp(\varepsilon \|m\|^2_{\mathcal{H}_M}).$$
2. *Para todo $r > 0$, existe $K = K(r) > 0$ tal que, para todo par $m_1, m_2 \in \mathcal{H}_M$ acotados por $\max\{\|m_1\|_{\mathcal{G}(m)}, \|m_2\|_{\mathcal{G}(m)}\} < r$,*
$$\|\mathcal{G}(m_1) - \mathcal{G}(m_2)\|_{\mathcal{H}_U} \leq K\|m_1 - m_2\|_{\mathcal{H}_M}.$$

Primero notemos que las hipótesis 2.3.3 prácticamente piden que el operador esté acotado y sea Lipschitz continuo, respecto a los parámetros. Como ejemplo tenemos que si $\mathcal{G}$ es un operador lineal continuo, entonces ambas propiedades se satisfacen automáticamente.

Con la hipótesis 2.3.3 ya es posible probar el siguiente teorema de existencia.

Teorema 2.3.4. *Sea $\mathbb{P}_{pr} = \mathcal{N}(m_{pr}, \mathcal{C}_{pr})$ de manera que $\mathbb{P}_{pr}(\mathcal{H}_M) = 1$, supongamos que la función potencial Φ satisface la hipótesis 2.3.3; entonces $Z(\tilde{u})$ es acotada y en consecuencia la medida de la posterior 2.48 está bien definida.*

Demostración. Se puede consultar la prueba en [36]. □

Ahora hablaremos de la estabilidad con respecto a los datos.

Estabilidad

Para poder hablar de estabilidad o bien de Lipschitz continuidad, iniciaremos por definir la métrica sobre la cual será definida. En este caso corresponde a una métrica que compara la distancia entre medidas de probabilidad (qué tanto podrían parecerse dos distribuciones). Esta distancia se conoce como distancia de Hellinger y se define a continuación.

Definición 2.3.1. Sean dos medidas de probabilidad $\mathbb{P}_1$ y $\mathbb{P}_2$ absolutamente continuas respecto a una medida $\mathbb{P}$, definimos la distancia de Hellinger entre ellas como

$$d_{\text{Hell}}\left(\mathbb{P}_1, \mathbb{P}_2\right) = \sqrt{\left(\frac{1}{2}\int\left(\sqrt{\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}}} - \sqrt{\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}}}\right)^2 \mathrm{d}\mathbb{P}\right)}. \tag{2.52}$$

Con esta distancia ya es posible definir estabilidad en la solución por medio del siguiente teorema.

Teorema 2.3.5. *Sea Φ una función potencial que satisface la hipótesis 2.3.3 y sean los datos $\tilde{u}, \tilde{u}' \in \mathcal{H}_U$ acotados en la forma $\max\{\|\tilde{u}\|_{\mathcal{H}_U}, \|\tilde{u}'\|_{\mathcal{H}_U}\} < r$. Supongamos que Φ define con $\tilde{u}$ y $\tilde{u}'$ las medidas posteriores $\mathbb{P}_{post}$ y $\mathbb{P}'_{post}$ como en el teorema 2.3.4, respectivamente. Entonces la medida posterior es Lipschitz continua respecto a los datos, es decir, existe $C(r) > 0$ tal que*

$$d_{Hell}\left(\mathbb{P}_{post}, \mathbb{P}'_{post}\right) \leqslant C\left\|\tilde{u} - \tilde{u}'\right\|_{\mathcal{H}_U}.$$

Demostración. Para ver la prueba consulte [36]. □

Discretización por metodos numéricos

Si bien el planteamiento de esta sección está hecha en espacios de Hilbert con dimensión infinita, en un problema real el mapeo $\mathcal{G}$ tiene que ser discretizado por métodos numéricos a una versión de dimensión finita $\mathcal{G}^N$. Esta versión debe de mimetizar de manera adecuada la solución continua en función del número de nodos N de la discretización, labor que está sustentada por la teoría del método numérico elegido.

Así, el mapeo $\mathcal{G}^N$ definirá una nueva función potencial Φ^N que a su vez esperamos construya una distribución posterior $\mathbb{P}^N_{post}$ con la distribución a priori original 2.47. Por lo que la nueva derivada de Radon-Nykodim de nuestro problema inverso es de la forma

$$\frac{d\mathbb{P}^N_{post}}{d\mathbb{P}_{pr}} = \frac{1}{Z^N(\tilde{u})}\exp\left(-\Phi^N(m, \tilde{u})\right).$$

Nota 1. La discretización del operador $\mathcal{G}^N$ podría no heredar las hipótesis 2.3.3 si $\mathcal{G}$ es no lineal.

Teorema 2.3.6. *Sean las medidas posteriores $\mathbb{P}_{post}$ y $\mathbb{P}^N_{post}$, absolutamente continuas respecto a $\mathbb{P}_{pr}$ por medio de Φ y Φ^N respectivamente. Supongamos ahora que para todo $\varepsilon > 0$ existe $K'(\varepsilon) > 0$ tal que*

$$\|\mathcal{G}(m) - \mathcal{G}^N(m)\| \leq K' \exp(\varepsilon\|m\|_{\mathcal{H}_m})\psi(N),$$

donde $\psi(N) \to 0$ cuando $N \to \infty$. Donde $\mathcal{G}$ y $\mathcal{G}^N$ cumplen la hipótesis 2.3.3 de manera uniforme respecto a N, entonces se cumplen las siguientes afirmaciones.

1. *Existe $C \geq 0$ (independiente de N) tal que*

$$d_{Hell}(\mathbb{P}_{post}, \mathbb{P}^N_{post}) \leq C\psi(N).$$

2. *La distancia entre las medias de las posteriores*

$$\|m_{post} - m^N_{post}\|_{\mathcal{H}_M} = \mathcal{O}(\psi(N)).$$

3. *Para todo $z \in \mathcal{H}_M$*

$$\|\mathcal{C}_{post}(z) - \mathcal{C}^N_{post}(z)\|_{\mathcal{H}_M} = \mathcal{O}(\psi(N)).$$

Demostración. Las pruebas a estas afirmaciones pueden ser consultadas en [36]. □

2.3.5. Distribución a priori

En la sección anterior dimos las herramientas necesarias para construir una distribución posterior que cumple dos propiedades del buen planteamiento de Hadamard 2.1.1. Sin embargo, la distribución a priori sigue siendo una medida gaussiana de dimensión infinita; en consecuencia nuestra distribución posterior también podría tenerla aunque exista una discretización $\mathcal{G}^N$.

Por otro lado, es necesario definir un operador de covarianzas coherente para nuestro problema inverso. En particular nosotros estamos interesados en problemas de ecuaciones diferenciales parciales sobre dominios Ω acotados y conexos, donde las observaciones y los parámetros se encuentran en espacios de Sobolev y algunos subespacios de $L^2(\Omega)$.

Un candidato para esto podría ser el operador $\mathcal{C}_{pr}^{-1} = -\Delta : H^2(\Omega) \bigcap H_0^1(\Omega) \to L^2(\Omega)$; sin embargo, este operador no es continuo [36]. Una forma de solucionar este problema es por medio de las proposiciones B.0.6 y B.0.7, con las cuales podemos construir a $\mathcal{C}_{pr}$ por medio del operador $\mathcal{A} := -\Delta + I$ con frontera nula. Así el operador de covarianzas a priori es de la forma

$$\mathcal{C}_{pr} := \mathcal{A}^{-\alpha}; \qquad \alpha > d/2. \tag{2.53}$$

Remarcaremos que este operador tiene la ventaja de ofrecer regularidad sobre elementos de la distribución a priori, es decir si $m \sim \mathbb{P} = \mathcal{N}(0, \mathcal{C}_{pr})$, entonces m es s-Hölder continua con $s < \min\{1, \alpha - d/2\}$ y además $m \in \mathcal{H}^s_M$. Para ello este operador debe satisfacer las hipótesis B.0.4, las cuales se heredan del operador $-\Delta$ (puede consultar [30] y [33]).

Ya que hemos concretado cuál es el operador de covarianzas que debemos usar, es necesario hablar sobre la correcta discretización de este operador. Como ya mencionamos anteriormente, usaremos FEM para el problema directo, por lo que el método numérico para discretizar $\mathcal{C}_{pr}$ también debe empatar con los nodos de discretización.

La discretización del operador $\mathcal{A}$ obtenida con el método de elemento finito es estándar y está dada por

$$\mathbf{A} := \mathbf{M}^{-1}\mathbf{K} + \mathbf{I},$$

donde $\mathbf{M}$ y $\mathbf{K}$ son las matrices de masa y rigidez respectivamente, con condición de frontera nula (para ver más detalles de esta discretización puede consultar [34]).

A partir de esta discretización, expondremos a continuación el enfoque propuesto en [6] para construir la versión fraccionaria del operador. Notemos primero que la matriz $\mathbf{A}$ se puede escribir de la siguiente manera

$$\mathbf{A} = \mathbf{M}^{-\frac{1}{2}}\left[\mathbf{M}^{-\frac{1}{2}}(\mathbf{K}+\mathbf{M})\mathbf{M}^{-\frac{1}{2}}\right]\mathbf{M}^{\frac{1}{2}}.$$

Esto indica que $\mathbf{A}$ es similar a una matriz SPD, en consecuencia es diagonalizable y tiene una descomposición espectral de la forma

$$\mathbf{A} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^{-1}.$$

Por lo que el operador fraccional queda dado entonces como

$$\mathbf{A}^{s/2} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^{-1}, \tag{2.54}$$

donde la matriz diagonal se define como $\mathbf{\Lambda} = \Sigma^{-s/2} := \text{diag}(\sigma_1^{-s/2}, \sigma_2^{-s/2}, \cdots, \sigma_N^{-s/2})$.

Con esto finalmente se puede dar una definición congruente del operador de covarianzas en dimensión finita.

Definición 2.3.2. Sea la discretización 2.54 obtenida con el método de elemento finito del operador $\mathcal{A}^{s/2}$, entonces definimos la matriz de covarianzas como

$$\mathbf{C}_{pr} := \frac{1}{\alpha}\mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T. \tag{2.55}$$

Más aún, es posible dar una expresión concreta para la matriz de precisión.

Teorema 2.3.7. *La matriz de precisión asociada a la matriz* $\mathbf{C}_{pr}$ *es*

$$\mathbf{C}_{pr}^{-1} = \alpha \mathbf{M}\mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^T\mathbf{M}. \tag{2.56}$$

Demostración. Consulte [6]. □

Necesitamos hacer algunos comentarios respecto a este operador de covarianzas y su precisión. El primer detalle es que es costoso computacionalmente, ya que implica la inversión de la matriz de masa. Así mismo es necesario calcular numéricamente la base de eigenvectores y eigenvalores. Sin embargo, el teorema 2.56 evita el cálculo de inversas de manera numérica.

Debemos ahora mencionar que la elección de la media debe encontrarse en el espacio de Cameron-Martin (revisar la definición B.0.6 y el teorema B.0.2 del apéndice B para más detalles). Si bien sabemos que $m_{pr} \in Im(\mathcal{C}_{pr}^{1/2})$, puede ser complicado caracterizar cualquier elemento que esté en este conjunto. Una opción es usar una función constante, ya que el dominio es un conjunto compacto y en consecuencia tenemos una media que es suave y se encuentra en cualquier subespacio de $L^2(\Omega)$. En particular $m_{pr} = 0$ es una opción aceptable en muchos problemas.

Finalmente la medida de la distribución a priori, tiene una discretización cuya medida a su vez tiene asociada una densidad gaussiana en dimensión finita y es de la forma

$$\mathbb{P}_{pr} \approx \mathcal{N}(\mathbf{m}_{pr}, \mathbf{C}_{pr}). \tag{2.57}$$

2.3.6. Distribución posterior

En este trabajo nos enfocaremos en el caso donde el operador $\mathcal{G}$ es lineal; dicha suposición es válida para una gran cantidad de problemas inversos, en particular los estudios de caso que hemos propuesto.

Recordemos primero que el problema directo está definido por el mapeo $\mathcal{G} : \mathcal{H}_M \to \mathbb{R}^q$. Por lo que al discretizar por métodos numéricos, el operador puede discretizarse en dimensión finita como $\mathbf{G} : \mathbb{R}^N \to \mathbb{R}^q$ y el modelo

$$\tilde{\mathbf{u}} = \mathbf{G}\mathbf{m} + \eta, \tag{2.58}$$

donde $\eta \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{obs})$.

Usando el paradigma de Stuart con la distribución a priori 2.57, el modelo 2.58 y el teorema 2.3.2, tiene distribución posterior

$$\pi_{post}(\mathbf{m}|\tilde{\mathbf{u}}) = \mathcal{N}(\mathbf{m}_{post}, \mathbf{C}_{post}), \tag{2.59}$$

con parámetros

$$\mathbf{C}_{post}^{-1} = \mathbf{H} + \mathbf{C}_{pr}^{-1}, \qquad \mathbf{m}_{post} = \mathbf{C}_{post}(\mathbf{G}^T\mathbf{C}_{obs}^{-1}\tilde{\mathbf{u}} + \mathbf{C}_{pr}^{-1}\mathbf{m}_{pr}) \tag{2.60}$$

y donde

$$\mathbf{H} = \mathbf{G}^T\mathbf{C}_{obs}^{-1}\mathbf{G}. \tag{2.61}$$

Como ya mencionamos en el mismo ejemplo, el cálculo de esta distribución posterior es muy costoso al tener que calcular varias inversas. Una forma de remediar este problema, es por medio del paradigma propuesto en [35], el cual consiste en hacer una aproximación de rango bajo para el operador de covarianzas y de precisión posterior.

La primer hipótesis en la modelación para usar esta técnica es que

$$\mathbf{m}_{pr} := 0.$$

Una forma de tomar ventaja a este problema es demostrar que la matriz de covarianzas posterior y la matriz de precisión se pueden escribir de la siguiente manera.

$$\mathbf{C}_{post} = \mathbf{C}_{pr} - \mathbf{K}\mathbf{K}^T, \qquad \mathbf{C}_{post}^{-1} = \mathbf{C}_{pr}^{-1} + \mathbf{Z}\mathbf{Z}^T.$$

En el caso de la precisión esto es fácil de probar, ya que $\mathbf{H}$ es SSPD por construcción, mientras que en el de la matriz de covarianzas, haremos uso de la identidad de Woodbury.

Teorema 2.3.8 (Sherman-Morrison-Woodbury)**.** *Supongamos que las matrices* $\mathbf{C}$*,* $\mathbf{D}$*,* $\mathbf{D}^{-1} + \mathbf{A}\mathbf{C}^{-1}\mathbf{B}$ *y* $\mathbf{C} + \mathbf{B}\mathbf{D}\mathbf{A}$ *son invertibles. Entonces tenemos la identidad*

$$(\mathbf{C} + \mathbf{B}\mathbf{D}\mathbf{A})^{-1} = \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{B}\left(\mathbf{D}^{-1} + \mathbf{A}\mathbf{C}^{-1}\mathbf{B}\right)^{-1}\mathbf{A}\mathbf{C}^{-1}.$$

Demostración. Puede revisar completa la prueba en [13]. □

Aplicando este teorema a la matriz de covarianzas tenemos el siguiente corolario.

Corolario 2.3.9. *Existe matriz* $\mathbf{K}$ *tal que*

$$\mathbf{C}_{post} = \mathbf{C}_{pr} - \mathbf{K}\mathbf{K}^T.$$

Demostración. Definiendo $\mathbf{A} := \mathbf{G}$, $\mathbf{B} := \mathbf{G}^T$, $\mathbf{C} := \mathbf{C}_{pr}^{-1}$ y $\mathbf{D} := \mathbf{C}_{obs}^{-1}$

$$(\mathbf{C}_{pr}^{-1} + \mathbf{H})^{-1} = \mathbf{C}_{pr} - \mathbf{C}_{pr}\mathbf{G}^T\left(\mathbf{C}_{obs} + \mathbf{G}\mathbf{C}_{pr}\mathbf{G}^T\right)^{-1}\mathbf{G}\mathbf{C}_{pr}.$$

Por construcción $\mathbf{C}_{obs} + \mathbf{G}\mathbf{C}_{pr}\mathbf{G}^T$ es SPD, por lo que su inversa también lo es. Así tenemos que $\mathbf{C}_{pr}\mathbf{G}^T \left(\mathbf{C}_{obs} + \mathbf{G}\mathbf{C}_{pr}\mathbf{G}^T\right)^{-1} \mathbf{G}\mathbf{C}_{pr}$ es SSPD, y en consecuencia esta última matriz es factorizable por $\mathbf{K}\mathbf{K}^T$. □

Una vez probado lo anterior, la estrategia a seguir es hacer la aproximación de rango bajo sobre las factorizaciones $\mathbf{K}\mathbf{K}^T$ y $\mathbf{Z}\mathbf{Z}^T$. Por lo tanto, es conveniente definir los espacios donde la aproximación es coherente.

Definición 2.3.3. Definimos:

1. El conjunto de actualizaciones semidefinidas negativas de $\mathbf{C}_{pr}$

$$\mathcal{M}_r = \left\{\mathbf{C}_{pr} - \mathbf{K}\mathbf{K}^T \succ 0 : \operatorname{rank}(\mathbf{K}) \leq r\right\}.$$

2. El conjunto de actualizaciones semidefinidas positivas de $\mathbf{C}_{pr}^{-1}$

$$\mathcal{M}_r^{-1} = \left\{\mathbf{C}_{pr}^{-1} + \mathbf{J}\mathbf{J}^T \succ 0 : \operatorname{rank}(\mathbf{J}) \leq r\right\}.$$

Donde la notación $\succ 0$ es usada para indicar que la matriz es SPD.

Ahora, denotaremos a las nuevas distribuciones de rango bajo de la siguiente manera

$$\widehat{\pi}_{post}(\mathbf{m}|\mathbf{u}) := \mathcal{N}(\widehat{\mathbf{m}}_{post}, \widehat{\mathbf{C}}_{post}).$$

Lo que esperamos de esta nueva distribución es que cumplan lo siguiente:

1. $\widehat{\mathbf{C}}_{post} \in \mathcal{M}_r$.
2. $\widehat{\mathbf{C}}_{post}^{-1} \in \mathcal{M}_r^{-1}$.
3. $\widehat{\pi}(\mathbf{m}|\mathbf{u})$ y $\pi(\mathbf{m}|\mathbf{u})$ deben definir medidas de probabilidad cercanas con la distancia de Hellinger.
4. La distancia entre $\widehat{\mathbf{C}}_{post}$ y $\mathbf{C}_{post}$ así como la que existe entre $\widehat{\mathbf{C}}_{post}^{-1}$ y $\mathbf{C}_{post}^{-1}$ debe ser pequeña para alguna métrica.

Debemos comentar que con las condiciones 1 y 2 tenemos matrices que aunque se consideran de rango bajo, son SPD y por ende invertibles. Para la condición 4 necesitamos definir otros conceptos. El primero es una generalización de la idea de eigenpar, ya que con ello es posible obtener información de manera simultánea de dos matrices.

Definición 2.3.4. Sean $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{N\times N}$, entonces

1. Llamaremos **pencil** [1] a la función

$$f(\lambda) := \mathbf{A} - \lambda\mathbf{B} =: (\mathbf{A}, \mathbf{B})$$

2. Eigenvalores generalizados al conjunto

$$\lambda(\mathbf{A}, \mathbf{B}) := \{\mathbf{z} \in \mathbb{C} | det(\mathbf{A} - \mathbf{z}\mathbf{B}) = \mathbf{0}\}.$$

3. Eigenvectores generalizados a los vectores $\mathbf{x}$ no nulos tales que

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}.$$

Notemos que esta definición generaliza el concepto de eigenvalor y eigenvector cuando $\mathbf{B}$ es la identidad. Por otro lado, esta idea permite definir un solo espectro para dos matrices.

Nota 2. Si $\mathbf{A} \in \mathbb{C}^{N\times N}$ es SSPD y $\mathbf{B} \in \mathbb{C}^{N\times N}$ es SPD, entonces el pencil $(\mathbf{A}, \mathbf{B})$ se reduce al problema clásico de eigenvalores.

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \tag{2.62}$$

Una vez hecho esto, podemos definir una métrica con la que podemos medir la distancia entre matrices. Si bien existen diferentes normas como la p o la de Frobenius, estamos interesados en una métrica que penalice fuertemente una aproximación que sea cercana a singular o con un espectro muy grande. El candidato ideal en este caso es la distancia de Förstner.

Definición 2.3.5 (Distancia de Förstner)**.** Sean $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{N\times N}$ SPD, entonces definimos la distancia de Förstner como

$$d^2_{\mathcal{F}}(\mathbf{A}, \mathbf{B}) = \mathrm{tr}\left[\ln^2\left(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}\right)\right]$$

Cabe decir que es fácil probar que esta métrica está bien definida; además podemos dar algunas propiedades de esta métrica.

Proposición 2.3.10. *Sean las matrices $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{N\times N}$ SPD; entonces la distancia de Förstner cumple lo siguiente:*

1. $d^2_{\mathcal{F}}(\mathbf{A}, \mathbf{B}) = \sum_{i=1}^{N} \ln^2(\lambda_i)$*, donde* $\{\lambda_i\}_{i=1}^{N}$ *es el conjunto de eigenvalores generalizados asociados al pencil* $(\mathbf{A}, \mathbf{B})$.

[1] Este término proviene del inglés y en un contexto de matemáticas no tiene traducción al español.

2. $d_{\mathcal{F}}(\mathbf{A},\mathbf{B}) = d_{\mathcal{F}}\left(\mathbf{A}^{-1},\mathbf{B}^{-1}\right)$.
3. $d_{\mathcal{F}}(\mathbf{A},\mathbf{B}) = d_{\mathcal{F}}\left(\mathbf{M}\mathbf{A}\mathbf{M}^{\top},\mathbf{M}\mathbf{B}\mathbf{M}^{\top}\right)$. *Para toda matriz no singular* $\mathbf{M}$.
4. $d_{\mathcal{F}}\left(\mathbf{C}_{post},\alpha\widehat{\mathbf{C}}_{post}\right) \to \infty$, *si* $\alpha \to 0,\infty$.

Demostración. Puede consultar estos resultados y más propiedades en [12]. □

La propiedad 1 es una caracterización de la distancia de Förstner con la que intuimos que si el pencil es cercano a ser singular o tiene un radio espectral demasiado grande, entonces la distancia es muy grande. Mientras que las propiedades 2 y 3 son de invarianza. Finalmente la propiedad 4 penaliza la distancia cuando se quiere aproximar con múltiplos escalares muy contrastantes.

Una vez mencionado lo anterior, es posible dar la proximación óptima de la covarianza posterior en la distancia de Förstner y la distancia de Hellinger.

Teorema 2.3.11 (Covarianza posterior óptima)**.** *Sean* $\{(\delta_i^2,\widehat{w}_i)\}_{i=1}^N$ *los eigenpares del pencil* $(\mathbf{H},\mathbf{C}_{pr}^{-1})$*, de tal manera que están ordenados como* $\delta_i^2 \geq \delta_{i+1}^2$*. Si definimos la matriz*

$$\widehat{\mathbf{C}}_{post,r} := \mathbf{C}_{pr} - \mathbf{K}\mathbf{K}^T, \quad \mathbf{K}\mathbf{K}^T = \sum_{i=1}^{r} \delta_i^2\left(1+\delta_i^2\right)^{-1}\widehat{w}_i\widehat{w}_i^{\top},$$

entonces se cumplen las siguientes propiedades:

1. $\widehat{\mathbf{C}}_{post,r} \in \mathcal{M}_r$
2. $\widehat{\mathbf{C}}_{post,r}$ *es un minimizador de* $d_{\mathcal{F}}(\mathbf{A},\mathbf{C}_{post})$ *sobre el espacio* $\mathcal{M}_r$*, más aún, la distancia mínima está dada por*

$$d_{\mathcal{F}}^2(\widehat{\mathbf{C}}_{post,r},\mathbf{C}_{post}) = \sum_{i=r+1}^{N} \ln^2\left(1+\delta_i^2\right).$$

3. *El minimizador* $\widehat{\mathbf{C}}_{post,r}$ *es único cuando los primeros* δ_i^2 *no son repetidos.*

Demostración. La prueba completa se encuentra en [35]. □

En el caso de la matriz de precisión óptima, ésta se puede determinar por el siguiente teorema.

Teorema 2.3.12 (Precisión posterior óptima)**.** *Sean* $\{(\delta_i^2,\widehat{w}_i)\}_{i=1}^N$ *como en el teorema 2.3.11*

$$\widehat{\mathbf{C}}_{post,r}^{-1} := \mathbf{C}_{pr}^{-1} + \mathbf{U}\mathbf{U}^T, \quad \mathbf{U}\mathbf{U}^T = \sum_{i=1}^{r} \delta_i^2\widetilde{w}_i\widetilde{w}_i^T, \quad \widetilde{w}_i = \mathbf{C}_{pr}^{-1}\widehat{w}_i. \tag{2.63}$$

Dicha matriz cumple lo siguiente:

1. $\widehat{\mathbf{C}}^{-1}_{post,r} \in \mathcal{M}_r^{-1}$
2. $\widehat{\mathbf{C}}^{-1}_{post,r}$ *es un minimizador de* $d_{\mathcal{F}}(\mathbf{A}, \mathbf{C}^{-1}_{post})$ *sobre el espacio* $\mathcal{M}_r^{-1}$.
3. $\widehat{\mathbf{C}}^{-1}_{post,r} = \left(\widehat{\mathbf{C}}_{post,r}\right)^{-1}$
4. *El minimizador* $\widehat{\mathbf{C}}_{post}$ *es único cuando los primeros* δ_i^2 *no son repetidos.*

Demostración. Consulte [35] para ver la prueba original. □

Dado que también nos interesan otras formas de medir similitud entre las distribuciones originales y aquellas en las que el operador se aproxima por uno de rango bajo, la siguiente proposición nos garantiza optimalidad cuando la media no cambia salvo el operador de covarianzas.

Proposición 2.3.13. *Sea* $\widehat{\mathbf{C}}_{post,r}$ *como en el teorema 2.3.11; entonces dicha aproximación minimiza en* $\mathcal{M}_r$ *la distancia de Hellinger y la divergencia de Kullback-Leibler entre las distribuciones* $\mathcal{N}(\mathbf{m}_{post}, \mathbf{C}_{post})$ *y* $\mathcal{N}(\mathbf{m}_{post}, \widehat{\mathbf{C}}_{post,r})$.

Demostración. Consulte [35]. □

Un inconveniente claro de las matrices de rango bajo óptimas para la covarianza y precisión posteriores es que se tiene que hacer el cálculo de los eigenvalores; si bien el problema se reduce a calcular el espectro de $\mathbf{C}_{pr}\mathbf{H}$, esto puede ser costoso computacionalmente por la asimetría de la matriz. Una forma de arreglar este problema es usando el hecho de que la matriz de covarianzas a priori, al ser SDP, se puede factorizar como

$$\mathbf{C}_{pr} = \mathbf{S}_{pr}\mathbf{S}_{pr}^T. \tag{2.64}$$

Con dicha factorización se tiene la siguiente factorización para $\widehat{\mathbf{C}}_{post,r}$.

Proposición 2.3.14. *Supongamos que* $\mathbf{C}_{pr}$ *se puede factorizar como en 2.64, entonces la matriz* $\widehat{\mathbf{C}}_{post,r}$ *se puede calcular como*

$$\widehat{\mathbf{C}}_{post,r} = \widehat{\mathbf{S}}_{post}\widehat{\mathbf{S}}_{post}^T,$$

donde

$$\widehat{\mathbf{S}}_{post} = \mathbf{S}_{pr}\left(\sum_{i=1}^{r}\left[\left(1+\delta_i^2\right)^{-1/2} - 1\right] w_i w_i^T + \mathbf{I}\right),$$

y los eigenpares $\{(\delta_i^2, w_i)\}_{i=1}^N$ *corresponden a la matriz* $\mathbf{S}_{pr}\mathbf{H}\mathbf{S}_{pr}^T$.

Más aún se tiene la transformación

$$w_i = \mathbf{S}_{pr}^{-1}\widehat{w}_i. \tag{2.65}$$

Demostración. Puede consultar [35] para más detalles. □

Con la proposición 2.3.14, los eigenpares por calcular correponden a la matriz $\mathbf{S}_{pr}\mathbf{H}\mathbf{S}_{pr}^T$ la cual es SSPD por construcción. Con esta característica se pueden usar algoritmos ya conocidos para matrices simétricas como Lanczos o descomposición QR simétrica (para varios ejemplos puede consultar [14]). Notemos además que los eigenvalores siguen siendo los mismos que en el pencil original, esto se debe a que $\mathbf{C}_{pr}\mathbf{H}$ y $\mathbf{S}_{pr}\mathbf{H}\mathbf{S}_{pr}^T$ son matrices similares. De esta manera el cálculo de la precisión posterior en la expresión 2.63 puede hacerse con la transformación 2.65.

Nota 3. Es claro que la matriz de covarianzas a priori 2.55 se puede factorizar y $\mathbf{S}_{pr} := \frac{1}{\sqrt{\alpha}}\mathbf{V}\mathbf{\Lambda}$.

Finalmente cabe decir que la media también depende del rango r, pues se calcula por medio de la expresión 2.60. Por lo que al usarla aproximación de rango bajo se puede redefinir como

$$\widehat{\mathbf{m}}_{post,r} = \widehat{\mathbf{C}}_{post,r}^{-1}(\mathbf{G}^T\mathbf{C}_{obs}^{-1}\tilde{\mathbf{u}} + \mathbf{C}_{pr}^{-1}\mathbf{m}_{pr}). \tag{2.66}$$

2.4. Diseño óptimo en problemas inversos

En esta sección nos adentraremos en el campo del diseño óptimo, abordaremos el concepto de manera general como se encuentra en la literatura estadística, su aplicación en problemas inversos bayesianos y finalmente el desarrollo de los algoritmos necesarios para su implementación.

2.4.1. Motivación

Cuando resolvemos una PDE por métodos numéricos, es común introducir un número considerable de nodos para asegurar convergencia y resolución del resultado. Es por esto que si el problema directo $\mathbf{Gm}$ usa muchos nodos en el parámetro, la solución a la PDE también se hará sobre muchos nodos (generalmente sobre los mismos de $\mathbf{m}$).

Lo anterior implicaría que si queremos hacer q observaciones, éstas serían más fieles si se hicieran sobre todos los nodos del mapeo directo, es decir $q = N$. Esto sin lugar a dudas es casi siempre inviable en un problema real; las razones pueden variar, sin embargo, las más comunes son que: puede ser físicamente imposible medir en ciertas coordenadas espaciales o bien no se pueden hacer tantas mediciones cuando los nodos solución están hechos sobre una rejilla fina. Por ello distingamos algunos conceptos clave:

1. **Nodos del espacio parametral**

   Son los nodos $\{\overline{\mathbf{x}}_j\}_{j=1}^{N_M} \subset \mathbb{R}^d$ que usa el método numérico para evaluar el parámetro $\mathbf{m}$. La discretización del parámetro se evalua sobre estos nodos

$$\mathbf{m} = [m_1, m_2, \cdots, m_{N_M}], \quad m_j = m(\overline{\mathbf{x}}_j).$$

2. **Nodos del problema directo**

   Aquellos nodos $\{\mathbf{x}_j\}_{j=1}^{N} \subset \mathbb{R}^d$ donde se evalua la solución numérica a la PDE

$$\mathbf{u} = [u_1, u_2, \cdots, u_N], \quad u_j = u(\mathbf{x}_j).$$

3. **Nodos de observación**

   Se trata de los nodos $\{\mathbf{x}_{\tau_k}\}_{k=1}^{q} \subset \mathbb{R}^d$ donde se hacen las mediciones. Estos nodos pueden estar ubicados en cualquier punto del dominio Ω

$$\tilde{\mathbf{u}} = [\tilde{u}_1, \tilde{u}_2, \cdots, \tilde{u}_q], \quad \tilde{u}_k = \tilde{u}(\mathbf{x}_{\tau_k}).$$

De esta manera podemos definir el concepto de diseño óptimo experimental de la siguiente manera.

Definición 2.4.1. Sea $p < q$. Un *Diseño Óptimo Experimental* es un subconjunto de $\{\mathbf{x}_{\tau_k}\}_{k=1}^{q}$ con p elementos, tal que el error en la estimación de $\mathbf{m}$ es mínimo.

En este trabajo nos enfocaremos en una versión de FEM en la que los nodos del espacio parametral y del mapeo directo coinciden, es decir $\{\mathbf{x}_j\}_{j=1}^{N} = \{\overline{\mathbf{x}}_j\}_{j=1}^{N_M}$, más aún, usaremos nodos que se encuentran sobre una rejilla. Esto por simplicidad en la simulación. Gráficamente esto podría verse como en la figura 2.4.1

2.4.2. Diseño óptimo experimental

La teoría del OED ha sido bastante desarrollada en el campo de la inferencia estadística. Principalmente la idea es aplicada en modelos de regresión que dependen de: los nodos de medición x, un parámetro θ y ruido gaussiano η aditivo; como en la expresión 2.67

$$y = \nu(x, \theta) + \eta. \tag{2.67}$$

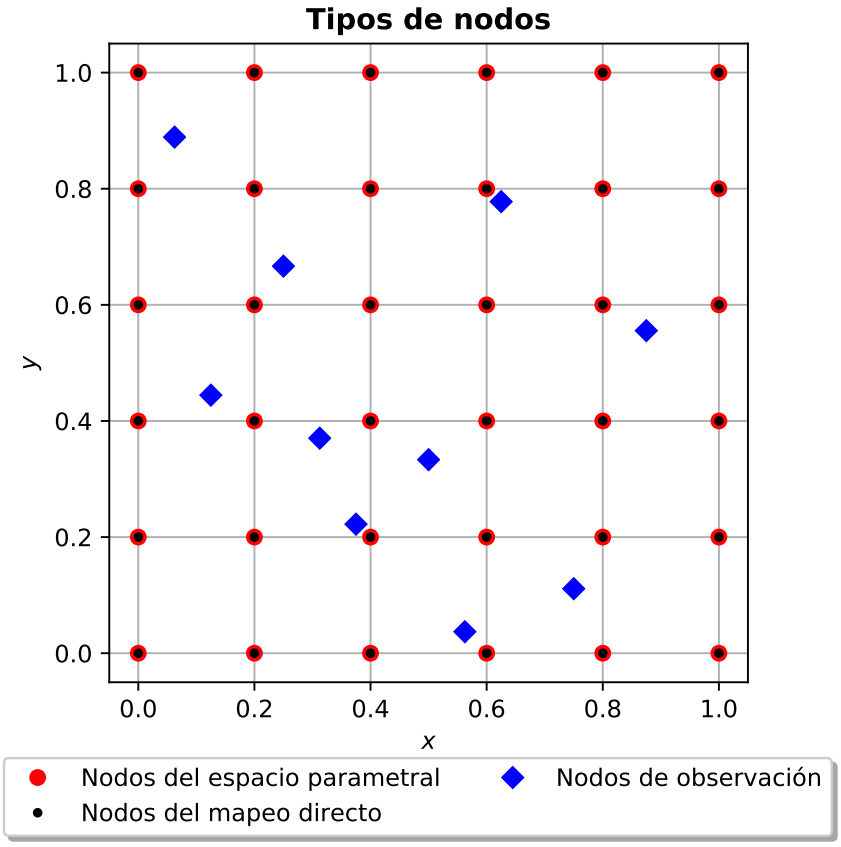

Figura 2.4.1: Tipos de nodos.

Históricamente la base de la teoría fue desarrollada en un principio para estadística frecuentista, una vez hecho esto, la idea fue mimetizada a la estadística bayesiana. Por lo que es común encontrar denominaciones similares en uno u otro caso.

Recordemos que en inferencia estadística frecuentista, el modelo de regresión calcula estimadores de θ, los cuales comunmente son denotados por $\widehat{\theta}$. Centrándonos en el caso de modelos lineales, el estimador tiene una matriz de covarianzas, la cual denotaremos por $\mathbf{C}$, por lo que la teoría del OED se basa en minimizar una función $\psi : \mathbb{R}^{N\times N} \to [0, \infty)$ que dependa de $\mathbf{C}$.

Dependiendo de quien sea ψ, es común encontrar en la literatura una letra para asignarle nombre a dicho OED. Algunos de los más conocidos son los siguientes:

1. **D**-OED

   $\psi(\mathbf{C}) := \det(\mathbf{C})$.

2. **A**-OED

   $\psi(\mathbf{C}) := \mathrm{tr}(\mathbf{C})$.

3. **c**-OED

   $\psi(\mathbf{C}) := \mathbf{c}^T(\mathbf{C})\mathbf{c}, \qquad \mathbf{c} \in \mathbb{R}^N$.

4. **E**-OED

   $\psi(\mathbf{C}) := \rho(\mathbf{C})$.

   Donde $\rho(\mathbf{A})$ es el radio espectral de $\mathbf{A}$.

Varias de estas funciones fueron deducidas al minimizar la Entropía de Shanon, maximizar diferentes funciones de utilidad, así como el cálculo de secciones geométricas que limitan el elipsoide de concentración. Todos estos conceptos están fuera del alcance de este trabajo, por lo que puede consultar [4], [27] y [28] para más detalles.

En estadística bayesiana la estimación del parámetro es por medio de su distribución posterior $\pi_{post}(\theta|y)$. Concretamente, en el caso de modelos gaussianos como el presentado en el teorema 2.3.2, la forma de definir la función ψ está en función de la covarianza de su distribución posterior $\mathbf{C}_{post}$. Es mencionado en [8] que existe una versión bayesiana de los conceptos de **D, A, c, E**-OED, en los que solamente es sustituida la matriz $\mathbf{C}$ por $\mathbf{C}_{post}$. La mayoría deducidos de manera análoga a su parte frecuentista.

Finalmente, mencionaremos que es recomendado en [2] el uso de la **A**-optimalidad, ya que el cálculo de $\mathrm{tr}(\mathbf{C_{post}})$ es muy sencillo. Por lo que a partir de ahora cuando mencionemos un OED haremos referencia a la **A** optimalidad.

2.4.3. Formulación en problemas inversos

Estamos interesados en un problema OED que contenga las siguientes características:

1. Dada una propuesta de nodos (coordenadas espaciales y/o temporales) $\{\mathbf{x}_j\}_{j=1}^q$ sobre las cuales se pretenden hacer mediciones, el usuario pueda elegir un subconjunto p de mediciones, de manera que $p < q$.
2. Que minimice la traza de la matriz de covarianzas posterior en función de dichos nodos.
3. El proceso de optimización debe mimetizar de manera adecuada un problema inverso de dimensión infinita en uno de dimensión finita.
4. La selección de nodos no debe hacerse de manera combinatoria.

Dado que $\mathbf{G}$ mapea hacia los nodos de observación, entonces implícitamente este mapeo se puede descomponer como

$$\mathbf{G} = \mathbf{B}\widehat{\mathbf{G}}, \tag{2.68}$$

donde $\widehat{\mathbf{G}} : \mathbb{R}^N \to \mathbb{R}^N$ actua sobre el parámetro y mapea hacia los nodos del mapeo directo, mientras que $\mathbf{B} : \mathbb{R}^N \to \mathbb{R}^q$ mapea de los nodos del mapeo directo hacia los nodos de observación. Al operador $\mathbf{B}$ se le conoce como **mapeo de observación**.

Para entender cómo funciona este mapeo, consideremos los siguientes casos:

El primero es cuando $\{\mathbf{x}_{\tau_k}\}_{k=1}^q \subseteq \{\mathbf{x}_j\}_{j=1}^N$, entonces el mapeo queda dado por

$$(\mathbf{B})_{i,j} = \begin{cases} 0, & i \neq j \\ 1, & i = j,\ i \in \{\tau_k\}_{k=1}^q \\ 0, & i = j,\ i \notin \{\tau_k\}_{k=1}^q \end{cases}$$

más aún, si $q = N$, entonces $\mathbf{B}$ es la matriz identidad.

En el segundo caso, que es más general, es cuando existen nodos de observación que no coinciden con los del mapeo directo, esto como en la figura 2.4.1. En este caso $\mathbf{B}$ es una matriz cuyas filas tienen pesos que actúan sobre $\mathbf{u}$. Dichos pesos surgen ya sea de algún método de **interpolación** o alguno de **aproximación**. En nuestro caso nos enfocaremos por sencillez en una interpolación lineal, esto debido a que cualquier nodo $\mathbf{x}_{\tau_k}$ de medición está contenido en la cerradura de al menos un triángulo como en la figura 2.4.2.

En este caso, la interpolación es de la forma

$$u(\mathbf{x}_{\tau_k}) \approx \alpha_{j_1} u(\mathbf{x}_{j_1}) + \alpha_{j_2} u(\mathbf{x}_{j_2}) + \alpha_{j_3} u(\mathbf{x}_{j_3}),$$

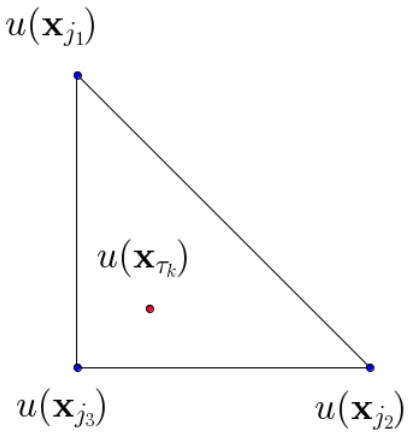

Figura 2.4.2: Interpolación.

donde

$$(\alpha_{j_1}\ \alpha_{j_2}\ \alpha_{j_3}) = (1\ x_{\tau_k}\ y_{\tau_k}) \begin{pmatrix} 1 & x_{j_1} & y_{j_1} \\ 1 & x_{j_2} & y_{j_2} \\ 1 & x_{j_3} & y_{j_3} \end{pmatrix}^{-1}.$$

Por lo que el mapeo de observación es entonces

$$(\mathbf{B})_{k,j} = \alpha_{j_1}\delta_{j_1,j} + \alpha_{j_2}\delta_{j_2,j} + \alpha_{j_3}\delta_{j_3,j},$$

donde $\delta_{i,j}$ es la función delta de Kronecker.

Una vez que hemos hemos construido el mapeo de observación hacia los nodos de medición, el diseño óptimo se reduce en escoger un subconjunto de $\{\tau_k\}_{k=1}^{q}$ con p elementos. De esta manera, las mediciones en un OED se podrían reformular como

$$\tilde{\mathbf{u}} = \tilde{\mathbf{G}}\mathbf{m} + \eta, \tag{2.69}$$

con $\tilde{\mathbf{G}} = \mathbf{W}^{1/2}\mathbf{G}$, donde $\mathbf{G}$ es como en 2.68 y $\mathbf{W} = \text{diag}(w_1, w_2, \cdots, w_q)$. Estos pesos w_k valen 1 si $\mathbf{x}_{\tau_k}$ está en el diseño óptimo y 0 en otro caso. Así el OED se reduce a elegir los pesos $\mathbf{w} := [w_1, w_2, \cdots, w_q]$ de manera que el error en el problema inverso sea mínimo.

Nota 4. El exponente $1/2$ es por convención y no es relevante, ya que el dominio es $\{0,1\}^q$.

Como ya dijimos, el criterio para elegir los nodos de manera que su aportación sea máxima, es minimizando la traza de $\mathbf{C}_{post}$ que ahora es función de $\mathbf{w}$. Por lo que el problema se puede escribir como

$$\min_{\mathbf{w}\in\{0,1\}^q} \text{tr}\left[\mathbf{C}_{post}(\mathbf{w})\right]. \tag{2.70}$$

Desafortundamente esto es muy costoso computacionalmente, ya que el número de veces que tendría que calcularse $\mathbf{C}_{post}(\mathbf{w})$ es:

$$\begin{pmatrix} q \\ p \end{pmatrix}.$$

Esto claramente es inviable; es por ello que se usó la propuesta hecha en [2]. Dicho paradigma consiste en suponer que $\mathbf{w} \in [0,1]^q$, de esta manera el problema de optimización 2.70 se transforma en uno continuo y que puede ser resuelto por algún método de optimización clásico.

Con dicho contexto, entenderemos por solución al problema OED a los p pesos w_k de mayor magnitud. Cabe mencionar que la solución puede ser difícil de interpretar si los pesos de mayor magnitud son muy pequeños respecto al valor 1. En [2] la solución a este conflicto se basa en suponer una regularización de la forma

$$\min_{\mathbf{w}\in[0,1]^q} \left[\operatorname{tr}\left[\mathbf{C}_{post}(\mathbf{w})\right] + J(\mathbf{w})\right]. \tag{2.71}$$

donde $J : [0,1]^q \to [0,\infty)$ y es convexa. El objetivo de esta penalización es discriminar nodos, es decir que los pesos que son representativos estén muy cercanos a 1 y los que no lo son que sean muy próximos a cero.

Se puede probar que $\operatorname{tr}\left[\mathbf{C}_{post}(\mathbf{w})\right]$ es una función convexa (puede consultar [27]), por lo que al sumar a la traza el término J se tiene una función convexa. Esto debido a que la suma de dos funciones convexas es convexa en un dominio convexo $[0,1]^q$. Con las propiedades anteriores el problema de optimización 2.71 tiene solución única (consulte [26]).

2.4.4. Estimadores de traza

Cabe mencionar que el cálculo de la traza no es convencional, es decir, no se reduce a la suma de la diagonal principal. Esto debido a que $\mathbf{C}_{post}(\mathbf{w})$ está en función de $\mathbf{w}$ y no se conoce una regla de correspondencia para la acción de cualquier $\mathbf{w}$.

Por lo anterior, podemos asociar el concepto de traza de este operador a su generalización en dimensión infinita. Por ejemplo, si tenemos un operador $\mathcal{K}$ y sus eigenfunciones son $\{\phi_k\}_{k\in\mathbb{K}}$, entonces su traza es de la forma 2.72 (puede revisar más detalles en el apéndice B.3).

$$\operatorname{tr}(\mathcal{K}) := \sum_{k=1}^{\infty} \langle \mathcal{K}\phi_k, \phi_k \rangle_{\mathcal{H}}. \tag{2.72}$$

Ahora, dado que en un problema de optimización con restricciones, se necesita la evaluación de la función objetivo y la acción de su matriz Jacobiana, calcular la acción de la función objetivo

resulta muy complicado y aún más la acción de su derivada. Una forma de mitigar este problema, es hacer una aproximación estocástica de la traza por medio de funciones gaussianas.

Basado en la defínción anterior de traza, en [5] se propuso una estimación de la traza para una matriz SPD dada.

Definición 2.4.2. Sea $\mathbf{A} \in \mathbb{R}^{N\times N}$ SPD con respecto al producto interior canónico; definimos un estimador de traza Gaussiano como

$$\Theta_{N_{tr}} = \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \mathbf{y}_i^T \mathbf{A} \mathbf{y}_i,$$

donde $\mathbf{y}_i$ son vectores aleatorios i.i.d. distribuidos $\sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Claramente la definición 2.4.2 es una forma de estimar la traza por medio de una simulación Monte-Carlo, por lo que

$$\mathbb{E}[\mathbf{y}^T \mathbf{A} \mathbf{y}] \approx \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \mathbf{y}_i^T \mathbf{A} \mathbf{y}_i,$$

más aún, este estimador es insesgado. Para probar esto consideremos la siguiente proposición.

Proposición 2.4.1. *Sea la matriz* $\mathbf{A}$ *y sea* $\mathbf{y} \in \mathbb{R}^N$ *un vector aleatorio con media* μ *y matriz de covarianzas* $\mathbf{\Sigma}$*. Entonces*

$$\mathbb{E}[\mathbf{y}^T \mathbf{A} \mathbf{y}] = \operatorname{tr}(\mathbf{A}\mathbf{\Sigma}) + \mu^T \mathbf{A} \mu.$$

Demostración. Puede consultar la prueba en [31]. □

Así tenemos el siguiente corolario.

Corolario 2.4.2. *El estimador 2.4.2 es un estimador insesgado.*

Demostración.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \mathbf{y}_i^T \mathbf{A} \mathbf{y}_i\right] &= \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \mathbb{E}\left[\mathbf{y}_i^T \mathbf{A} \mathbf{y}_i\right] \\
&= \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \left[\operatorname{tr}(\mathbf{A}\mathbf{I}) + \mathbf{0}^T \mathbf{A} \mathbf{0}\right] \\
&= \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \left[\operatorname{tr}(\mathbf{A})\right] \\
&= \operatorname{tr}(\mathbf{A})
\end{aligned}$$

□

2.4.5. Función objetivo

Por otro lado, cabe mencionar que al usar FEM puede ser más conveniente usar un producto interior no canónico, que esté pesado por una matriz. En [2] se propuso el uso de la matriz de masa $\mathbf{M}$, dando lugar al producto

$$\langle \cdot, \cdot \rangle_{\mathbf{M}^{-1}} := \langle \mathbf{M}^{1/2} \cdot, \mathbf{M}^{1/2} \cdot \rangle. \tag{2.73}$$

Con dicho producto interior el operador adjunto de $\mathbf{G}$ ya no es $\mathbf{G}^T$ sino

$$\mathbf{G}^* = \mathbf{M}^{-1}\mathbf{G}^T, \tag{2.74}$$

por lo que la matriz $\mathbf{H}$ (ver sección 2.3.6) ahora es de la forma

$$\mathbf{H}_{\mathbf{M}}(\mathbf{w}) = \mathbf{G}^*\mathbf{W}^{1/2}\mathbf{C}_{obs}^{-1}\mathbf{W}^{1/2}\mathbf{G},$$

y la matriz de covarianzas posterior se calcula como

$$\mathbf{C}_{\mathbf{M},post}(\mathbf{w}) = \left(\mathbf{H}_{\mathbf{M}}(\mathbf{w}) + \mathbf{C}_{pr}^{-1}\right)^{-1}. \tag{2.75}$$

Cabe mencionar que se prueba en [2] que dicha matriz cumple con ser autoadjunta con el producto interior 2.73, es decir

$$\mathbf{C}_{\mathbf{M},post}(\mathbf{w}) = \mathbf{M}^{-1}\mathbf{C}_{\mathbf{M},post}^T(\mathbf{w})\mathbf{M}.$$

Para mimetizar la idea de un estimador de traza gaussiano en un producto interior pesado, se definen los vectores aleatorios $\mathbf{z}_i = \mathbf{M}^{-1/2}\mathbf{y}_i$, con $\mathbf{y}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ i.i.d., para $i = 1, 2, \cdots, N_{tr}$. Con esto ya podemos redefinir el problema de optimización como

$$\min_{\mathbf{w}\in[0,1]^q} \left[\Theta(\mathbf{w}) + J(\mathbf{w})\right], \tag{2.76}$$

donde la parte de la función objetivo asociada a la traza está dada por

$$\Theta(\mathbf{w}) := \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \langle \mathbf{z}_i, \mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{z}_i \rangle_{\mathbf{M}^{-1}}.$$

Por otro lado, es necesario considerar la acción del gradiente de $\Theta(\mathbf{w})$, por lo que es necesario derivar $\mathbf{C}_{\mathbf{M},post}(\mathbf{w})$ respecto a las componentes de w_i. Para ello consideremos la siguiente proposición.

Proposición 2.4.3. *La derivada de* $\mathbf{C}_{\mathbf{M},post}(\mathbf{w})$ *está dada por la expresión*

$$\frac{\partial}{\partial w_i}\left(\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\right) = -\frac{1}{\sigma^2}\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{G}^*\mathbf{e}_i\mathbf{e}_i^T\mathbf{G}\mathbf{C}_{\mathbf{M},post}(\mathbf{w}).$$

Demostración. De la definición de inversa, se tiene la relación

$$\begin{aligned}\frac{\partial}{\partial w_i}\left(\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{C}_{\mathbf{M},post}(\mathbf{w})^{-1}\right) &= \frac{\partial}{\partial w_i}(\mathbf{I}) \\ &= \mathbf{0},\end{aligned} \tag{2.77}$$

al mismo tiempo por la regla de la cadena

$$\frac{\partial}{\partial w_i}\left(\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{C}_{\mathbf{M},post}(\mathbf{w})^{-1}\right) = \frac{\partial}{\partial w_i}\left(\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\right)\mathbf{C}^{-1}_{\mathbf{M},post}(\mathbf{w}) + \mathbf{C}_{\mathbf{M},post}(\mathbf{w})\frac{\partial}{\partial w_i}\left(\mathbf{C}^{-1}_{\mathbf{M},post}(\mathbf{w})\right). \tag{2.78}$$

Así al juntar las expresiones 2.77 y 2.78, llegamos a

$$\frac{\partial}{\partial w_i}\left(\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\right)\mathbf{C}^{-1}_{\mathbf{M},post}(\mathbf{w}) = -\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\frac{\partial}{\partial w_i}\left(\mathbf{C}^{-1}_{\mathbf{M},post}(\mathbf{w})\right), \tag{2.79}$$

postmultiplicamos por $\mathbf{C}_{\mathbf{M},post}(\mathbf{w})$ y tenemos la ecuación

$$\frac{\partial}{\partial w_i}\left(\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\right) = -\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\frac{\partial}{\partial w_i}\left(\mathbf{C}^{-1}_{\mathbf{M},post}(\mathbf{w})\right)\mathbf{C}_{\mathbf{M},post}(\mathbf{w}). \tag{2.80}$$

Desarrollando la derivada del lado derecho y asumiendo $\mathbf{C}_{obs} = \sigma^2\mathbf{I}$ tenemos

$$\begin{aligned}\frac{\partial}{\partial w_i}\left(\mathbf{C}^{-1}_{\mathbf{M},post}(\mathbf{w})\right) &= \frac{\partial}{\partial w_i}\left(\mathbf{G}^*\mathbf{W}^{1/2}\frac{1}{\sigma^2}\mathbf{I}\mathbf{W}^{1/2}\mathbf{G} + \mathbf{C}^{-1}_{pr}\right) \\ &= \frac{1}{\sigma^2}\frac{\partial}{\partial w_i}\left(\mathbf{G}^*\mathbf{W}\mathbf{G}\right),\end{aligned} \tag{2.81}$$

para derivar el término $\mathbf{W}$, es posible usar una identidad ya conocida para matrices diagonales y la cual está dada por

$$\mathbf{W} = \sum_{k=1}^{N} w_k\mathbf{e}_k\mathbf{e}_k^T, \tag{2.82}$$

donde $\mathbf{e}_k$ son los vectores canónicos de $\mathbb{R}^N$.

Con lo anterior tenemos entonces que

$$\begin{aligned}\frac{\partial}{\partial w_i}\left(\mathbf{G}^*\mathbf{W}\mathbf{G}\right) &= \frac{\partial}{\partial w_i}\left(\mathbf{G}^*\sum_{k=1}^{N} w_i\mathbf{e}_k\mathbf{e}_k^T\mathbf{G}\right) \\ &= \mathbf{G}^*\mathbf{e}_i\mathbf{e}_i^T\mathbf{G}.\end{aligned} \tag{2.83}$$

□

De la proposición anterior podemos calcular ya el gradiente de $\Theta(\mathbf{w})$, así

$$\begin{aligned}
\frac{\partial}{\partial w_i}\left\langle \mathbf{z}_k, \mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{z}_k\right\rangle_{\mathbf{M}} &= \left\langle \mathbf{z}_k, \frac{\partial}{\partial w_i}\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{z}_k\right\rangle_{\mathbf{M}^{-1}} \\
&= -\frac{1}{\sigma^2}\left\langle \mathbf{z}_k, \mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{G}^*\mathbf{e}_i\mathbf{e}_i^T\mathbf{G}\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{z}_k\right\rangle_{\mathbf{M}^{-1}} \\
&= -\left\langle \mathbf{M}^{1/2}\mathbf{z}_k, \mathbf{M}^{1/2}\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{G}^*\mathbf{e}_i\mathbf{e}_i^T\mathbf{G}\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{z}_k\right\rangle \\
&= -\frac{1}{\sigma^2}\mathbf{z}_k^T\mathbf{M}\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{M}^{-1}\mathbf{G}^T\mathbf{e}_i\mathbf{e}_i^T\mathbf{G}\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{z}_k \\
&= -\frac{1}{\sigma^2}\mathbf{z}_k^T\mathbf{C}_{\mathbf{M},post}^T(\mathbf{w})\mathbf{G}^T\mathbf{e}_i\mathbf{e}_i^T\mathbf{G}\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{z}_k \\
&= -\frac{1}{\sigma^2}(\mathbf{d}^k)_i^2,
\end{aligned} \tag{2.84}$$

donde $\mathbf{d}^k := \mathbf{G}\mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{z}_k$.

Así el gradiente queda dado por la expresión

$$\nabla_{\mathbf{w}}\Theta(\mathbf{w}) = -\frac{1}{\sigma^2}(\mathbf{d}^k)^2.$$

Con todo lo anterior podemos definir el algoritmo 1 para calcular a $\Theta(\mathbf{w})$ y $\nabla_{\mathbf{w}}\Theta(\mathbf{w})$. Cabe mencionar que muchos algoritmos de optimización con restricciones necesitan la acción del Hessiano, sin embargo, las paqueterías de python3 pueden calcularlo de manera numérica a través del método de diferencias finitas por medio de su gradiente.

Algorithm 1 Función objetivo $\Theta(\mathbf{w})$ y su gradiente $\nabla_{\mathbf{w}}\Theta(\mathbf{w})$

Require: N_{tr}, σ^2, $\mathbf{M}$, $\mathbf{M}^{-1/2}$, $\mathbf{G}$, $\mathbf{C}_{\mathbf{M},post}(\mathbf{w})$.

$\mathbf{Q} := \mathbf{G}\mathbf{C}_{\mathbf{M},post}(\mathbf{w})$, $\Theta(\mathbf{w}) := 0$, $\nabla_{\mathbf{w}}\Theta(\mathbf{w}) := \mathbf{0}$
for $k = 1, 2, 3, ..., N_{tr}$ **do**
  $\mathbf{y}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{z}_k := \mathbf{M}^{-1/2}\mathbf{y}_k$
  $\mathbf{a}_k := \mathbf{M}\mathbf{z}_k$
  $\mathbf{d}^k := \mathbf{Q}\mathbf{z}_k$
  $\Theta(\mathbf{w}) += \mathbf{a}_k^T \mathbf{C}_{\mathbf{M},post}(\mathbf{w})\mathbf{z}_k$
  $\nabla_{\mathbf{w}}\Theta(\mathbf{w}) += (\mathbf{d}^k)^2$
end for
$\Theta(\mathbf{w}) *= \dfrac{1}{N_{tr}}$
$\nabla_{\mathbf{w}}\Theta(\mathbf{w}) *= -\dfrac{\sigma^{-2}}{N_{tr}}$
return $\Theta(\mathbf{w})$, $\nabla_{\mathbf{w}}\Theta(\mathbf{w})$

2.4.6. Algoritmo de discriminación sucesiva

Respecto a la función de penalización $J(\mathbf{w})$, en [2] se eligió esta función en dos etapas. La primera consistió en elegir una penalización con la norma l_1

$$J(\mathbf{w}) := \alpha\|\mathbf{w}\|_1, \qquad \alpha \in \mathbb{R}^+.$$

Con esta función, el problema 2.71 es conocido en la literatura estadística como regresión LASSO. Esta función tiene la cualidad de discriminar componentes de $\mathbf{w}$ cuando $\alpha \to \infty$, es decir, mientras más grande es α, más componentes innecesarias de $\mathbf{w}$ tienden a cero y las necesarias tienden a 1. Para más detalles de esto puede consultar [17].

Una vez obtenida una solución previa, ésta se usó como condición inicial en el mismo problema de optimización con una sucesión de funciones polinomiales a trozos y suaves. Dichas funciones necesitaban de la elección de una sucesión decreciente de parámetros $\{\varepsilon_j\}$.

El problema de esta metodología es que resolver el problema de optimización depende de los parámetros α y ε_j. La elección de cada uno de estos puede ser complicada y los criterios como por ejemplo tomar α muy grande puede discriminar abruptamente los pesos o causar inestabilidad numérica en el método de optimización elegido.

En nuestro caso, se tomó la decisión de discriminar los nodos de manera sucesiva con la optimi-

zación LASSO. Para ello el problema se puede reformular en etapas sucesivas como

$$\min_{\mathbf{w}^j \in [0,1]^{q_j}} \left(\Theta(\mathbf{w}^j) + J_j(\mathbf{w}^j) \right), \tag{2.85}$$

donde $J_j(\mathbf{w}^j) = \alpha_j \|\mathbf{w}^j\|$ y donde la sucesión $\{\alpha_j\}_{j=0}^{\infty}$ cumple con ser creciente y tal que $\alpha_j \to \infty$. Por otro lado, si definimos $q_0 := q$ y resolvemos el problema de optimización para $j = 0$, los nodos $\mathbf{w}^0 \in \mathbb{R}^{q_0}$ que no sean mayores a una tolerancia $tol \ll 1$ son descartados, dejando q_1 nodos a discriminar. De esta manera, la dimensión del problema se reduce y el problema 2.85 ahora minimiza $\mathbf{w}^1 \in \mathbb{R}^{q_1}$.

Este proceso se repite sucesivamente hasta que se discriminen todos los nodos que el usuario desee, es decir, hasta que $q_j \leq q_{min}$ o bien se alcance un máximo de iteraciones j_{max} permitidas por el usuario. Lo anterior se puede resumir en el algoritmo 2.

Algorithm 2 OED

Require: q, q_{min}, j_{max}, $\mathbf{w}_{ini}$, $\mathbf{G}$, $\mathbf{M}$, $\mathbf{M}^{-1}$, $\mathbf{C}_{pr}^{-1}$, tol.

1: $\mathbf{G}_0 := \mathbf{G}$, $\mathbf{G}_0^* := \mathbf{G}^*$, $q_0 := q$, $j := 0$

2: $\mathbf{L}_0 := [1, 2, \cdots, q]$

3: **while** $q_j > q_{min}$ and $j < j_{max}$ **do**

4: **Resuelve el problema**:
$\mathbf{w}_{min}^j = \min_{\mathbf{w}^j \in [0,1]^{q_j}} (\Theta(\mathbf{w}^j) + J_j(\mathbf{w}^j))$, con la condición inicial $\mathbf{w}_{ini}$.
En cada iteración del algoritmo de optimización (con j fijo) se calcula:
- $\mathbf{w}_j$
- $\mathbf{C}_{\mathbf{M},post}(\mathbf{w}^j)$ *con* $\mathbf{G}_j$ *y* $\mathbf{G}_j^*$
- *Manda a llamar al Algoritmo* 1
- $J_j(\mathbf{w}^j)$ y $\nabla J_j(\mathbf{w}^j)$

5: **Selecciona los pesos y sus índices**:
$w_{\iota_1}^j, w_{\iota_2}^j, \cdots, w_{\iota_{N_j}}^j > tol$ de $\mathbf{w}_{min}^j$.

6: **Reduce los operadores**:
$\mathbf{L} = [\iota_1, \iota_2, \cdots, \iota_{N_j}]^T$
$\mathbf{G}_{j+1} = \mathbf{G}_j[\mathbf{L}, :]$
$\mathbf{G}_{j+1}^* = \mathbf{G}_j^*[:, \mathbf{L}]$

7: **Discrimina los nodos del total**
$\mathbf{L}_0 = \mathbf{L}_0[\mathbf{L}]$

8: **Redefine la condición inicial**:
$\mathbf{w}_{ini} = [w_{\iota_1}^j, w_{\iota_2}^j, \cdots, w_{\iota_{N_j}}^j]$

9: **Actualiza la dimensión**
$q_{j+1} = N_j$.

10: **Actualiza el paso**:
$j += 1$

11: **end while**

12: **return** $\mathbf{L}_0$

2.4.7. Cálculo de la distribución posterior en OED

Cerraremos esta sección y el capítulo remarcando que la forma en la que se ha definido la matriz de covarianzas posterior del OED 2.75 es diferente a la de la distribución 2.59. En el **primer caso**, ésta solo se usa en el **diseño óptimo**, mientras que **la segunda** se usa hasta que ya **se han discriminado los nodos**. Así mismo debemos notar que no es posible usar la aproximación de rango bajo 2.3.11 para el OED.

Es por esta razón que en el algoritmo 1 se sugiere hacer el cálculo de la matriz 2.75 por medio de una aproximación de rango bajo, por ejemplo una SVD. Para ello consideremos la siguiente factorización de dicha matriz

$$\begin{aligned}\mathbf{C}_{\mathbf{M},post}(\mathbf{w}) &= \left(\mathbf{H}_{\mathbf{M}}(\mathbf{w}) + \mathbf{C}_{pr}^{-1}\right)^{-1} \\ &= \left[\mathbf{C}_{pr}^{-1/2}\left(\mathbf{C}_{pr}^{1/2}\mathbf{H}_{\mathbf{M}}(\mathbf{w})\mathbf{C}_{pr}^{1/2} + \mathbf{I}\right)\mathbf{C}_{pr}^{-1/2}\right]^{-1} \\ &= \mathbf{C}_{pr}^{1/2}\left(\mathbf{C}_{pr}^{1/2}\mathbf{H}_{\mathbf{M}}(\mathbf{w})\mathbf{C}_{pr}^{1/2} + \mathbf{I}\right)^{-1}\mathbf{C}_{pr}^{1/2}.\end{aligned} \tag{2.86}$$

Con dicha factorización, el cálculo de la matriz de covarianzas se reduce a calcular la inversa de la matriz $\left(\mathbf{C}_{pr}^{1/2}\mathbf{H}_{\mathbf{M}}(\mathbf{w})\mathbf{C}_{pr}^{1/2} + \mathbf{I}\right)$. Si bien este cálculo parece ser costoso computacionalmente, dicha factorización ha precondicionado la matriz $\mathbf{H}_{\mathbf{M}}$ y como consecuencia el número de condición de esta matriz es muy bajo (de un orden menor a 10^1). Con este buen condicionamiento es muy rápida una descomposición SVD o algún otro método iterativo para calcular inversas.

Finalmente notemos que el cálculo de $\mathbf{C}_{pr}^{1/2}$ solo se hace una vez en todo el proceso, por lo que la parte pesada en el cálculo de la función objetivo y su gradiente se reduce a la acción de 2.86 en la traza.

Capítulo 3

Metodología

3.1. Estudios de caso

3.1.1. Determinación de la fuente en la ecuación de Poisson

Consideremos el **estudio de caso 1** dado por la expresión 2.5 con los parámetros y la solución mencionados. Definimos el problema inverso de estimar el parámetro

$$m := f(\mathbf{x}).$$

Recordemos que la solución a la PDE con FEM está dada por la expresión 2.16, con $\mathbf{K}$ la matriz de rigidez y $\mathbf{b}$ una regla de cuadratura sobre el vector $\mathbf{m}$. Por lo que el problema se puede reescribir como

$$\mathbf{KU} = \mathbf{Qm}.$$

Denotando a $\mathbf{U}$ como $\mathbf{u}$, tenemos que el mapeo directo de este problema es ahora

$$\mathbf{u} = \mathbf{Gm},$$

con $\mathbf{G} := \mathbf{K}^{-1}\mathbf{Q}$.

3.1.2. Determinación de la fuente en la ecuación elíptica

Este problema es tomado del **estudio de caso 2** del problema 2.7 con los mismos parámetros presentados y con la solución dada. De igual manera que el caso anterior, el problema inverso

consiste en estimar el parámetro

$$m := f(\mathbf{x}).$$

Como ya hemos visto, la solución a la PDE con FEM está dada por la expresión 2.21 y también puede escribirse su mapeo directo como

$$\mathbf{u} = \mathbf{G}\mathbf{m},$$

donde $\mathbf{G} := (\mathbf{K} + \mathbf{M}_{a_0})^{-1}\mathbf{Q}$.

3.1.3. Condición inicial en la ecuación de calor

Este caso corresponde al **estudio de caso 3** dado por la ecuación de calor 2.10. El problema inverso asociado a este problema es la estimación de la condición inicial

$$m := g(\mathbf{x}).$$

Una vez dada una discretización del tiempo $t = t_0, t_1, \cdots, t_{N_t}$, la solución a la PDE estaba dada de manera sucesiva por la expresión 2.31. Por lo que en cada instante k del tiempo, la solución a la PDE puede ser calculada como

$$\mathbf{U}^k = \mathbf{A}\mathbf{U}^{k-1}, \tag{3.1}$$

donde $\mathbf{A} := \left(\mathbf{M} + \frac{A\Delta t}{2}\mathbf{K}\right)^{-1}\left(\mathbf{M} - \frac{A\Delta t}{2}\mathbf{K}\right)$.

Aplicando el operador $\mathbf{A}$ de manera recursiva se tiene entonces que

$$\mathbf{U}^k = \mathbf{A}^k\mathbf{U}^0.$$

Sin embargo, si se tienen observaciones en los tiempos $\mathbf{u}(\mathbf{x}, t_{\kappa_1}), \mathbf{u}(\mathbf{x}, t_{\kappa_2}), \cdots, \mathbf{u}(\mathbf{x}, t_{\kappa_m})$, con $\{t_{\kappa_j}\}_{j=1}^{m} \subset \{t_j\}_{j=1}^{N_t}$. Entonces el mapeo directo del problema se escribe como

$$\mathbf{u} = \mathbf{G}\mathbf{m},$$

donde

$$\mathbf{u} = \begin{bmatrix} \mathbf{U}^{\kappa_1} \\ \mathbf{U}^{\kappa_2} \\ \vdots \\ \mathbf{U}^{\kappa_m} \end{bmatrix} \quad \text{y} \quad \mathbf{G} = \begin{bmatrix} \mathbf{A}^{\kappa_1} \\ \mathbf{A}^{\kappa_2} \\ \vdots \\ \mathbf{A}^{\kappa_m} \end{bmatrix}.$$

3.2. Parámetros generales usados

3.2.1. Problema directo

En todos los casos el dominio Ω se particionó sobre una rejilla con 50 nodos en la dirección x y 50 nodos en la dirección y. Con esto tenemos un total de $N = 2500$ **nodos del mapeo directo** tal y como se muestra en la figura 3.2.1.

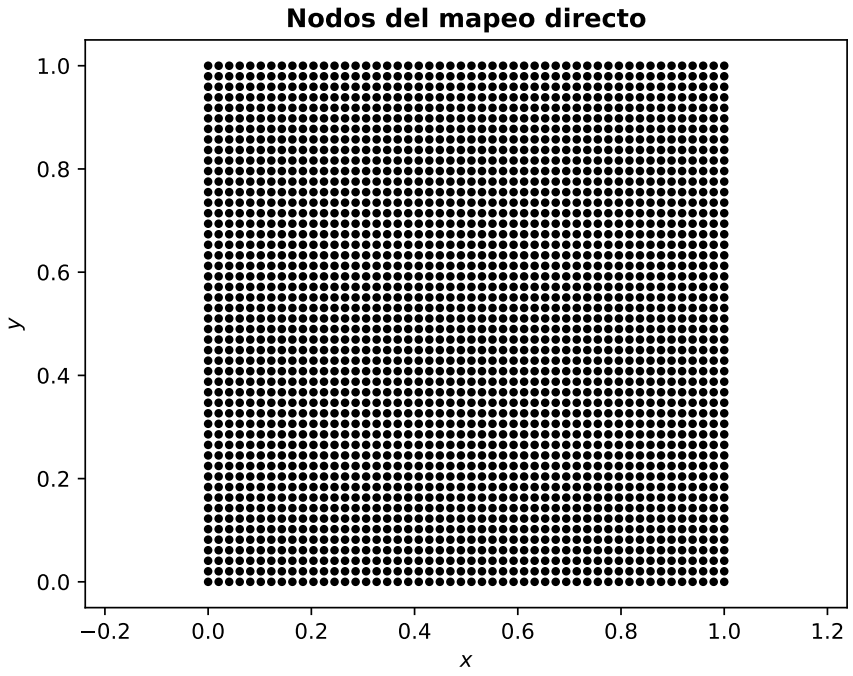

Figura 3.2.1: Nodos del mapeo directo.

Para la discretización temporal se tomó $T = 1$ y se hizo una partición con $N_t = 1001$.

3.2.2. Distribución a priori

En la construcción de la distribución a priori 2.55 son necesarios los parámetros α y s, los cuales deben satisfacer las hipótesis del teorema B.0.7. Por lo que se escogieron como $\alpha = 8.0$ y $s = 1.4$ en todos los casos.

Con los parámetros anteriores, la distribución a priori queda caracterizada por su media y su matriz de covarianzas. Dichos parámetros de la distribución pueden observarse en la figura 3.2.2.

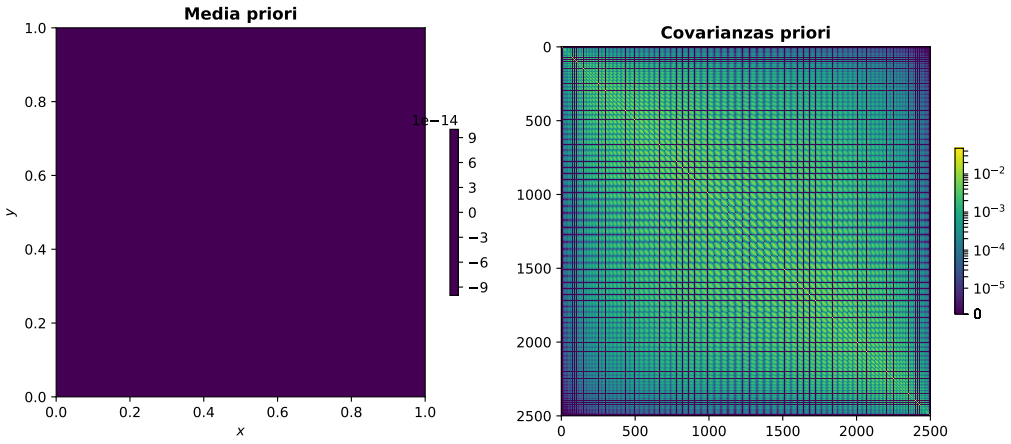

Figura 3.2.2: Distribución a priori

.

3.2.3. Verosimilitud

La verosimilitud queda definida completamente por el operador de covarianzas del ruido, el cual elegimos por simplicidad como

$$\mathbf{C}_{obs} = \sigma^2 \mathbf{I}.$$

Cabe mencionar que se necesita un nivel de ruido bajo para que la estimación del parámetro sea factible. Una forma de cuantificar el ruido es con la relación de señal a ruido, la cual se define como

$$SNR = \frac{\text{máx}\{u\}}{\sigma}.$$

Es mencionado en [6] que este parámetro debe satisfacer $SNR > 100$, por lo que en los estudios de caso 1 y 2 se crearon datos sintéticos con

$$\sigma = 10^{-4}, \tag{3.2}$$

mientras que en el estudio de caso 3 se usó

$$\sigma = 0.05. \tag{3.3}$$

3.2.4. Distribución posterior

En el cálculo de la matriz de covarianzas posterior, se necesita dar una tolerancia r en la aproximación de rango bajo. Empíricamente se determinó que con $r = 200$ es suficiente si se toman a lo más 100 observaciones.

3.2.5. Diseño óptimo

Para la construcción de la matriz de covarianzas posterior, se necesitan proponer los **nodos de observación**. En este caso se optó por una configuración que no coincidiera con los **nodos del mapeo directo** y que a su vez rellenaran el espacio de manera adecuada, por lo que usamos los nodos con la configuración conocida como Halton [16].

Tomando $q = 100$ nodos iniciales Halton, los nodos a discriminar se ven como en la figura 3.2.3.

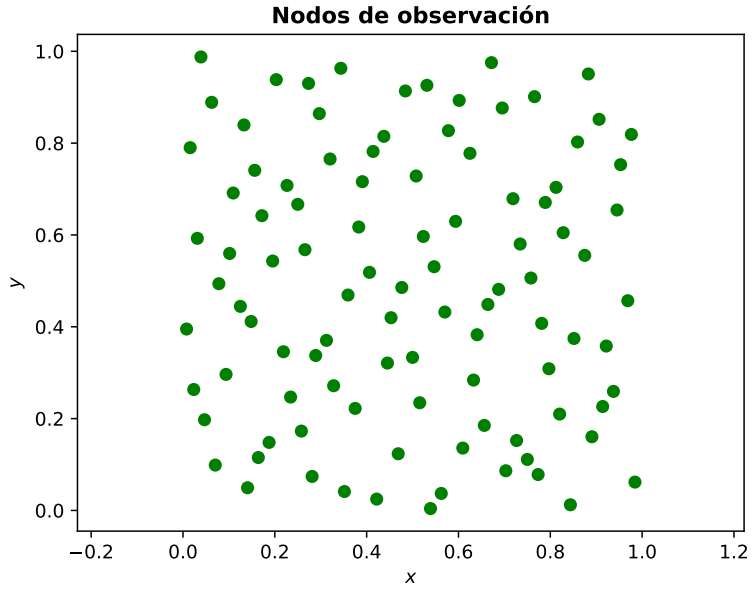

Figura 3.2.3: Nodos de observación.

Por otro lado, para la construcción de la matriz de covarianzas posterior 2.75, en [2] se recomienda usar σ no muy pequeño, por ejemplo $\sigma = 1$. En nuestro caso tambien tomamos $\sigma = 1$ **sólo para el problema OED**, mientras que el cálculo de la distribución posterior 2.3.11 se usaron los niveles de ruido 3.2 y 3.3 **una vez que terminó el proceso de OED**. Si bien no se puede probar invarianza en el problema OED respecto a σ, si tenemos la certeza de que con dicho valor de σ se tiene un escenario más pesimista, por lo que los resultados obtenidos se podrían considerar aceptables.

En el caso de la regularización LASSO, la elección de parámetros α_j fueron elegidos como

$$\alpha_{j+1} = \alpha_j + 2, \qquad \alpha_0 = 1.$$

Respecto a la tolerancia con la que se discriminan los pesos $\mathbf{w}$ se usó $tol = 10^{-4}$. Y finalmente tenemos que el número máximo de nodos a discriminar se asignaró como $q_{min} = 15$ para los estudios de caso 1 y 2, mientras que en el estudio de caso 3 se usó $q_{min} = 45$, así mismo el numéro máximo de iteraciones fue $j_{max} = 2500$.

3.3. Algoritmos y paquetería

El lenguaje de programación fué en su totalidad **python 3**, de donde se usaron diferentes bibliotecas para cada problema.

La primera fué **FENICS** [21], la cual se especializa en el método de elemento finito. Con dicha librería es posible construir las matrices de masa $\mathbf{M}$, rigidez $\mathbf{K}$ y $\mathbf{K}$. En segundo lugar se usó la librería **Scipy** [40] para el manejo vectorial de la información, la cual a su vez contiene varias bibliotecas.

Para el cálculo de las matrices $\mathbf{G}$ y $\mathbf{G}^*$ es necesario el cálculo de inversas de matrices ralas por lo se usaron las bibliotecas **scipy.sparse** y **scipy.sparse.linalg**.

En el caso de la matriz de covarianzas a priori 2.55 y su precisión 2.56, está involucrado el cálculo de eigenvalores y eigenvectores por lo que se usó la biblioteca **scipy.linalg**, la cual se especializa en álgebra lineal numérica para matrices densas. De igual manera, el cálculo de la distribución posterior 2.3.11 usa dicha librería para calcular de manera eficiente la aproximación de rango bajo por medio de algoritmos de matrices simétricas. Por otro lado, la matriz de covarianzas del diseño óptimo se calculó haciendo la descomposición SVD de la matriz 2.86; dicha descomposición también usó **scipy.linalg**.

Finalmente, tenemos que el problema de optimización 2.85, se resolvió para cada j por medio de la biblioteca de scipy llamada **scipy.optimize**. Dicha biblioteca está diseñada para la solución

a problemas de optimización. En dicha librería se usa como parámetro el método a elegir, en nuestro casó se optó por la sugerencia dada en [2] de usar el método conocido como *L-BFGS-B* [7].

Capítulo 4

Resultados

4.1. Estudio de caso 1

Una vez hecha la discriminación sucesiva por medio del algoritmo OED, los resultados de comparar las distribuciones posteriores q^k contra la q^0 están dados en la figura 4.1.1.

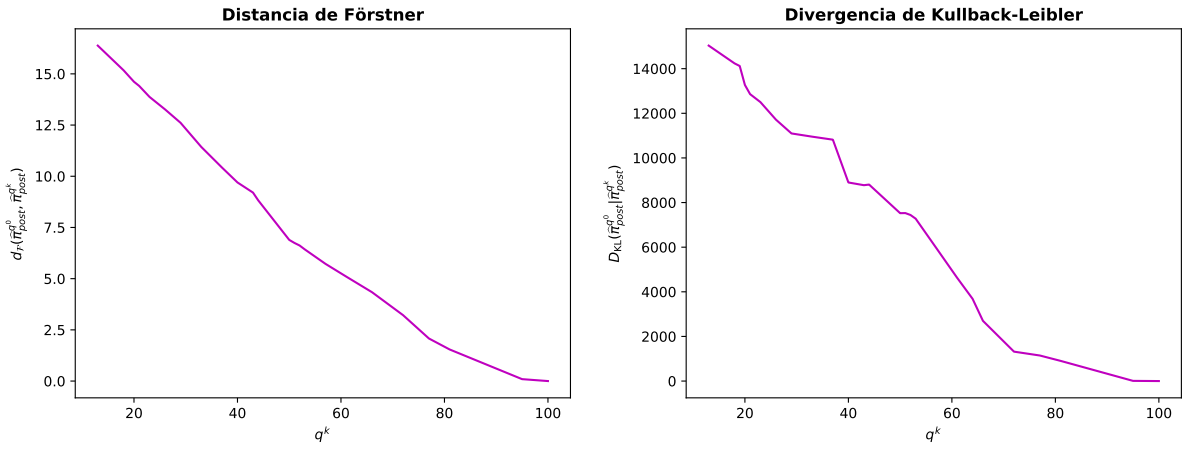

Figura 4.1.1: Comparación de las distribuciones. Estudio de caso 1.

En primer lugar consideremos la distancia de Förstner, en la que la distancia entre en las matrices de covarianzas parece ser semilineal y mantiene una escala no muy grande. Para el caso de la divergencia de Kullback-Leibler, tampoco se observan cambios muy abruptos; sin embargo, la escala es mayor, lo cual también se refleja en el hecho de que esta divergencia considera la media de la distribución.

Teniendo lo anterior en cuenta, se eligieron de manera arbitraria las distribuciones posteriores

cuando $q^k = 100, 52$ y 13. Los resultados de las configuraciones resultantes así como sus pesos se muestra en la figura 4.1.3; en ella podemos ver que existe una discriminación hacia el centro de la configuración. Así mismo, los pesos se van haciendo más contrastantes hacia los extremos del intervalo $[0, 1]$.

Por otro lado, en la figura 4.1.4 se puede apreciar que no existe un cambio sustancial de las medias de la distribución posterior, esto a pesar de haber realizado las mediciones en tan pocos nodos. Así mismo, se puede apreciar que las matrices de covarianzas se encuentran en un rango de $[0, 10^{-1})$, además de presentarse más dispersión fuera de la diagonal principal a menor número de nodos. Esto indica mayor incertidumbre a menor número de nodos; así mismo, mayor correlación entre las variables de la distribución. Lo anterior es fácil de interpretar, ya que la verosimilitud disminuye su aportación con un número pequeño de nodos y la distribución a priori es la que rellena la información.

En general se tienen resultados bastante aceptables al comparar la media de la distribución posterior con el parámetro verdadero 2.6, el cual puede observarse en la figura 4.1.2.

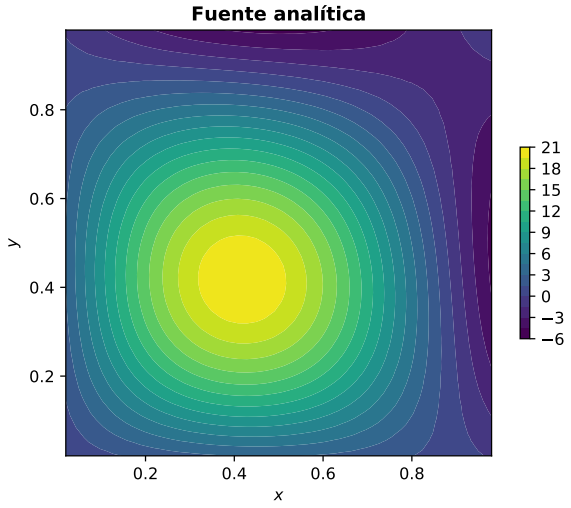

Figura 4.1.2: Parámetro verdadero. Estudio de caso 1.

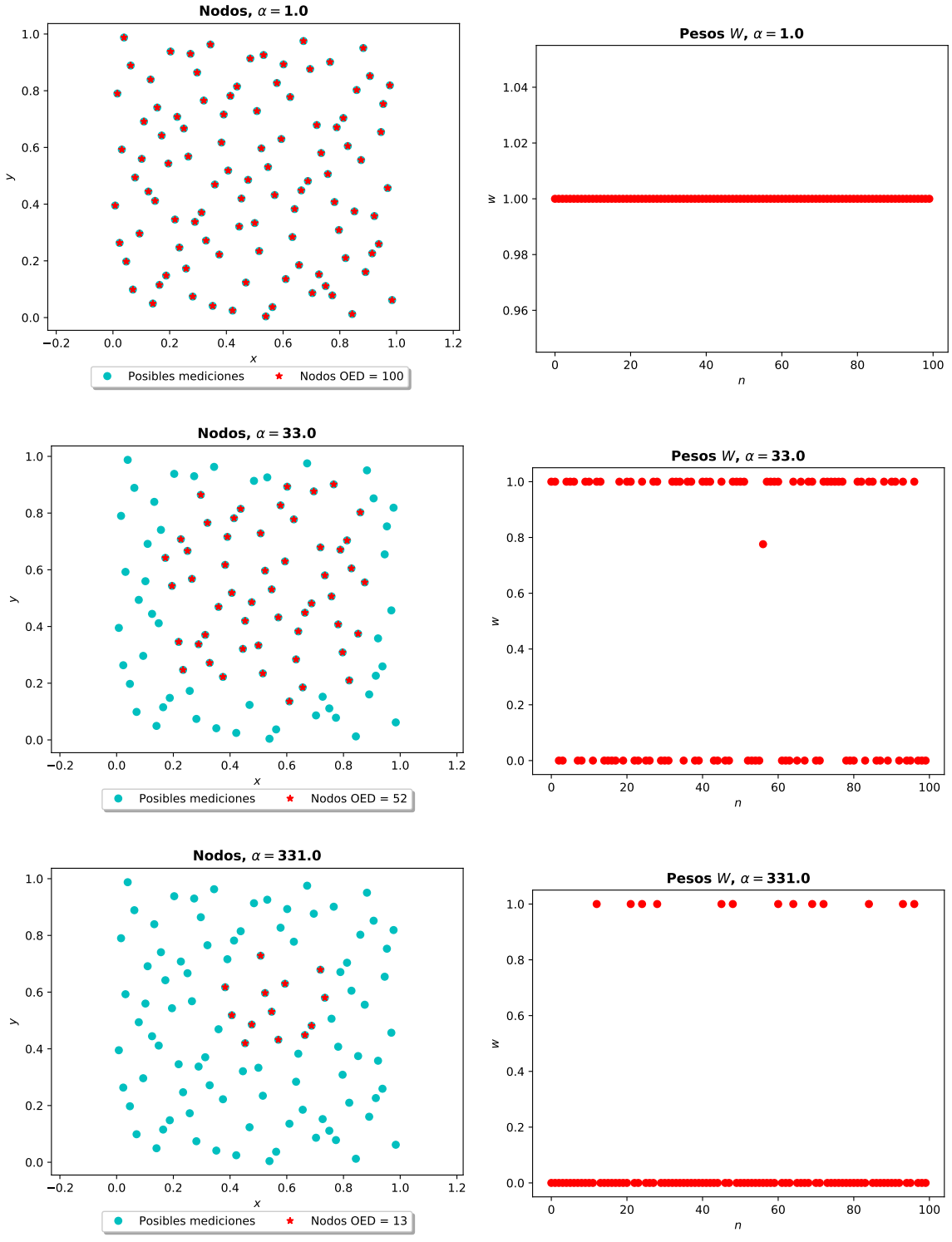

Figura 4.1.3: Nodos y pesos del OED. Estudio de caso 1.

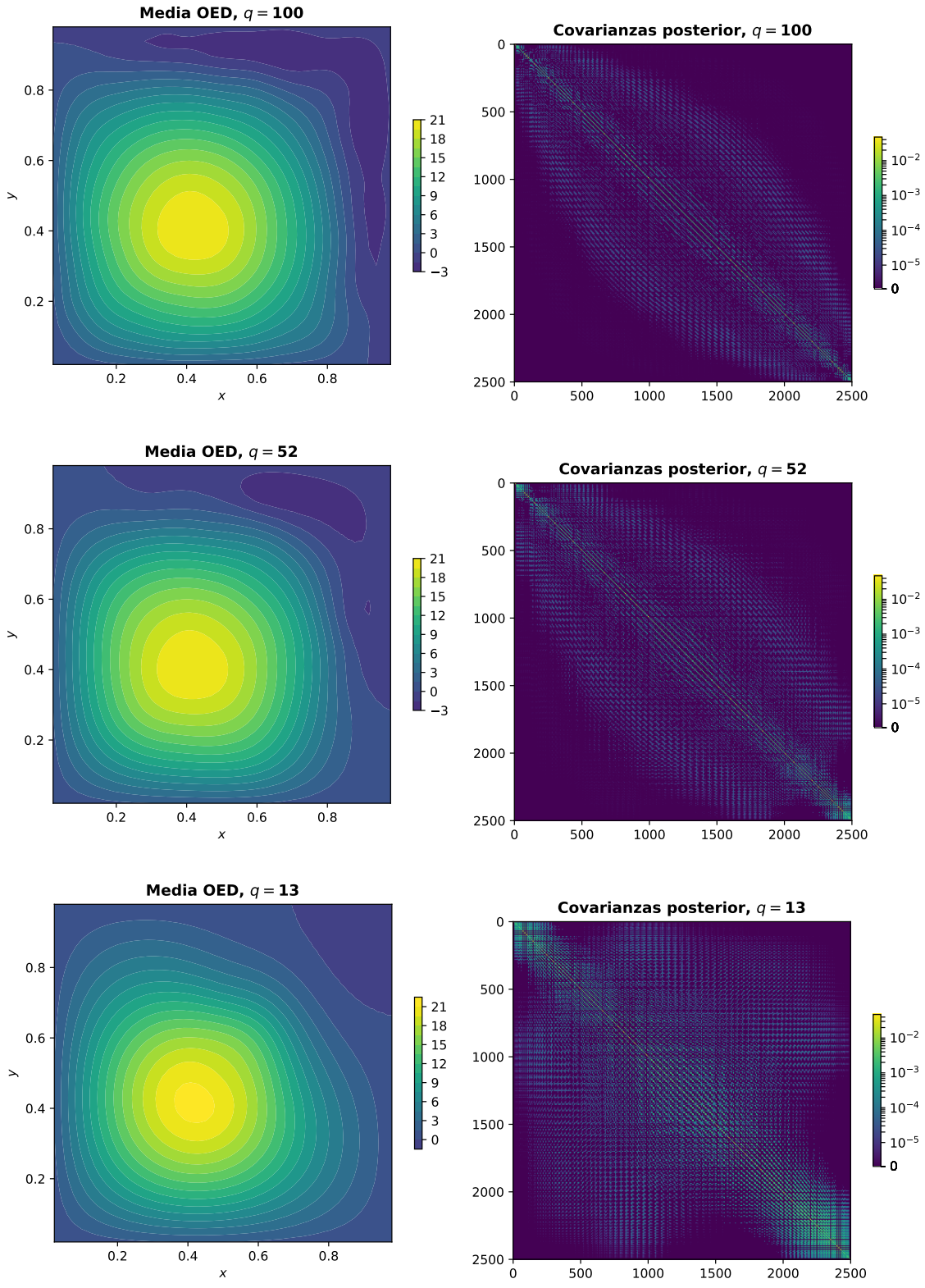

Figura 4.1.4: Distribución posterior. Estudio de caso 1.

4.2. Estudio de caso 2

Para este caso, los resultados de la comparación entre los nodos discriminados y sin discriminar se encuentra en la figura 4.2.1.

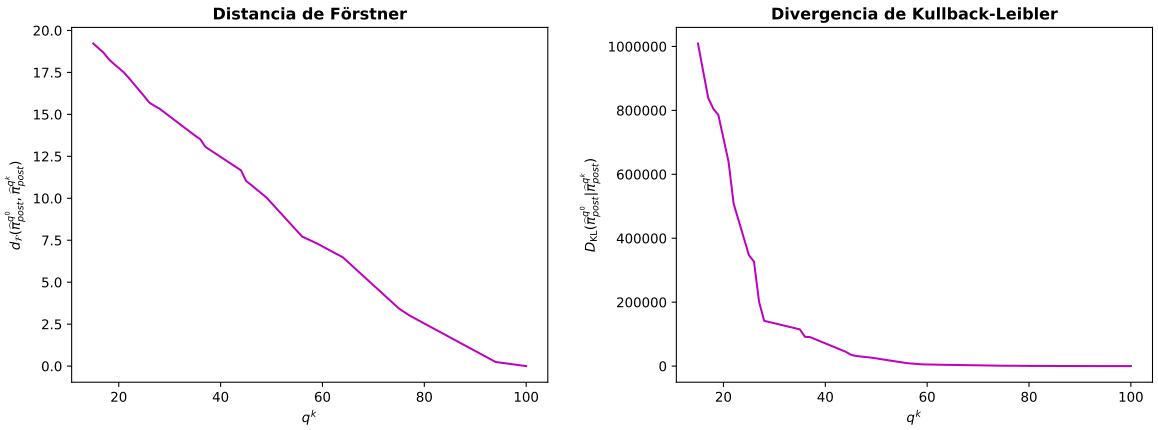

Figura 4.2.1: Comparación de las distribuciones. Estudio de caso 2.

Respecto a la distancia de Förstner observamos un comportamiento muy parecido al caso anterior, ya que la distancia entre las matrices tiene una forma semilineal y se mantiene en una escala de $[0, 20)$. Sin embargo, en la divergencia de Kullback-Leibler es claro que existe un cambio abrupto cuando se tienen $q^k = 28$ nodos, además de una escala mucho mayor. Esto nos indica que a partir de este punto la media empieza a cambiar considerablemente; para verificar este resultado observamos las distribuciones posteriores en $q^k = 100, 28$ y 15.

En la figura 4.2.3 tenemos los resultados de la discriminación sucesiva con el OED. Notemos que al igual que en el estudio de caso 1, la discriminación sucesiva se hace hacia el centro de la configuración. Así mismo, los pesos se contrastan hacia los extremos de manera muy semejante al caso anterior.

La razón por la cual los nodos y los pesos se comportan en forma muy parecida al caso anterior puede ser debido a la estructura del operador diferencial, ya que éste se diferencia del anterior con solo por el término de reacción.

Al observar la figura 4.2.4, podemos notar que desde un principio con $q^k = 100$ se nota una diferencia entre el parámetro verdadero 4.2.2 y la media; si bien no es significativa esta diferencia, sí es notable al comparar con la figura 2.8. Dicha diferencia se acentua un poco en 28 nodos y donde finalmente se nota una diferencia relativamente significativa aunque aceptable en 15 nodos. Por otro lado, las matrices de covarianzas también presentan dispersión fuera de la

diagonal principal cuando se disminuyen los nodos de observación, además de mantenerse en los mismos rangos que el estudio de caso anterior.

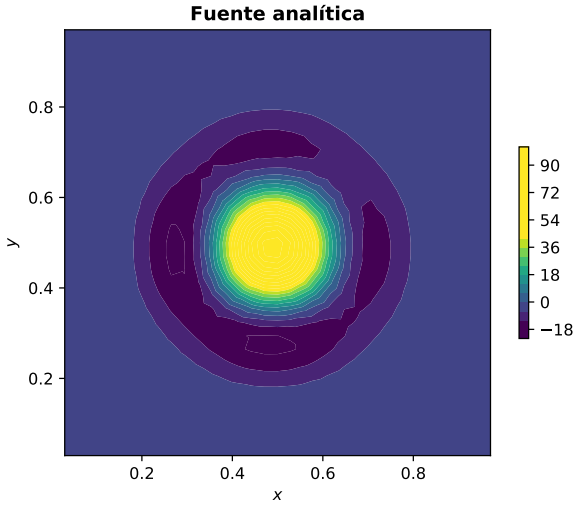

Figura 4.2.2: Parámetro verdadero. Estudio de caso 2.

Es claro que las estimaciones de la media son un poco más abruptas respecto al caso anterior; esto se debe a que la solución analítica $u(\mathbf{x})$ de los datos dada por la ecuación 2.9, contiene zonas de muy alto gradiente cuando $p = 10$, las cuales se modelan bien con elemento finito bajo nodos adaptativos (consulte [22] para más detalles). Lo anterior provoca que el operador del mapero directo $\mathbf{G}$ no pueda capturar de manera adecuada los valores que más aportan información en la verosimilitud, por lo que a pesar de que en la formulación OED son nodos óptimos, la verosimilitud necesita aportar más información.

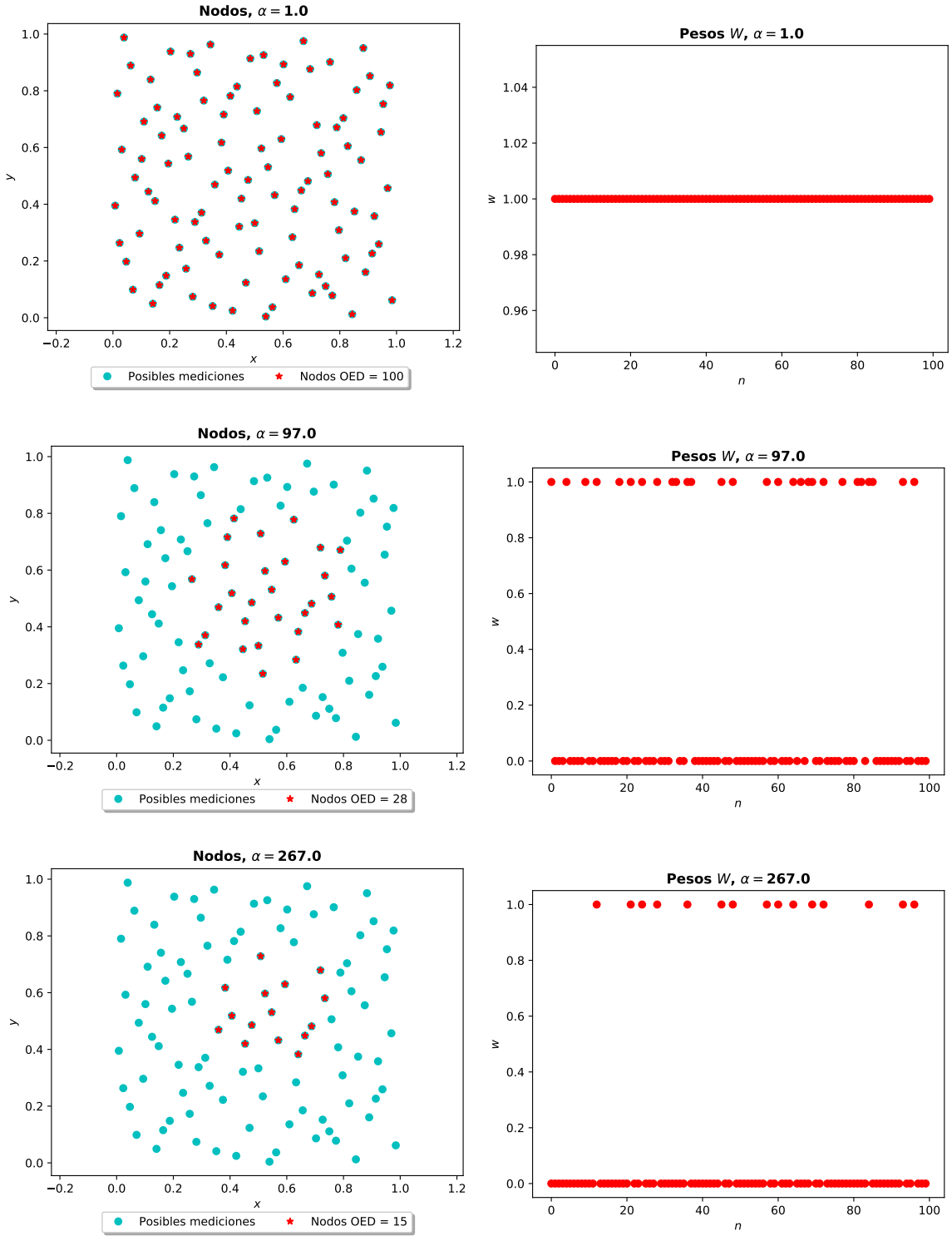

Figura 4.2.3: Nodos y pesos del OED. Estudio de caso 2.

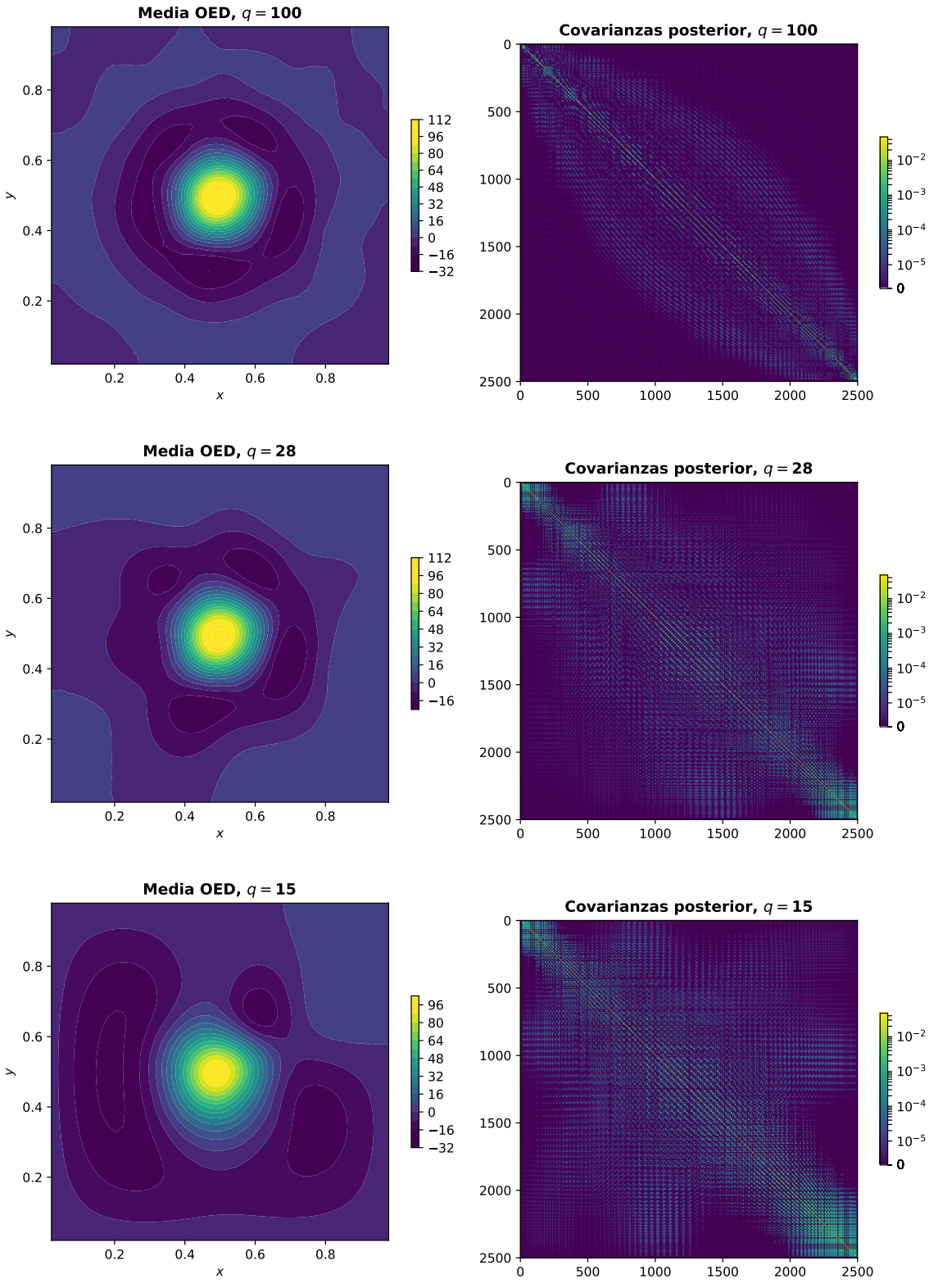

Figura 4.2.4: Distribución posterior. Estudio de caso 2.

4.3. Estudio de caso 3

Finalmente para el estudio de caso 3, las gráficas de comparación entre las distribuciones son mostradas en la figura 4.3.1. En ella podemos notar que existe una discriminación abrupta en $q^k = 151$ nodos para la distancia de Förstner; después de eso ya no se observa un cambio sustancial en dicha distancia.

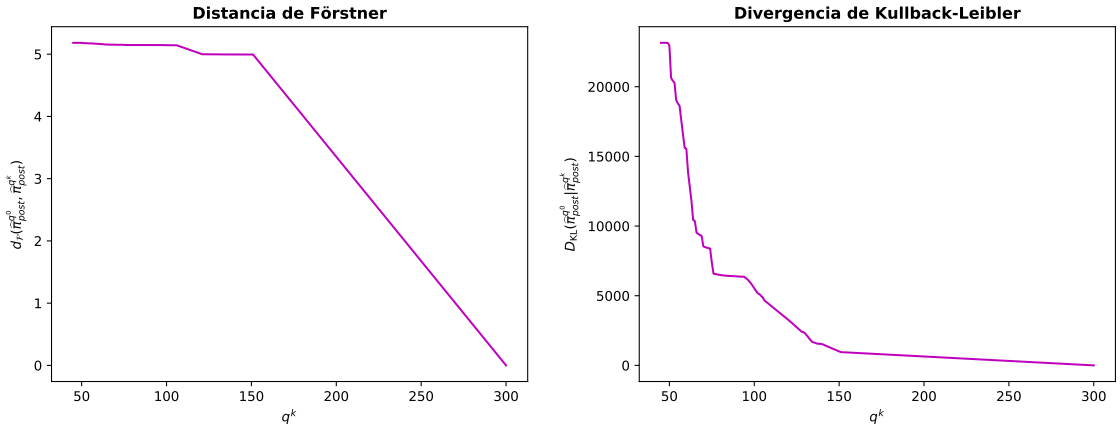

Figura 4.3.1: Comparación de las distribuciones. Estudio de caso 3.

Por otro lado, en la divergencia de Kullback-Leibler además del salto en 151 nodos, podemos ver otro en 94, a partir del cual se tiene un incremento significativo en la divergencia. Nuevamente esto es un indicador de cambios en la media de la distribución. Debido a lo anterior se examinaron los datos en $q^k = 300, 151, 94$ y 49.

Como se mencionó en el capítulo anterior, los nodos espaciales a discriminar en este caso eran los mismos para los tiempos $t = 0.5, 0.75$ y 1 como se muestra en la figura 4.3.3. Observemos ahora la figura 4.3.4; en ella tenemos que se han descartato todos los nodos en el tiempo $t = 1.0$ y varios más en el tiempo $t = 0.75$ dejando solo 151 nodos dispobibles.

Respecto al caso en que tenemos solo 94 nodos como en la figura 4.3.5, vemos que se han descartado todos los nodos en el tiempo $t = 0.75$. Al seguir con la discriminación sucesiva observamos 49 nodos en el tiempo $t = 0.5$, es claro que el proceso de discriminación tardó demasiado, pues notamos que esto fue en $\alpha_k = 6085$.

Lo anterior se debe a que la ecuación de difusión aplasta la solución conforme avanza el tiempo, esto provoca que la verosimilitud definida por los datos $\mathbf{u}$ no sea informativa para valores grandes del tiempo. Dicha información ya se encuentra implícita en el operador $\mathbf{G}$. Así mismo, los nodos se van centrando salvo algunas excepciones debido a que un operador parabólico en esencia es muy parecido a un elíptico en su discretización numérica.

Al observar los correspondientes pesos de esta discriminación de nodos en las mismas figuras, podemos notar que los pesos se marginzalizan hacia la izquierda, ya que los 300 nodos están enumerados de izquierda a derecha en función del tiempo.

Ahora, al analizar las figuras 4.3.7, 4.3.8 y comparar las medias con respecto al parámetro verdadero 2.11 representado en la figura 4.3.2, podemos notar que en realidad hay un cambio apenas visible en la discriminación a 45 nodos. En general los resultados en sí son bastante aceptables.

Por último, al analizar las matrices de covarianzas en las figuras 4.3.7 y 4.3.8, queda claro que no hay un cambio perceptible en la matriz de covarianzas, lo cual ya se ve en la poca variación de la distancia de Förstner y en la divergencia de Kullback-Leibler, la cual preserva una escala no mayor a 25000; esto es muy poco en comparación con el estudio de caso anterior.

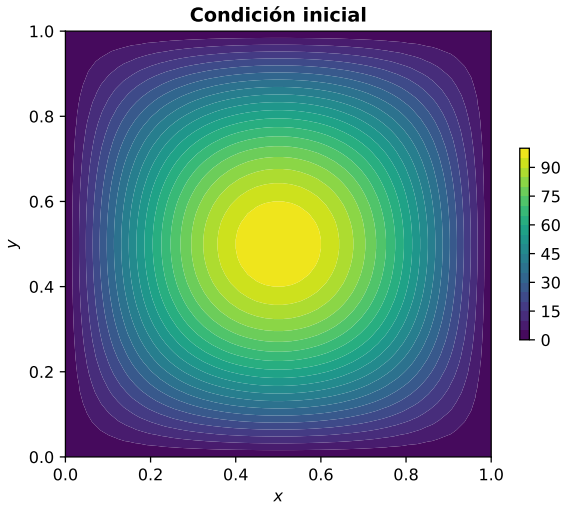

Figura 4.3.2: Parámetro verdadero. Estudio de caso 3

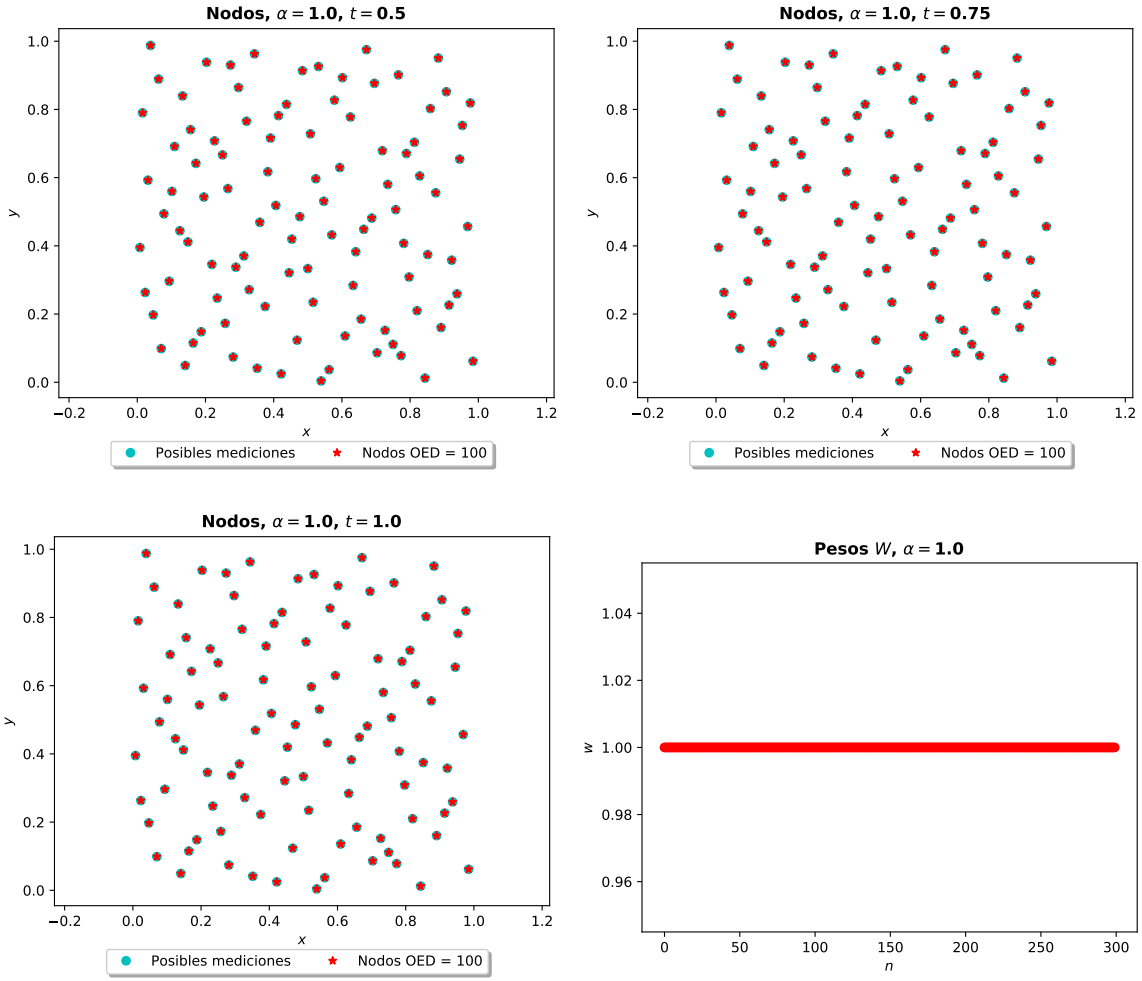

Figura 4.3.3: Nodos y pesos del OED. Estudio de caso 3, $\alpha = 1$.

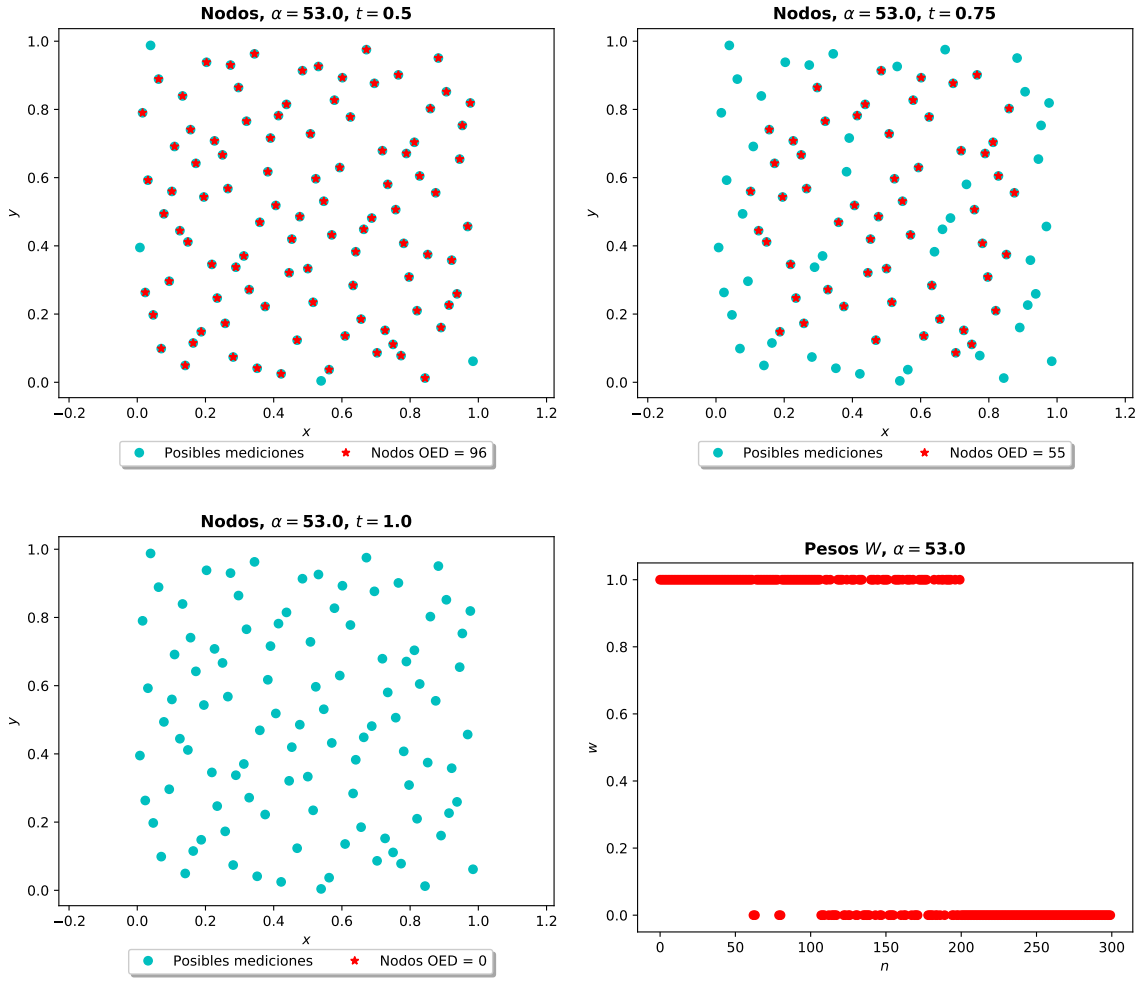

Figura 4.3.4: Nodos y pesos del OED. Estudio de caso 3, $\alpha = 53$.

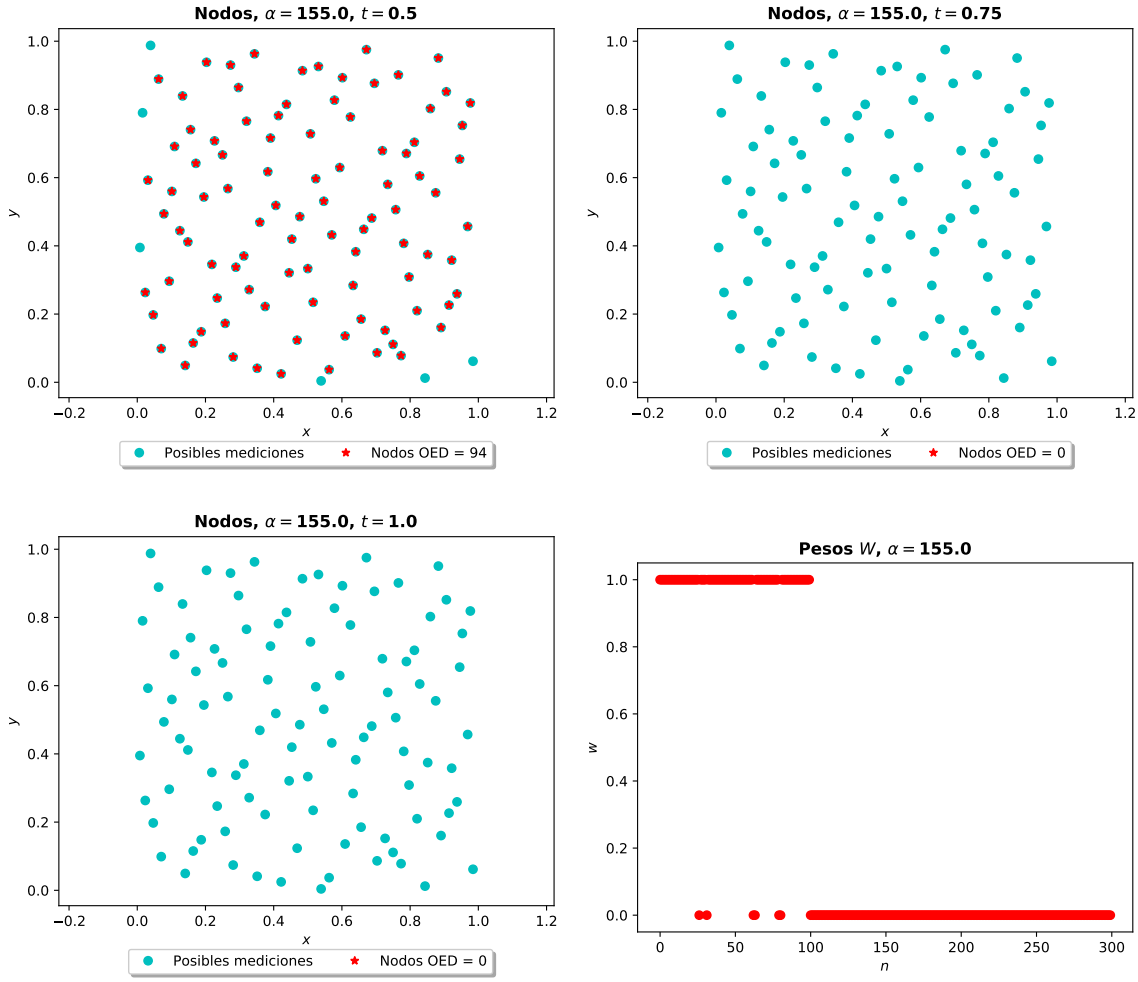

Figura 4.3.5: Nodos y pesos del OED. Estudio de caso 3, $\alpha = 155$.

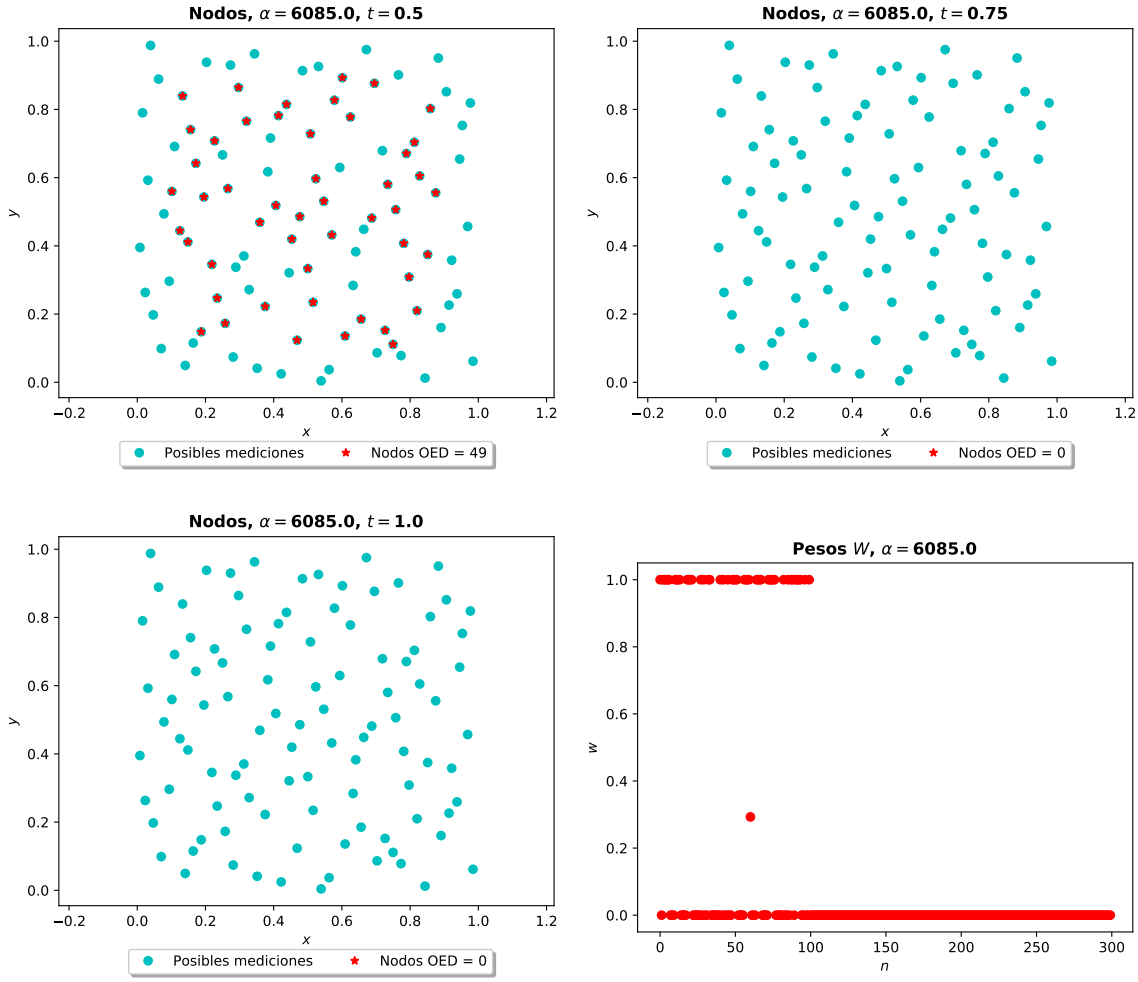

Figura 4.3.6: Nodos y pesos del OED. Estudio de caso 3, $\alpha = 6085$.

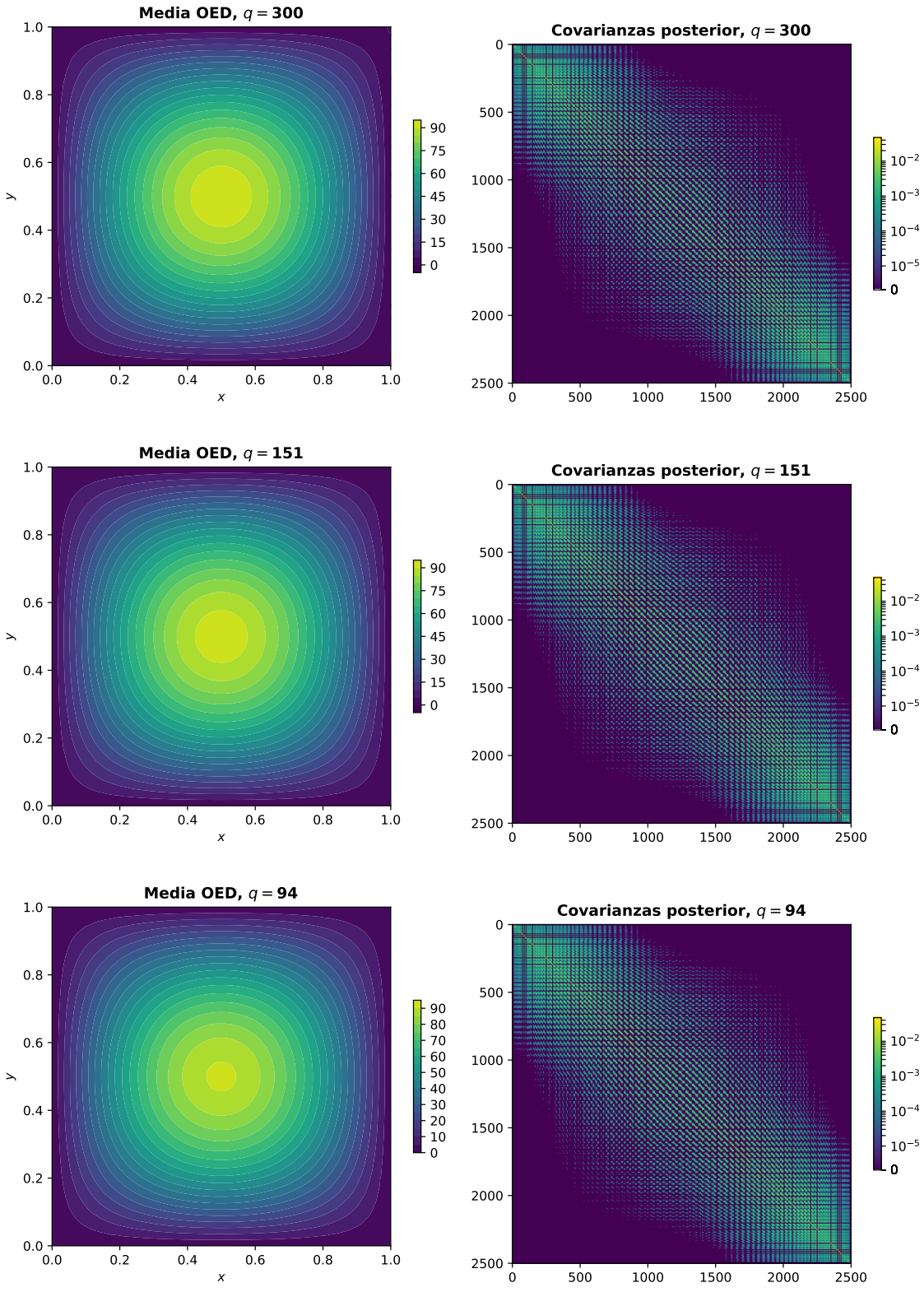

Figura 4.3.7: Distribución posterior. Estudio de caso 3.

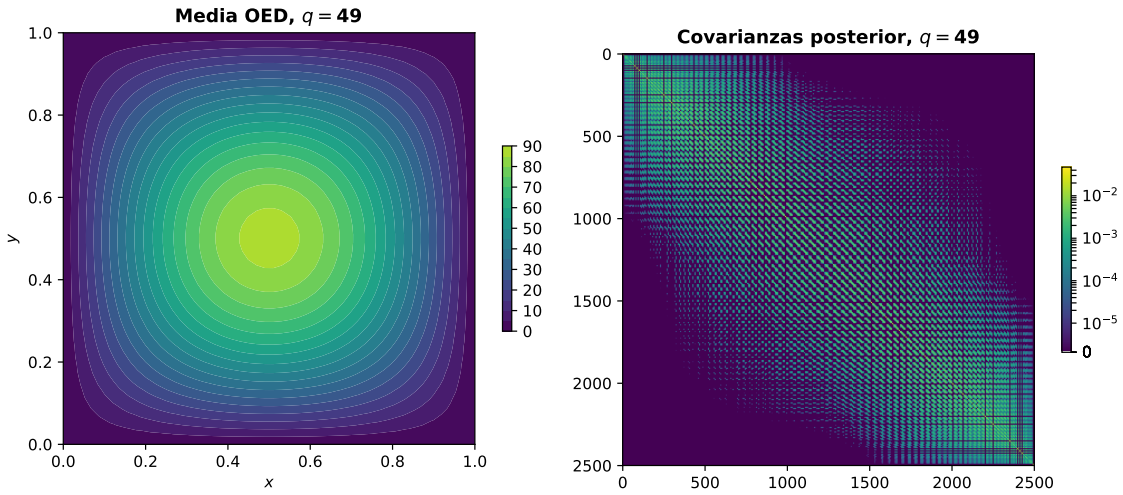

Figura 4.3.8: Distribución posterior. Estudio de caso 3.

Capítulo 5

Conclusiones

Los resultados obtenidos en el capítulo 4 nos indican que la metodología propuesta del algoritmo OED puede considerarse aceptable, ya que exhibe el poco efecto de los nodos en los últimos pasos del tiempo para el estudio de caso 3.

Así mismo, tenemos en todos los casos una centralización de los nodos, lo cual reafirma la idea de que los nodos que se encuentran en la frontera o cerca de ella, son aquellos que aportan menos información al tener una frontera nula. Es decir, la interacción entre los nodos con la frontera está presente en la discriminación.

Por otro lado, el estudio de caso 2 demuestra que a pesar de que la traza de una matriz de covarianzas es un buen criterio para definir optimalidad sin conocer los datos u, éstos también aportan información valiosa y necesaria para muchos problemas. Sin embargo, como un primer intento de experimentación es recomendable una discriminación OED ya que, una vez hechas las mediciones sobre los nodos, si los resultados no son convincentes y se desea a futuro medir de nuevo, la información previa se puede reutilizar simplemente con asignar de manera fija en nuevo OED a los pesos medidos como 1 y redefinir los nodos de observación con la evidencia acumulada hasta el momento.

Finalmente diremos que el usar la formulación del problema de optimización con regularizaciones LASSO sucesivas, tiene la ventaja de guardar toda la información de manera secuencial y analizar de manera previa la distancia de Förstner. Lo anterior nos indicaría la forma en la que la incertidumbre del problema cambia en función de los nodos de medición.

Apéndice A

Análisis Funcional

Dentro de los elementos teóricos que se abordan en esta tesis, se encuentran el estudio de Ecuaciones Diferenciales Parciales, el Método de Elemento Finito y Problemas Inversos. En el primer caso, la teoría base se encuentra en los espacios de Sobolev, los cuales se abordarán más adelante. Por otro lado la teoría de elemento finito hace uso de estas mismas herramientas cuando recurre nuevamente a la teoría de las parciales, así mismo los conceptos de interpolación involucrados requieren también del estudio de los espacios de Sobolev. Finalmente, la parte correspondiente a los problemas de regularización clásicos, hacen uso exhaustivo del Teorema de Proyecciones Ortogonales y teoría de operadores, ambos temas tratados aquí mismo. Para el lector interesado en las pruebas de éstos y otros temas relacionados puede consultar los libros de [33], [30], [10] y [29].

A.1. Espacios de Hilbert

Definición A.1.1. Sea $(X, \|\cdot\|_X)$ un espacio vectorial normado; diremos que una sucesión $\{x_k\}$ es de Cauchy si se cumple que

$$\lim_{k,l\to\infty} \|x_k - x_l\|_X = 0.$$

Si bien en $\mathbb{R}$ las sucesiones de Cauchy implican convergencia de la sucesión, esto no siempre es así. Existen númerosos ejemplos de espacios donde esto no ocurre, por ejemplo: El espacio de las funciones continuas en un intervalo cerrado $C^0([a,b])$ con la norma L^2.

Definición A.1.2. Sea X un espacio vectorial normado, diremos que es completo o de Banach si toda sucesión de Cauchy $\{x_k\}_{k=0}^{\infty}$ converge a un elemento $x \in X$.

En particular estamos interesados en espacios con producto interior, pues éstos definen algunos espacios de interés en ecuaciones diferenciales parciales.

Definición A.1.3. Sea $\mathcal{H}$ un espacio de Banach, diremos que es de Hilbert si su norma $\|\cdot\|_{\mathcal{H}}$ proviene de un producto interior $\langle\cdot,\cdot\rangle_{\mathcal{H}}$.

Uno de los teoremas célebres en espacios de Hilbert sobre caracterización de espacios es el siguiente.

Teorema A.1.1 (Proyecciones Ortogonales)**.** *Sea $\mathcal{H}$ un subespacio de Hibert y sea $V \subset \mathcal{H}$ un subespacio cerrado; entonces se cumple lo siguiente.*

1. *Existe un único $w \in V$ tal que es solución al problema de minimización para el funcional*

$$\|w - x\|_{\mathcal{H}} = \inf_{v \in V} \|v - x\|_{\mathcal{H}} .$$

2. *$w = x$ si y solo si $x \in V$.*

3. *Existe un único $z \in V^{\perp}$ tal que*

$$x = w + z.$$

Esto es equivalente a que $\mathcal{H} = V \bigoplus V^{\perp}$ y además

$$\|x\|_{\mathcal{H}} = \|w\|_{\mathcal{H}} + \|z\|_{\mathcal{H}} .$$

Definición A.1.4. Diremos que $\mathcal{H}$ es separable si existe un subconjunto denso en $\mathcal{H}$ que sea numerable.

Ejemplos destacados de estos espacios son $\mathbb{R}^n$ con el producto interior usual y $L^2(\mathbb{R}^n)$. Este tipo de conjuntos son de vital importancia, ya que admiten una **base ortonormal**, es decir, para cada $x \in \mathcal{H}$ existe una sucesión $\{z_k\} \subset \mathcal{H}$ ortonormal tal que

$$x = \sum_{k=0}^{\infty} \langle x, z_k \rangle_{\mathcal{H}} z_k.$$

Otro concepto importante es el de dualidad, el cual introduciremos a continuación.

Definición A.1.5. Sean el funcional lineal $T : \mathcal{H} \to \mathbb{R}$, diremos que es continuo si para cualquier $x \in \mathcal{H}$, existe $C > 0$ tal que

$$|Tx| \leq C \|x\|_{\mathcal{H}} . \tag{A.1}$$

Todos estos operadores pueden englobarse en un solo espacio, el cual definimos a continuación.

Definición A.1.6. Definimos a $\mathcal{H}^*$ como el conjunto de todos los funcionales linales continuos; a este espacio le llamaremos **espacio dual de** $\mathcal{H}$.

La importancia del espacio dual es su relación de isomorfismo con el propio espacio $\mathcal{H}$, esto queda dado por el siguiente teorema.

Teorema A.1.2 (Representación de Riesz)**.** *Sea $\mathcal{H}$ un subespacio de Hibert; entonces para todo $T \in \mathcal{H}^*$ existe un único $y \in \mathcal{H}$*

$$Tx = \langle x, y \rangle_{\mathcal{H}}, \qquad \forall x \in \mathcal{H}.$$

Nota 5. La acción de T sobre x usualmente se representa como $\langle T, x \rangle_*$.

A.2. Teorema de Lax-Milgram

En el estudio de las ecuaciones diferenciales parciales, es común usar la forma variacional de las ecuaciones; dicha estructura generamente es equivalente a una forma bilineal. Por lo que empezaremos por definir este concepto.

Definición A.2.1. Sea W un espacio vectorial, decimos que la función $B : W \times W \to \mathbb{R}$ es una forma bilineal si su mapeo $B(x, y)$ es lineal en cada entrada.

Estamos interesados en formas bilineales que cumplan ciertas características de regularidad, las cuales definimos a continuación.

Definición A.2.2. Sea $B(x, y)$ una forma bilineal asociada al espacio de Hilbert $\mathcal{H}$. Definimos las propiedades de:

1. **Coercividad**
   Si existe una constante $C_1 > 0$ tal que
   $$C_1 \|x\|^2_{\mathcal{H}} \leq B(x, x).$$
2. **Continuidad**
   Si existe una constante $C_2 > 0$ tal que
   $$|B(x, y)| \leq C_2 \|x\|_{\mathcal{H}} \|y\|_{\mathcal{H}}.$$

Con lo anterior podemos establecer ahora el famoso teorema conocido como Lax-Milgram.

Teorema A.2.1 (Lax-Milgram)**.** *Sea B una forma bilineal coerciva y continua sobre $\mathcal{H}$ y sea tambien $f \in \mathcal{H}^*$, entonces existe un único $x \in \mathcal{H}$ tal que*

$$B(x, y) = \langle f, y \rangle_* \quad \forall y \in \mathcal{H}.$$

A.3. Espacios de Sobolev

Una forma de compactar la notación de la derivada en todos los ordenes posibles es usando la notación mulitiíndice; para ello consideremos el elemento $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n) \in \mathbb{N}^n$ (considerando al 0 en los naturales). A dicho vector se le conoce como un multiíndice de orden $|\alpha| := \sum_{j=1}^{n}$.

Definición A.3.1. Sea $\Omega \subset \mathbb{R}^n$ abierto y $u \in C^k(\Omega)$, definimos la derivada multíndice como

$$D^\alpha u(x) := \frac{\partial^{|\alpha|} u(x)}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}} = \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n} u,$$

y para $k = |\alpha|$

$$D^k u(x) := \left\{ D^\alpha u(x) || \alpha \mid = k \right\}.$$

Definición A.3.2. Sea $\gamma \in (0, 1]$; diremos que una función es γ-*Hölder continua* si existe $C > 0$ tal que

$$|u(x) - u(y)| \leq C|x - y|^\gamma \qquad x, y \in \Omega,$$

se puede asociar una norma.

Definición A.3.3. Sea $u \in C^k(\Omega)$ cuyas derivadas están acotadas, definimos la *norma Hölder* como

$$\|u\|_{C^{k,\gamma}(\bar{\Omega})} := \sum_{|\alpha| \leq k} \|D^\alpha u\|_{L^\infty(\Omega)} + \sum_{|\alpha|=k} [D^\alpha u]_{C^{0,\gamma}(\bar{\Omega})},$$

donde

$$[u]_{C^{0,\gamma}(\bar{\Omega})} := \sup_{x,y \in \Omega} \left\{ \frac{|u(x) - u(y)|}{|x - y|^\gamma} \right\}$$

Las funciones cuyas dervadas acotadas son todas Hölder continuas pueden caracterizarse en el siguiente espacio

Definición A.3.4. Definimos el espacio k-Hölder como

$$C^{k,\gamma}(\bar{\Omega}) = \{u \in C^k(\Omega) \,|\, \|u\|_{C^{k,\gamma}(\bar{\Omega})} < \infty\}.$$

Finalmente se puede afirmar que estos espacios son de Banach.

Algunos otros espacios de funciones de interés son los siguientes.

Definición A.3.5. Definimos el espacio de las *funciones de prueba* $C_c^\infty(\Omega)$ como aquellas que son continuamente diferenciables de todos los órdenes y además cumplen que cualquier $\phi \in C_c^\infty(\Omega)$ tiene un soporte compactamente contenido en Ω.

Otro espacio de interés es el de la funciones localmente integrables, el cualde finimos a continuación.

Definición A.3.6. Llamaremos espacio de las funciones localmente integrable como aquel dado por

$$L^1_{\text{loc}}(\Omega) = \left\{u \in L^1(V) \,|\quad \text{para todo} \quad V \subset\subset \Omega\right\},$$

donde $V \subset\subset \Omega$ indica que V está compactamente contenido en Ω.

Es con estas funciones que ahora podemos definir la derivada débil, como una generalización de la derivada clásia.

Definición A.3.7. Sean $u, v \in L^1_{\text{loc}}(\Omega)$, definimos a v como la α-ésima derivada débil de u si se cumple que

$$\int_\Omega u D^\alpha \phi d\mathbf{x} = (-1)^{|\alpha|} \int_\Omega v\phi d\mathbf{x}, \qquad \forall \phi \in C_c^\infty(\Omega).$$

A dicha derivada la denotaremos como $D^\alpha u := v$.

Es con este concepto que podemos definir ahora los espacios de Sobolev.

Definición A.3.8. Definimos el *Espacio de Sobolev* $W^{k,p}(\Omega)$ como el conjunto de funciones tales que D^α existe en sentido débil y está en $L^p(\Omega)$ para todo multiíndice $|\alpha| \leq k$.

Definición A.3.9. Sea una función con derivadas débiles hasta orden $|\alpha| = k$, definimos la norma Sobolev $\|u\|_{W^{k,p}(\Omega)}$ como

1. Para $p \in [1, \infty)$
$$\|u\|_{W^{k,p}(\Omega)} := \left(\sum_{|\alpha| \leq k} \|u\|^p_{L^p(\Omega)} \right)^{1/p}.$$
2. Para $p = \infty$
$$\|u\|_{W^{k,\infty}(\Omega)} := \sum_{|\alpha| \leq k} \|u\|_{L^\infty(\Omega)}.$$

Cuando $p = 2$, usaremos la notación $H^k(\Omega) := W^{k,2}(\Omega)$.

Nota 6. Por convención $H^0(\Omega) := L^2(\Omega)$.

Teorema A.3.1. *Los espacios de Sobolev $W^{k,p}(\Omega)$ son espacios de Banach con la norma Sobolev, más aún $H^k(\Omega)$ es un espacio de Hilbert separable con producto interior*
$$\langle u, v \rangle_{H^k(\Omega)} = \sum_{|\alpha| \leq k} \langle D^\alpha u, D^\alpha v \rangle_{L^2(\Omega)}.$$

Defimos ahora el siguiente espacio.

Definición A.3.10. Definimos el espacio $H^k_0(\Omega)$ como la cerradura de $C^\infty_c(\Omega)$ con la norma Sobolev de $H^k(\Omega)$.

Claramente por construcción $H^k_0(\Omega) \subset H^k(\Omega)$, por lo que puede considerarse a dicho espacio como los elementos de $H^k(\Omega)$ con traza cero.

Como observación final notemos que al ser $H^k_0(\Omega)$ un subespacio cerrado con la norma Sobolev de uno completo, entonces es un espacio de Hilbert. Sin embargo se puede definir una nueva norma sobre este espacio, la cual se conoce como la *norma de energía*; dicha norma la definimos por la expresión A.2.

$$\|v\|_{H^k_0(\Omega)} = \left(\sum_{|\alpha| = k} \|D^\alpha v\|^2 \right)^{1/2}. \tag{A.2}$$

A.4. Espacios de Sobolev para problemas de evolución

Dado que uno de los casos de estudio comprendidos en este trabajo fue la ecuación de calor, es necesario entrar en los problemas que dependen del tiempo, también conocidos como de

evolución. Debido a que estamos interesados en ecuaciones parabólicas, la teoría desarrollada en la sección anterior sirve de base para definir los espacios de Sobolev para ecuaciones parabólicas y sobre las cuales se define la solución.

Empezaremos por redefinir los espacios $L^p(\Omega)$ en sentido de evolución.

Definición A.4.1. Sea X un espacio de Banach, y sean las funciones medibles $u : [0,T] \to X$. Definimos las normas

1. Para $p \in [1,\infty)$ como
$$\|u\|_{L^p(0,T;X)} := \left(\int_0^T \|u(t)\|_X^p dt\right)$$
2. Para $p = \infty$ como
$$\|u\|_{L^p(0,T;X)} := ess\sup_{[0,T]} \|u(t)\|_X$$

Es con dichas normas definimos el siguiente espacio.

Definición A.4.2. Sea $u : [0,T] \to X$ medible con X espacio de Banach. Definimos el espacio

$$L^p(0,T;X) := \{u : [0,T] \to X \mid \|u\|_{L^p(0,T;X)} < \infty\}$$

Nuestro segundo espacio de interés es el de lo espacios que generalizan a las funciones que mapean del tiempo a funciones continuas. Dicho espacio se define como

$$C^0(0,T;X) := \{u : [0,T] \to X \mid u(t) \text{ continua, } \max_{[0,T]} \|u\|_X < \infty\}.$$

Y finalmente mencionaremos la versión temporal de los espacios de Sobolev. Para ello primero definiremos la derivada débil en sentido temporal.

Definición A.4.3. Sea X un espacio de Banach y sea u una función en $L^1(0,T;X)$. Llamaremos derivada débil a la función $v \in L^1(0,T;X)$ tal que

$$\int_0^T \phi'(t)\mathbf{u}(t)dt = -\int_0^T \phi(t)\mathbf{v}(t)dt, \qquad \forall \phi(t) \in C_c^\infty(0,T).$$

A dicha derivada la denotaremos por $\dfrac{d}{dt}u(t) := v$.

Es con esta definición de derivada que podemos definir algunas normas necesarias para construir nuestro espacio de interés.

Definición A.4.4. Sea $u \in L^p(0,T;X)$ y sea $\dfrac{d}{dt}u(t)$ su derivada débil. Definimos las normas

1. Para $p \in [1,\infty)$

$$\|u\|_{W^{1,p}(0,T;X)} := \int_0^T \left(\|u\|_X^p + \left\| \frac{d}{dt}u \right\|_X^p dt \right)^{1/p}$$

2. Para $p = \infty$

$$\|u\|_{W^{1,p}(0,T;X)} := ess \sup_{[0,T]} \left(\|u\|_X + \left\| \frac{d}{dt}u \right\|_X \right)$$

Con dichas normas ya es posible definir el siguiente espacio de Sobolev.

$$W^{1,p}(0,T;X) := \{u \in L^p(0,T;X) \mid \|u\|_{W^{1,p}(0,T;X)} < \infty\}.$$

Apéndice B

Elementos de probabilidad en espacios de Hilbert

En esta sección abordaremos los elementos de probabilidad usados en espacios de Hilbert sobre la tripleta $(\mathcal{H}, \mathfrak{H}, \mathbb{P})$, donde $\mathcal{H}$ es el espacio de Hilbert (nuestro espacio de medida), $\mathfrak{H}$ es una σ-álgebra sobre $\mathcal{H}$ y la medida de probabilidad $\mathbb{P}$. Para consultar las pruebas de este apéndice, así como profundizar en propiedades asociadas a estos espacios, puede consultar Stuart [36].

Empezaremos por definir las propiedades que caracterizan algunas distribuciones, en especial las medidas gaussianas, las cuales son de nuestro interés.

Definición B.0.1. Definimos la *media* o la *esperanza* de la medida $\mathbb{P}$ como el elemento $m \in \mathcal{H}$ tal que cumple

$$\langle u, m\rangle_{\mathcal{H}} = \int_{\mathcal{H}} \langle u, w\rangle_{\mathcal{H}} \mathbb{P}(dw), \qquad \forall u \in \mathcal{H}. \tag{B.1}$$

Definición B.0.2. Diremos que la medida $\mathbb{P}$ tiene un *operador de covarianzas* asociado, si existe $\mathcal{C} : \mathcal{H} \to \mathcal{H}$ lineal, simétrico y positivo tal que

$$\langle \mathcal{C}(u), v\rangle_{\mathcal{H}} = \int_{\mathcal{H}} \langle w - m, u\rangle_{\mathcal{H}} \langle w - m, v\rangle_{\mathcal{H}} \mathbb{P}(dw), \qquad \forall u, v \in \mathcal{H}. \tag{B.2}$$

Más aún, si $\mathcal{C}$ es invertible, llamaremos precisión al operador $\mathcal{C}^{-1}$.

Otro concepto importante es el de la *traza* asociada a un operador lineal. Si bien en dimensión finita es muy simple, en dimensión infinita este concepto está definido de la siguiente manera.

Definición B.0.3. Sea $\mathcal{K} : \mathcal{H} \to \mathcal{H}$ un operador lineal y sea $\{\phi_k\}_{k\in\mathbb{K}}$ una base ortonormal de eigenfunciones, entonces llamaremos:

1. Traza de $\mathcal{K}$ a
$$tr(\mathcal{K}) := \sum_{k=1}^{\infty} \langle \mathcal{K}\phi_k, \phi_k \rangle_{\mathcal{H}}. \tag{B.3}$$
2. Operador nuclear a $\mathcal{K}$ si
$$tr(\mathcal{K}) < \infty. \tag{B.4}$$

Es sabido que algunas medidas de probabilidad se pueden construir a partir de otra, esto generalmente es a partir de aquellos conjuntos donde ambas medidas se anulan, esta caracterización se define a continuación.

Definición B.0.4. Sean $\mathbb{P}_1$ y $\mathbb{P}_2$ dos medidas de probabilidad en $\mathfrak{H}$, diremos que $\mathbb{P}_1$ es absolutamente continua respecto a $\mathbb{P}_2$ si para todo $A \in \mathfrak{H}$ tal que $\mathbb{P}_2(A) = 0$ se tiene que $\mathbb{P}_1(A) = 0$. Denotaremos a esta propiedad como $\mathbb{P}_2 >> \mathbb{P}_1$.

Con esto ya es posible definir una medida por medio de otra con el siguiente teorema.

Teorema B.0.1 (Radon-Nikodym)**.** *Sean las tripletas $(\mathcal{H}, \mathfrak{H}, \mathbb{P}_1)$ y $(\mathcal{H}, \mathfrak{H}, \mathbb{P}_2)$, con $\mathbb{P}_2 >> \mathbb{P}_1$, entonces existe la función medible $d\mathbb{P}_2/d\mathbb{P}_1$ respecto a $\mathbb{P}_1$ tal que*
$$\mathbb{P}_2(A) = \int_A \frac{d\mathbb{P}_2}{d\mathbb{P}_1}(z)\mathbb{P}_1(dz), \qquad \forall A \in \mathfrak{H}. \tag{B.5}$$

En particular estaremos interesados en trabajar con la generalización de las distribuciones gaussianas de dimensión finita; esto es posible por medio de la siguiente definición.

Definición B.0.5. Sea la tripleta $(\mathcal{H}, \mathfrak{H}, \mathbb{P})$. Diremos que $\mathbb{P}$ es una medida gaussiana si existen $m_u \in \mathbb{R}$ y $\sigma_u^2 \geq 0$, tales que la función dada por $\langle u, \cdot \rangle_{\mathcal{H}}$ se distribuye $\mathcal{N}(m_u, \sigma_u^2)$ para todo $u \in \mathcal{H}$. Más aún, denotaremos a dicha medida como $\mathbb{P} = \mathcal{N}(m, \mathcal{C})$.

Dentro de las medidas gaussianas es posible caracterizar las propiedades del operador de covarianzas y de precisión; empezaremos por hablar de los espacios de Hilbert que se pueden construir por medio de la precisión.

Definición B.0.6. Dada una medida gaussiana $\mathbb{P} = \mathcal{N}(m, \mathcal{C})$, llamaremos espacio de Cameron-Martin E sobre $\mathcal{H}$ como el dado por el conjunto

$$E = \bigcap_{j \in \mathbb{J}} \mathcal{H}_j, \tag{B.6}$$

donde los $\mathcal{H}_j$ son subespacios vectoriales de $\mathcal{H}$ y que cumplen que $\mathbb{P}(\mathcal{H}_j) = 1$.

Nota 7. Observemos que el espacio de Cameron-Martin también es subespacio vectorial de $\mathcal{H}$.

La relación entre el espacio de Cameron-Martin y una medida gaussiana está dada por el siguiente teorema.

Teorema B.0.2. *Sea la medida gaussiana $\mathbb{P} = \mathcal{N}(0, \mathcal{C})$ y E el espacio de Cameron-Martin sobre $\mathcal{H}$. Entonces dicho espacio es de Hilbert y se caracteriza como*

$$E = Im(\mathcal{C}^{1/2}), \tag{B.7}$$

más aún, el producto interior asociado a este espacio es de la forma

$$\langle \cdot, \cdot \rangle_{\mathcal{C}} := \langle \mathcal{C}^{-1/2} \cdot, \mathcal{C}^{-1/2} \cdot \rangle_{\mathcal{H}} = \langle \mathcal{C}^{-1} \cdot, \cdot \rangle_{\mathcal{H}}. \tag{B.8}$$

La importancia del espacio de Cameron-Martin E sobre una medida gaussiana está dada por la regularidad de este espacio sobre $\mathcal{H}$ y su propiedad de invarianza bajo traslaciones. Esto se caracteriza por la siguiente proposición.

Proposición B.0.3. *Dada la medida gaussiana $\mathbb{P} = \mathcal{N}(0, \mathcal{C})$ y E su espacio de Cameron-Martin sobre $\mathcal{H}$. Entonces se cumple lo siguiente:*

1. *E está encajado continuamente sobre $\mathcal{H}$, es decir, existe $C > 0$ tal que*

$$\|u\|_{\mathcal{H}} \leq C \|u\|_E. \tag{B.9}$$

2. *Sea $\mathbb{P}_m = \mathcal{N}(m, \mathcal{C})$, entonces $\mathbb{P}_m << \mathbb{P}$ y $\mathbb{P} << \mathbb{P}_m$ si y solo si $m \in E$.*

El teorema anterior nos indica que una medida gaussiana con media no nula está bien definida si la media pertenece a E, lo cual es importante para poder al menos definir de manera correcta la distribución a priori en estadística bayesiana . Si bien el espacio de Cameron-Martin nos dice en qué espacio debe estar la media, es necesario que el operador de Covarianzas esté bien definido. Consideremos entonces las siguientes hipótesis.

Hipótesis B.0.4. .
Sea $\mathcal{K}$ un operador denso sobre $\mathcal{H} \subset L^2(D, \mathbb{R}^d)$ que satisface lo siguiente:

1. *$\mathcal{K}$ es autoadjunto, definido positivo y con inversa $\mathcal{K}^{-1}$.*
2. *El conjunto de eigenfunciones $\{\phi_k\}_{k\in\mathbb{K}}$ de $\mathcal{K}$ con índices $\mathbb{K} \subset \mathbb{Z}^d \setminus \{0\}$ forman una base ortonormal en $\mathcal{H}$.*
3. *El conjunto de eigenvalores $\{\lambda_k\}_{k\in\mathbb{K}}$ asociados a $\{\phi_k\}_{k\in\mathbb{K}}$ están acotados por*

$$C_1 \leq \frac{\lambda_k}{|k|^2} \leq C_2,$$

   para algún par $C_1, C_2 > 0$.
4. *Existe $C > 0$ tal que*

$$\sup_{k\in\mathbb{K}} \left(\|\phi_k\|_{L^\infty} + \frac{1}{k}\|D\phi_k\|_{L^\infty} \right) \leq C.$$

Con estas hipótesis podemos construir el operador de covarianzas apropiado, así como los subespacios de $\mathcal{H}$ asociados a este operador. Por ello, definimos los siguientes conceptos.

Definición B.0.7. Sea $\mathcal{K}$ un operador que cumple las hipótesis 1 y 2

1. El operador faccional $\mathcal{K}^\alpha$ con $\alpha \in \mathbb{R}$ es aquel que cumple

$$\mathcal{K}^\alpha u := \sum_{k\in\mathbb{K}} \lambda_k^\alpha \langle u, \phi_k \rangle_{\mathcal{H}} \phi_k, \qquad \forall u \in \mathcal{H}.$$

2. El espacio de Hilbert $\mathcal{H}^s$ con $s \in \mathbb{R}$ está dado por

$$\mathcal{H}^s := \left\{ u \in \mathcal{H} : \sum_{k\in\mathbb{K}} \lambda_k^s \left| \langle u, \phi_k \rangle_{\mathcal{H}} \right|^2 < \infty \right\},$$

   más aún, $\mathcal{H}^s$ tiene norma dada por

$$\|u\|_s^2 = \sum_{k\in\mathbb{K}} \lambda_k^s \left| \langle u, \phi_k \rangle_{\mathcal{H}} \right|^2. \tag{B.10}$$

La importancia de los espacios $\mathcal{H}^s$ radica en su contención sobre $\mathcal{H}$ y cómo construirlos. Esta relación está dada por la siguiente proposición.

Proposición B.0.5. *Sea $\mathcal{K}$ que satisface las hipótesis B.0.4. Entonces*

1. *Si* $s > 0$*, entonces* $\mathcal{H}^s \subset \mathcal{H}$.
2. $\mathcal{H}^s = Dom(\mathcal{K}^{s/2})$.

El siguiente teorema nos indica cuál es el operador de covarianzas apropiado para $\mathcal{H}$ y la regularidad del espacio.

Proposición B.0.6. *Sea* $\mathcal{K}$ *que satisface las hipótesis B.0.4 y sea* $\alpha > d/2$*. Entonces* $\mathcal{C} = \mathcal{K}^{-\alpha}$ *está bien definido (*$Im(\mathcal{C}) \subset L^2(D, \mathbb{R}^d)$*), más aún, dado* $u \sim \mathbb{P} = \mathcal{N}(0, \mathcal{C})$*, tenemos que* u *es* s*-Hölder continua para todo* $s < \min\{1, \alpha - d/2\}$.

Claramente este operador de covarianzas ofrece continuidad en sentido Hölder para las variables aleatorias de $\mathbb{P}$. Por otro lado, también es posible decir más sobre el espacio en el que las variables aleatorias de $\mathbb{P}$ se encuentran.

Recordemos que $\mathcal{H} \subset L^2(D, \mathbb{R}^d)$, y en especial nos iteresan problemas definidos en espacios de Sobolev H. Puede probarse que $\mathcal{H}^s \subset H^s$, por ello, el siguiente teorema es de suma importancia cuando usemos operadores como las potencias del Laplaciano.

Proposición B.0.7. *Consideremos* $\mathcal{K}$ *que cumple B.0.4 y sea* $u \sim \mathbb{P} = \mathcal{N}(0, \mathcal{K}^{-\alpha})$ *con* $\alpha > d/2$*. Entonces* $u \in \mathcal{H}^s$ *c.s. para todo* $s \in [0, \alpha - d/2]$.

Bibliografía

[Sci] Scipy manual. `https://docs.scipy.org/doc/`. 23

[2] Alexanderian, A., Petra, N., Stadler, G., and Ghattas, O. (2014). A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized l_0-sparsification. *Journal on Scientific Computing*, 36(5):A2122–A2148. 2, 3, 47, 50, 52, 55, 64, 65

[3] Allaire, G. and Kaber, S. M. (2008). *Numerical linear algebra*, volume 55. Springer. 22

[4] Atkinson, A., Donev, A., and Tobias, R. (2007). *Optimum Experimental Designs, with SAS*. 47

[5] Avron, H. and Toledo, S. (2011). Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM (JACM)*, 58(2):1–34. 51

[6] Bui-Thanh, T. and Nguyen, Q. P. (2016). FEM-based discretization-invariant MCMC methods for PDE-constrained Bayesian inverse problems. *Inverse Problems and Imaging*, 10(4):943–975. 2, 37, 38, 62

[7] Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208. 65

[8] Chaloner, K. and Verdinelli, I. (1995). Bayesian Experimental Design: A review. *Statistical Science*, pages 273–304. 47

[9] Cotter, S. L., Dashti, M., Robinson, J. C., and Stuart, A. M. (2009). Bayesian inverse problems for functions and applications to fluid mechanics. *Inverse problems*, 25(11):115008. 1, 2

[10] Evans, L. C. (1998). *Partial differential equations. Graduate studies in mathematics*, volume 2. American mathematical society. 83

[11] Flath, H. P., Wilcox, L. C., Akçelik, V., Hill, J., van Bloemen Waanders, B., and Ghattas, O. (2011). Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations. *SIAM Journal on Scientific Computing*, 33(1):407–432. 2

[12] Förstner, W. and Moonen, B. (2003). A metric for covariance matrices. In *Geodesy-the Challenge of the 3rd Millennium*, pages 299–309. Springer. 42

[13] Golan, J. (2007). *The Linear Algebra a Beginning Graduate Student Ought to Know*, volume 27. Springer. 39

[14] Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU press. 44

[15] Griffiths, D. F. and Higham, D. J. (2010). *Numerical methods for ordinary differential equations: initial value problems*. Springer Science & Business Media. 21

[16] Halton, J. H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, 2(1):84–90. 63

[17] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media. 55

[18] Isakov, V. (2006). *Inverse problems for partial differential equations*, volume 127. Springer. 1, 26

[19] Kaipio, J. and Somersalo, E. (2006). *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media. 1, 27, 30

[20] Koval, K., Alexanderian, A., and Stadler, G. (2020). Optimal experimental design under irreducible uncertainty for linear inverse problems governed by PDEs. *Inverse Problems*. 2

[21] Langtangen, H. P. and Logg, A. (2016). *Solving PDEs in Python: The FEniCS Tutorial I*, volume 1. Springer. 64

[22] Larson, M. G. and Bengzon, F. (2013). *The finite element method: theory, implementation, and applications*, volume 10. Springer Science & Business Media. 13, 16, 22, 71

[23] Larsson, S. and Thomée, V. (2008). *Partial differential equations with numerical methods*, volume 45. Springer Science & Business Media. 13, 17, 18, 22

[24] Le, E. B., Myers, A., Bui-Thanh, T., and Nguyen, Q. P. (2017). A data-scalable randomized misfit approach for solving large-scale PDE-constrained inverse problems. *Inverse Problems*, 33(6):065003. 2

[25] Maz'ya, V. G. and Shaposhnikova, T. O. (1999). *Jacques Hadamard: a universal mathematician*. Number 14. American Mathematical Soc. 5

[26] Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media. 50

[27] Pázman, A. (1986). *Foundations of Optimum Experimental Design*, volume 14. Springer. 2, 47, 50

[28] Pronzato, L. and Pázman, A. (2013). Design of Experiments in Nonlinear Models: Asymptotic Normality, Optimality Criteria and Small-Sample Properties. 47

[29] Quarteroni, A. and Valli, A. (2008). *Numerical approximation of partial differential equations*, volume 23. Springer Science & Business Media. 8, 23, 83

[30] Renardy, M. and Rogers, R. C. (2006). *An introduction to partial differential equations*, volume 13. Springer Science & Business Media. 36, 83

[31] Rencher, A. C. and Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons. 51

[32] Saad, Y. (2003). *Iterative methods for sparse linear systems*, volume 82. SIAM. 22

[33] Salsa, S. (2016). *Partial differential equations in action: from modelling to theory*, volume 99. Springer. 7, 36, 83

[34] Simpson, D. P. (2008). *Krylov subspace methods for approximating functions of symmetric positive definite matrices with applications to applied statistics and anomalous diffusion*. PhD thesis, Queensland University of Technology. 2, 37

[35] Spantini, A., Solonen, A., Cui, T., Martin, J., Tenorio, L., and Marzouk, Y. (2015). Optimal low-rank approximations of Bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 37(6):A2451–A2487. 2, 39, 42, 43, 44

[36] Stuart, A. M. (2010). Inverse problems: a Bayesian perspective. *Acta numerica*, 19:451–559. 2, 28, 31, 33, 34, 35, 36, 91

[37] Thomée, V. (2006). *Galerkin finite element methods for parabolic problems*, volume 25. Springer Science & Business Media, 2 edition.

[38] Van der Pol, B. (1926). LXXXVIII. On "relaxation-oscillations". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):978–992. 24

[39] Virieux, J. (1986). P-SV wave propagation in heterogeneous media; velocity-stress finite-difference method. *Geophysics*, 51(4):889–901. 25

[40] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3):261–272. 64