ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

SOLUCIÓN DE UN PROBLEMA DE ASIMILACIÓN DE DATOS Y UN PROBLEMA DE LOCALIZACIÓN ÓPTIMA MEDIANTE MÉTODOS DE OPTIMIZACIÓN BINIVEL

TRABAJO PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGÍSTER EN OPTIMIZACIÓN MATEMÁTICA

TESIS

PAULA MONSERRATTE CASTRO CASTRO
paula.mc4@gmail.com

Director: DR. JUAN CARLOS DE LOS REYES BUENO
juan.delosreyes@epn.edu.ec

QUITO, MAYO 2017

DECLARACIÓN

Yo PAULA MONSERRATTE CASTRO CASTRO, declaro bajo juramento que el trabajo aquí escrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual, correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normatividad institucional vigente.

Paula Monserratte Castro Castro

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por PAULA MONSERRATTE CASTRO CASTRO, bajo mi supervisión.

Dr. Juan Carlos De los Reyes Bueno
Director del Trabajo

AGRADECIMIENTOS

Al Dr. Juan Carlos De los Reyes, por el tiempo invertido en la supervisión de este trabajo. A mi familia por su apoyo. A los colegas del Centro de Modelización Matemática por la oportunidad de ser parte de este grupo de trabajo.

Este trabajo fue parcialmente financiado por: el Instituto Nacional de Meteorología e Hidrología dentro del proyecto "*Sistema de Pronóstico del Clima y el Tiempo para todo el territorio ecuatoriano: Modelización numérica y estadística - fase dos*", la Escuela Politécnica Nacional dentro del proyecto "*Flujos de materiales viscoplásticos en la industria alimenticia: modelización matemática, simulación numérica y optimización*" y al programa de cooperación científico-tecnológica MATH-AmSud.

DEDICATORIA

A mis padres, Carlos y Paulina.

Índice general

Índice de cuadros

Índice de figuras

Resumen

Los problemas de asimilación de datos han sido ampliamente estudiados en la predicción numérica del tiempo y del clima, como una técnica para la reconstrucción del estado inicial de la atmósfera. Son importantes ya que mientras más precisa sea la reconstrucción de esta condición inicial, los pronósticos meteorológicos obtenidos serán más exactos para cualquier instante de tiempo.

Tomando este problema como motivación, la meta de este proyecto es encontrar la solución de un problema de localización óptima de ubicaciones en una ecuación parabólica, de tal manera que se reconstruya la condición inicial de la misma, es decir se resuelva un problema de asimilación de datos. Se consideró un modelo de optimización a dos niveles, donde el nivel inferior se encarga de la reconstrucción de la condición inicial del estado del sistema, es decir, de encontrar la solución a un problema de asimilación de datos, mientras el nivel superior resuelve el problema de localización óptima de observaciones. Además del modelo continuo de optimización, se consideró la utilización de la técnica de conjuntos de entrenamiento, ampliamente utilizada en Machine Learning, la cual consiste en predecir información de un dato desconocido basándose en la información aprendida de una muestra de datos conocidos. En nuestro caso, el conjunto de entrenamiento se construyó con simulaciones de la condición inicial y observaciones del estado del sistema parabólico estudiado. Lo que se obtuvo es un vector de localizaciones en promedio óptimo para todos los pares de entrenamiento considerados.

La resolución numérica del problema fue desarrollada también en dos niveles. El problema de asimilación de datos se resolvió utilizando el algoritmo BFGS, mientras que para el problema localización óptima se emplearon métodos Quasi-Newton proyectados, específicamente el método BFGS proyectado mediante la estimación de conjuntos $\epsilon-$activos. Se consideró la utilización de métodos de segundo orden debido a su velocidad de convergencia, ya que el sistema de optimalidad del problema del nivel inferior es la restricción del problema del nivel superior.

Abstract

Data assimilation problems have been widely studied in numerical weather prediction as a technique for the reconstruction of the atmosphere's initial condition. They are important because the more accurate the initial conditions are, the better quality of the forecast is achieved for any period of time.

Taking this problem as motivation, our goal in this project is finding the optimal placements of the locations of a data assimilation problem, represented by a parabolic equation. We worked in a bilevel optimization problem where the inner level solves the data assimilation problem, and the upper level solves the optimal placement problem. Moreover, to get a more robust result, we worked with training sets. It is a widely used technic in Machine Learning and tries to predict information from unknown data based on information learned from a sample of known data. In our case, the training set was formed by simulations of the initial condition and observations of the parabolic system's state. What we got was a optimal placement vector which is in average optimal for every training pair considered.

The numerical solution was also worked in two levels. The inner problem was solved by using the BFGS method while the upper level used the BFGS projected method through an estimation of $\epsilon-$active sets.

Capítulo 1

Introducción

De manera general, el proceso de asimilación de datos puede ser descrito como el procedimiento mediante el cual se analiza toda la información disponible tal como observaciones, fenómenos físicos, entre otros para tratar de estimar de mejor manera el estado de un sistema que evoluciona en el tiempo. Dependiendo del problema particular que se esté tratando, el proceso de asimilación de datos podría describir el estado del sistema en un tiempo determinado, explicar su evolución en un intervalo de tiempo u obtener aproximaciones de la condición inicial del mismo [Raschke and Jacob, 2013]. Es por esto último precisamente que este tipo de problemas son ampliamente estudiados en la predicción numérica del tiempo y del clima, ya que si se conociera la condición inicial exacta del estado de la atmósfera, sería posible obtener un pronóstico meteorológico preciso para cualquier instante de tiempo [Warner, 2010]. Haciendo de la estimación de la condición inicial del sistema que simula la atmósfera el objetivo principal del proceso de asimilación de datos en esta área [Kalnay, 2003].

Existen varios enfoques para la resolución del problema de asimilación de datos: la interpolación óptima, los filtros de Kalman, el enfoque variacional, entre otros. Este último consiste en la resolución de un problema de control óptimo y se divide en dos tipos de problemas, el $3D-$VAR y el $4D-$VAR. El primero considera observaciones en un solo instante del tiempo y el segundo, que es una extensión del primero, considera observaciones distribuidas en un intervalo de tiempo $[t_0, t_n]$[Kalnay, 2003].

En el contexto meteorológico, los datos necesarios para la asimilación pueden ser recopilados de diversas fuentes, entre ellas se tiene los provenientes de radio sondeos, imágenes satelitales, estaciones de medición, entre otras. Sin embargo, debido al costo o a la dificultad en la instalación de los equipos que permiten obtener estos datos se busca poder localizarlos de manera óptima, de tal forma que la información obtenida de ellos sea útil para el proceso de asimilación y con esto a la reconstrucción de la condición inicial del estado de la atmósfera.

Tomando en cuenta la resolución de este problema común en meteorología como motivación, lo que se busca en este proyecto es resolver un problema de localización óptima de observaciones de tal manera que se logre la reconstrucción de la condición inicial de un problema parabólico. Para ello, se considerará un modelo de optimización a dos niveles donde el nivel inferior se encargará de la reconstrucción de la condición inicial del estado del sistema, es decir, de encontrar la solución a un problema de asimilación de datos, mientras que el nivel superior resolverá el problema de localización óptima de observaciones.

Al tratarse de un problema de localización óptima, podría ser abordado con técnicas combinatorias y de programación lineal entera. En efecto, trabajos relacionados con la ubicación óptima de sensores han sido desarrollados con este enfoque. En [Krause and Guestrin, 2007] se hace uso de la idea intuitiva de la selección de objetos, donde se establece que añadir observaciones ayuda más si hasta el momento se han tomado pocas observaciones y por el contrario ayuda menos si ya se cuenta con una gran cantidad de observaciones. En dicho trabajo se define formalmente el fenómeno descrito bajo el concepto de *submodularidad*. A breves rasgos, en [Krause and Guestrin, 2007] la submodularidad queda definida como sigue: una función a valores reales F definida sobre un conjunto $\mathcal{A} \subset \mathcal{V}$ finito, donde $\mathcal{V}$ representa el conjunto de todas las posibles ubicaciones, se dice submodular si para todo $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ y para todo $s \in \mathcal{V} \setminus \mathcal{B}$ se tiene que

$$F(\mathcal{A} \cup \{s\}) - F(\mathcal{A}) \geq F(\mathcal{B} \cup \{s\}) - F(\mathcal{B}).$$

Aquí se establece además que muchos problemas de selección de observaciones, tal como el problema de hallar el óptimo número de sensores, pueden ser reducidos al problema de maximizar una función de conjunto submodular sujeto a algunas restricciones adicionales. Además se establece que aunque en general los problemas de optimización submodular con restricciones son problemas NP-difíciles son una forma de atacar al problema sin la necesidad de recurrir heurísticas.

Existe también una gran cantidad de trabajos que atacan al problema con técnicas de optimización continua, entre ellos podemos citar los trabajos de [Alexanderian et al., 2016], [Alexanderian et al., 2014], [Kang and Xu, 2012], entre otros. [Kang and Xu, 2012] trata directamente un problema de localización optima de sensores para la asimilación de datos. Aquí, se resalta el hecho de que el impacto que tengan las observaciones dependerá no solo del sistema que se use para la asimilación ($4D-$VAR, $3D-$VAR) sino también de la información contenida en cada observación a ser asimilada. Tanto la teoría como los algoritmos desarrollados en [Kang and Xu, 2012] se basan en el concepto de la *observabilidad* que a breves rasgos puede ser descrita como una medida de calidad de la información obtenida de un sensor. La observabilidad puede ser calculada

numéricamente basándose en la dinámica del sistema estudiado y como resultado se obtiene un funcional de costo para optimizar la locación de sensores.

En [Alexanderian et al., 2014] y en [Alexanderian et al., 2016] se considera un enfoque Bayesiano lineal y no lineal respectivamente, para determinar la ubicación óptima de sensores de los cuales se extraerá datos para experimentaciones. En estos trabajos se observa que la solución numérica de un problema Bayesiano inverso es un subproblema del diseño óptimo de experimentos, OED por sus siglas en inglés, y se lo trata desde esta perspectiva. En [Alexanderian et al., 2014] se considera un número finito de ubicaciones como candidatas para la colocación de sensores, a cada uno de estos candidatos se les asigna un valor no negativo que *pesa* las observaciones recolectadas por ese sensor. Mediante la utilización exhaustiva de la estructura del problema se observa que buscar un diseño óptimo es equivalente a escoger un vector de pesos con valores binarios, donde un peso de uno corresponde a que se tome la ubicación para la colocación de un sensor y cero caso contrario.

En este trabajo se propone la utilización de la técnica de conjuntos de entrenamiento, ampliamente utilizada en Machine Learning, la cual consiste en predecir información de un dato desconocido basándose en la información aprendida de una muestra de datos conocidos. En nuestro caso el conjunto de entrenamiento estará formado por simulaciones de la condición inicial y observaciones del estado del sistema parabólico estudiado respectivamente. Lo que se busca obtener es un vector de localizaciones que en promedio sea óptimo para todos los pares de entrenamiento considerados. Con este fin, se considera un modelo de optimización a dos niveles donde el nivel inferior resuelve el problema de asimilación de datos y el nivel superior el problema de localización óptima de observaciones. La ventaja de abordar el problema de localización de esta manera es que no se realizan supuestos sobre la distribución de probabilidad que siguen los datos a ser asimilados, como es el caso de las estimaciones Bayesianas. Tampoco nos enfrentamos a un problema del tipo NP - difícil al cual deberíamos hacer frente de formular el problema con técnicas combinatorias y de programación lineal entera.

Finalmente, la resolución numérica del problema también se desarrollará en dos niveles. Para el problema interior, es decir el de asimilación de datos, se utilizará el método BFGS. Se considera la utilización de un método de segundo orden debido a su velocidad de convergencia, ya que el sistema de optimalidad del problema del nivel inferior será la restricción del problema del nivel superior, es decir del problema de localización óptima. Para la resolución numérica del problema del nivel superior se emplearán métodos Quasi Newton proyectados, específicamente el método BFGS proyectado mediante la estimación de conjuntos $\epsilon-$activos.

La organización del trabajo es la siguiente: en el capítulo dos se abordará el problema de asimilación de datos, empezando con el planteamiento del problema con el enfoque 4D-VAR con operadores de covarianza iguales al operador identidad. Se demostrará la existencia y unicidad de la solución del problema de optimización, así como que el sistema de optimalidad está correctamente definido al demostrar la existencia y unicidad de la solución de la ecuación adjunta, que como se verá, tiene medidas regulares de Borel en el lado derecho.

El tercer capítulo abordará el estudio del problema de localización óptima de observaciones como un problema de optimización a dos niveles. Como una primera aproximación, centraremos nuestra atención a la resolución del problema cuando se trabaja únicamente con un par de entrenamiento, se presentará además el método de segundo orden utilizado para la resolución numérica del problema binivel, el BFGS proyectado. Como una subsección de este capítulo trabajaremos con un número, mayor a uno, fijo de pares de entrenamiento. Se incluye además la experimentación numérica realizada donde se muestra la reconstrucción de la condición inicial simulada y el estado observado. Finalmente, en el último capítulo se expondrán las conclusiones a las que se llegó de la realización del trabajo.

Capítulo 2

Problema de Asimilación de Datos

1. Planteamiento del problema

Para resolver el problema de asimilación de datos se utilizará el enfoque variacional mediante la resolución del problema $4D-$VAR que considera diferentes observaciones distribuidas en un intervalo de tiempo $[t_0, t_l]$ [Kalnay, 2003]. Matemáticamente resolver el problema $4D-$VAR significa resolver el siguiente problema de optimización

$$\begin{aligned} \min_u J(y,u) &= \frac{1}{2}\sum_{i=1}^{l}\left[H(y(t_i)-z_o(t_i))\right]^T R_i^{-1}\left[H(y(t_i)-z_o(t_i))\right] \\ &\quad +\frac{1}{2}\left[u-u^b\right]^T B^{-1}\left[u-u^b\right] \\ \text{sujeto a:} & \\ y(t_l) &= M(y(t_0)) \qquad \text{(Sistema de EDP's)} \\ y(t_0) &= u \qquad \text{(Condición inicial).} \end{aligned} \tag{2.1}$$

En la formulación anterior z_o representa el estado observado, u^b la información previa o *background*. H es un operador de observación, que transforma las variables del modelo en variables observables. Para todo $i = 1, \ldots, l$, los R_i representan matrices de covarianza respecto a los errores de las observaciones y B a la matriz de covarianza de los errores del background. Dependiendo si el análisis se realiza en dimensión finita o infinita se trabajará con matrices u operadores de covarianza respectivamente, para el caso de dimensión infinita los productos matriciales de la formulación (2.1) serán remplazados por productos internos definidos en espacios funcionales convenientes.

Para pasar de la formulación general de un problema $4D-$VAR al problema específico de asimilación de datos con el cual se trabajará, se fijan los operadores de covarianza del background y de los errores iguales al operador identidad, se toma además

$u^b \in L^2(\Omega)$ igual a cero. El operador de observaciones H viene dado por el vector en $\mathbb{R}^{\mu_E}$ de localizaciones $w = (w_k)$, $k = 1, \ldots, \mu_E$ de entradas binarias, donde μ_E representa el número total de puntos en el dominio espacial en el que se este trabajando. Los w_k toman el valor de uno si la posición x_k es seleccionada y cero caso contrario. El operador de observaciones H permite además la evaluación de las variables en entradas puntuales en el espacio. El sistema de EDP's que representa la restricción del problema corresponde a un problema parabólico lineal y continuo, específicamente la ecuación del calor con condiciones de Dirichlet homogéneas. Así, el problema de asimilación de datos en concreto que se va a resolver es el siguiente:

$$\min_u J(y,u) = \frac{1}{2}\sum_k \sum_i w_k \left(y(x_k,t_i) - z_o(x_k,t_i)\right)^2 + \frac{\alpha}{2}\int_\Omega |u|^2 \tag{2.2}$$

$$\text{s.a} \begin{cases} \dfrac{\partial y}{\partial t} - \Delta y = 0 & \text{en } Q = \Omega \times]0,T[\\ y = 0 & \text{en } \Sigma = \Gamma \times]0,T[\\ y(0) = u & \text{en } \Omega. \end{cases} \tag{2.3}$$

2. Existencia de solución del problema de optimización

Sean $\Omega \subset \mathbb{R}^n$ un dominio acotado tipo Lipschitz, con frontera Γ. Definimos $Q = \Omega \times (0,T)$ y $\Sigma = \Gamma \times (0,T)$ con un número real $T > 0$ fijo. Verifiquemos en primer lugar que la ecuación de estado del problema de asimilación de datos (2.3) está bien definida, es decir, que para cualquier $u = u(x) \in L^2(\Omega)$ existe una única solución débil $y = y(x,t)$ en un espacio funcional adecuado.

DEFINICIÓN 2.1. *Una función* $\mathbf{z} \in L^2(0,T;H_0^1(\Omega))$*, con* $\mathbf{z}' \in L^2(0,T;H^{-1}(\Omega))$ *es solución débil de* (2.3) *si:*

(i) $\langle \mathbf{z}', v\rangle_{H^{-1},H_0^1} + B[\mathbf{z},v;t] = 0, \quad \forall v \in H_0^1(\Omega)$ *y para c.t*,$0 \leq t \leq T$

(ii) $\mathbf{z}(0) = u.$

Donde

$$\begin{aligned} \mathbf{z} : [0,T] &\to H_0^1(\Omega), \\ t &\mapsto \mathbf{z}(t), \end{aligned}$$

tal que $[\mathbf{z}(t)](x) = z(x,t)$, donde B representa la forma bilineal dada por

$$B[\mathbf{z}, v; t] := \int_{\Omega} \sum_{i=1}^{n} z_{x_i} v_{x_i}$$

y donde

$$\mathbf{z}' = \frac{\partial \mathbf{z}}{\partial t}.$$

Notación 1. *Si* $\mathbf{z}$ *es tal que* $\mathbf{z} \in L^2(0,T; H_0^1(\Omega))$ *y además* $\mathbf{z}' \in L^2(0,T; H^{-1}(\Omega))$ *diremos que* $z \in W(0,T)$.

Notemos que $y \in W(0,T)$, solución de (2.3) verifica (i) y (ii). En efecto, tomando $v \in H_0^1(\Omega)$ una función test, multiplicándola por (2.3) expresado en términos de $\mathbf{y}$ e integrando se sigue que

$$\int_Q \mathbf{y}' v dx + \int_Q \nabla v . \nabla \mathbf{y} dx = 0$$

o equivalentemente

$$\left\langle \mathbf{y}', v \right\rangle_{H^{-1}, H_0^1} + B[\mathbf{y}, v; t] = 0$$

con lo cual obtiene (i). Por otro lado, para $x \in \Omega$ se tiene que $y(x,0) = u(x)$ con lo cual (ii) también se verifica. La última igualdad tiene sentido ya que $y \in W(0,T)$ es también elemento de $C([0,T]; L^2(\Omega))$ ([Evans, 1998], pp 287).

La existencia de una única $y \in W(0,T)$ solución débil de (2.3) queda garantizada mediante la utilización de métodos de Galerkin (ver [Evans, 1998],pp 356).

Se tiene además una estimación entre las normas de u y de y en sus respectivos espacios. Este resultado queda formalizado en el siguiente teorema, para la demostración del mismo referirse a ([Tröltzsch, 2010], pp. 150)

Teorema 1. *La solución débil del problema* (2.3) *satisface la estimación de la forma*

$$\| y \|_{W(0,T)} \leq c_w \| u \|_{L^2(\Omega)}, \text{ con } c_w > 0.$$

En otras palabras, la función solución que asigna $u \mapsto y$ define un operador lineal y continuo, notado G_0 que va de $L^2(\Omega)$ en $W(0,T)$, en particular en $C([0,T]; L^2(\Omega))$.

Ahora se probará que el problema de control óptimo (2.2) - (2.3) tiene solución, es

decir se demostrará la existencia de un control óptimo $\bar{u} \in L^2(\Omega)$ con $\bar{y} = y(\bar{u}) \in W(0,T)$ estado óptimo asociado.

De lo expuesto anteriormente, sabemos que para cualquier $u \in L^2(\Omega)$ existe una única solución débil $y \in W(0,T)$ para la ecuación de estado (2.3), que puede ser expresada como $y = G_0 u$, con G_0 un operador lineal y continuo de $L^2(\Omega)$ en $W(0,T)$. Consideremos el operador que representa la inyección canónica

$$E : W(0,T) \to L^2(0,T;L^2(\Omega))$$

también lineal y continuo que asigna a cada $y \in W(0,T)$ la misma función en $L^2(Q)$, con $L^2(0,T;L^2(\Omega)) \cong L^2(Q)$ y definiendo el operador S como la composición de G_0 con E, tenemos

$$\begin{aligned} S : L^2(\Omega) &\to L^2(Q) \\ u &\mapsto Su = y \end{aligned}$$

El operador S recibirá el nombre de operador control-estado. Notemos además que dicho operador es a su vez lineal y continuo por ser la composición de operadores lineales y continuos. Reescribiendo el funcional objetivo (2.2) tomando en cuenta el operador control-estado S obtenemos el funcional reducido

$$\min_u f(u) = \frac{1}{2} \sum_k \sum_i w_k \left(Su(x_k,t_i) - z_o(x_k,t_i) \right)^2 + \frac{\alpha}{2} \parallel u \parallel^2_{L^2(\Omega)} \tag{2.4}$$

Notemos que f así definida es convexa y continua. En efecto, para todo par (k,i) la función $(Su(.,.) - z_o(.,.))^2$ es convexa por ser cuadrática. Así f puede ser vista como la combinación cónica de funciones convexas, ya que los $w_k \geq 0$. Así mismo f es diferenciable y (continua) respecto a u, ya que tanto el operador S como la norma lo son.

Teorema 2. *Si $\alpha > 0$. Entonces:*

(i) $f : L^2(\Omega) \to \mathbb{R}$ *definida como en* (2.4) *es débilmente semicontinua inferior y radialmente no acotada.*

(ii) *El problema de minimización* (2.4) *tiene solución óptima única.*

Demostración.

(i) Notemos que

$$\begin{aligned} f(u) &= \frac{1}{2}\sum_k\sum_i w_k \left(Su(x_k,t_i) - z_o(x_k,t_i)\right)^2 + \frac{\alpha}{2} \parallel u \parallel^2_{L^2(\Omega)} \\ &\geq \frac{\alpha}{2} \parallel u \parallel^2_{L^2(\Omega)} . \end{aligned}$$

De donde se sigue inmediatamente que f es radialmente no acotada, pues $f(u) \to +\infty$ cuando $\parallel u \parallel \to +\infty$.

Por otro lado, debido a que f es continua y convexa es también semicontinua inferior (ver [Tröltzsch, 2010], pp.47).

(ii) Tomemos $\{u_n\}_{n\geq 1} \subset L^2(\Omega)$ una sucesión minimizante, es decir

$$\inf_{u\in L^2(\Omega)} f(u) = \lim_{n\to+\infty} f(u_n).$$

Notemos que la sucesión $\{u_n\}$ es acotada. Razonando por contradicción si suponemos que no lo es, se contradice que f sea radialmente no acotada, lo cual fue demostrado en (i). Por otro lado, como $L^2(\Omega)$ es un espacio reflexivo sabemos que existe una subsucesión $\{u_{n_k}\} \subset \{u_n\}$ tal que $u_{n_k} \rightharpoonup \bar{u}$ cuando $k \to +\infty$ (ver [Tröltzsch, 2010], pp.46). Luego, gracias a la semicontinuidad inferior de f se sigue que

$$f(\bar{u}) \leq \lim_{k\to\infty} \inf f(u_{n_k}) = \inf_{u\in L^2(\Omega)} f(u).$$

De donde $\bar{u} \in L^2(\Omega)$ es una solución óptima de (2.4).

La unicidad de esta solución se sigue de la convexidad estricta de la función f, lo cual se obtiene al exigir $\alpha > 0$.

□

Notemos además que al estar minimizando sobre todos los $u \in L^2(\Omega)$, necesitamos que el parámetro de regularización de Tikhonov sea estrictamente positivo para poder garantizar la existencia de la solución del problema de optimización, por lo cual no consideraremos el caso $\alpha = 0$.

Gracias al Teorema 2, se demuestra la existencia y unicidad de $\bar{u} \in L^2(\Omega)$ control óptimo del problema, con $\bar{y} = y(\bar{u}) \in W(0,T)$ estado óptimo asociado.

3. Sistema de optimalidad

Para iniciar esta sección definamos el siguiente espacio

$$L^2\left(0,T;W_0^{1,r}\right)=\left\{g:[0,T]\to W_0^{1,r}(\Omega)\text{ medibles }:\int_0^T\parallel g(t)\parallel^2_{W_0^{1,r}(\Omega)}dt<\infty\right\},$$

con $r\in\left[1,\frac{n}{n-1}\right[$.

De aquí en adelante, notaremos a este espacio como X y a su dual como X^*, es decir

$$\begin{aligned}X=&L^2\left(0,T;W_0^{1,r}\right)\\X^*=&L^2\left(0,T;W_0^{-1,r'}\right),\end{aligned}$$

donde r' representa el conjugado de r, es decir, $\frac{1}{r}+\frac{1}{r'}=1$.

Observación 1. *Si $z\in W(0,T)$, z_t por definición pertenecerá al espacio $L^2\left(0,T;H^{-1}\right)$. Por otro lado si $z\in X$, z_t denotará la derivada en el sentido de las distribuciones, es decir*

$$z_t\varphi=-\int_0^T z\varphi_t dt,\qquad\forall\varphi\in\mathcal{D}(I),$$

con $I=[0,T]$ ([Roubíček, 2013], pp.200).

De la sección anterior sabemos que existe un control óptimo único $\bar{u}\in L^2(\Omega)$ del problema (2.2) con $\bar{y}=y(\bar{u})\in W(0,T)$ su estado óptimo asociado. Notando como es usual a la ecuación de estado como $e(y,u)=0$, con

$$e:W(0,T)\times L^2(\Omega)\to X^*,$$

tal que

$$\langle e(y,u),p\rangle_{X^*,X}=\int_\Omega\left(y(T)p(T)-up(0)\right)dx-\int_\Omega\int_0^T(p_ty+\nabla y\nabla p)dxdt-\int_\Gamma\int_0^T p\partial_\nu y dxdt$$

para cualquier $p\in X$ y donde $\partial_\nu y$ representa la derivada de y en el sentido de la normal. Para obtener el sistema de optimalidad del problema de asimilación de datos se utilizará el enfoque Lagrangiano. Sea p el multiplicador de Lagrange asociado a la

restricción $e(y,u)=0$, el operador Lagrangiano queda expresado por

$$\begin{aligned}\mathcal{L}: W(0,T)\times L^2(\Omega)\times X &\to \mathbb{R}\\ (y,u,p) &\mapsto \mathcal{L}(y,u,p)=J(y,u)+\langle e(y,u),p\rangle_{X*,X}\end{aligned}$$

es decir

$$\begin{aligned}\mathcal{L}(y,u,p) &= \sum_k\sum_i w_k\left[y(x_k,t_i)-z_o(x_k,t_i)\right]^2+\frac{\alpha}{2}\parallel u\parallel^2_{L^2(\Omega)}-\int_\Omega\int_0^T\left(p_t y+\nabla y\nabla p\right)dxdt\\ &+\int_\Omega\left(y(T)p(T)-up(0)\right)dx-\int_\Gamma\int_0^T p\partial_\nu y dxdt\\ &=\sum_k\sum_i w_k\left[y(x_k,t_i)-z_o(x_k,t_i)\right]^2+\frac{\alpha}{2}\parallel u\parallel^2_{L^2(\Omega)}+\int_\Omega\int_0^T\left(-p_t-\Delta p\right)ydxdt\\ &+\int_\Omega\left(y(T)p(T)-up(0)\right)dx-\int_\Gamma\int_0^T p\partial_\nu y dxdt.\end{aligned}$$

El sistema

$$\begin{aligned}e(\bar{y},\bar{u})&=0,\\ \mathcal{L}_y(\bar{y},\bar{u},p)&=0,\\ \mathcal{L}_u(\bar{y},\bar{u},p)&=0,\end{aligned}$$

representa el sistema de optimalidad del problema estudiado ([De los Reyes, 2015], pp.34) y estará correctamente definido si $\mathcal{L}_y(\bar{y},\bar{u},p)=0$, que corresponde a la ecuación adjunta del sistema, tiene solución.

El estado adjunto se obtiene al derivar $\mathcal{L}(y,u,p)$ respecto a y en alguna dirección v e igualar el resultado a cero. Así, notando $\mathcal{L}_y=\mathcal{L}_y(y,u,p)$, tenemos

$$\begin{aligned}\mathcal{L}_y(v) &= \sum_k\sum_i w_k\left(y(x_k,t_i)-z_o(x_k,t_i)\right)v(x_k,t_i)+\int_\Omega\int_0^T\left(p_t-\Delta p\right)vdxdt\\ &-\int_\Gamma\int_0^T p\partial_\nu v dxdt+\int_\Omega v(T)p(T)dx=0.\end{aligned}$$

En la expresión anterior la doble sumatoria puede escribirse formalmente como sigue

$$\int_\Omega \int_0^T \sum_k \sum_i w_k \left(y(x,t) - z_o(x,y)\right) v(x,t)\delta(x-x_k)\delta(t-t_i)dxdt,$$

donde $\delta(x-x_k)$ y $\delta(t-t_i)$ representan la masa de Dirac concentrada en los puntos x_k y t_i respectivamente ([Lions, 1971], pp.79). Estas distribuciones devuelven el valor de un funcional cuando es evaluado en un punto, es decir

$$\int_\Omega g(x)\delta(x-x_k)dx = g(x_k).$$

De donde

$$\begin{aligned}\mathcal{L}_y(v) = &\int_\Omega \int_0^T \left(\sum_k \sum_i w_k \left[y(x,t) - z_o(x,y)\right] \delta(x-x_k)\delta(t-t_i) - p_t - \Delta p \right) v(x,t)dxdt \\ &- \int_\Gamma \int_0^T p\partial_\nu v dxdt + \int_\Omega p(T)v(T)dx = 0.\end{aligned}$$

De aquí,

(i) $\forall v \in X^*$ tal que $v(T) = 0$, $\partial_\nu v = 0$ se sigue que

$$\int_\Omega \int_0^T \left(-p_t - \Delta p\right) v dxdt = \int_\Omega \int_0^T \left(\sum_k \sum_i w_k \left[z_o(x,t) - y(x,y)\right] \delta(x-x_k)\delta(t-t_i) \right) v dxdt$$

(ii) Si $\partial_\nu v = 0$, entonces $p(x,T) = 0$

(iii) Si $v(T) = 0$, entonces $p|_\Sigma = 0$.

Así, la formulación fuerte de la ecuación adjunta está dada por

$$\begin{aligned} -p_t - \Delta p = &\sum_k \sum_i w_k \left[z_o(x,t) - y(x,t)\right] \otimes \delta(x-x_k) \otimes \delta(t-t_i) && \text{en } \Omega\times]0,T[\\ p = &\ 0 && \text{en } \Gamma\times]0,T[\\ p(x,T) = &\ 0 && \text{en } \Omega, \end{aligned} \tag{2.5}$$

donde

$$g(x) \otimes \delta(x-x_k) = \int_\Omega g(x)\delta(x-x_k)dx = g(x_k).$$

De manera similar, la ecuación del gradiente se obtiene al derivar el operador Lagrangiano con respecto a u en alguna dirección h e igualar a cero el resultado, es decir

$$\begin{aligned}\mathcal{L}_u(h) &= \alpha\langle u, h\rangle_{L^2(\Omega)} - \int_\Omega hp(0)dx \\ &= \int_\Omega (\alpha u - p(0))hdx = 0, \quad \forall h \in L^2(\Omega)\end{aligned}$$

de donde la ecuación del gradiente del problema (2.2) queda expresada por

$$\alpha u - p(0) = 0 \quad \text{en } L^2(\Omega). \tag{2.6}$$

Finalmente para que el sistema de optimalidad quede bien definido resta demostrar que (2.5) tiene solución en un espacio funcional apropiado.

3.1. Existencia de solución de la ecuación adjunta

El siguiente teorema establece un resultado de existencia y unicidad de una solución de la ecuación adjunta del problema de asimilación de datos.

TEOREMA 2.1. *Si $n \leq 3$, entonces existe una única solución muy débil $p \in L^2(Q)$ del sistema* (2.5).

Demostración. Si multiplicamos (2.5) por ψ, una función test en un espacio funcional conveniente, e integramos por partes de una manera formal, obtenemos que una solución muy débil de (2.5) es un $p \in L^2(Q)$ tal que verifica la siguiente formulación variacional

$$\int_\Omega \int_0^T (\psi_t - \Delta\psi)\, p dx dt = \sum_k \sum_i w_k \left[z_o(x_k, t_i) - y(x_k, t_i)\right] \psi(x_k, t_i) \tag{2.7}$$

$\forall \psi \in H^{2,1}(Q), \psi|_\Sigma = 0, \psi(x, 0) = 0.$
Con

$$H^{2,1}(Q) = \left\{ g, \frac{\partial g}{\partial x_i}, \frac{\partial^2 g}{\partial x_i x_j}, \frac{\partial g}{\partial t} \in L^2(Q) \right\},$$

equipado con la norma

$$\| g \|^2_{H^{1,2}(Q)} = \int_\Omega \int_0^T \left(|g|^2 + \sum_i \left(\frac{\partial g}{\partial x_i}\right)^2 + \sum_{i,j} \left(\frac{\partial^2 g}{\partial x_i \partial x_j}\right)^2 + \left(\frac{\partial g}{\partial t}\right)^2 \right) dx dt.$$

Para probar que en efecto $p \in L^2(Q)$, consideremos el siguiente problema auxiliar:

$$\begin{aligned} \varphi_t - \Delta\varphi = &\ \phi \quad \text{en } Q \\ \varphi = &\ 0 \quad \text{en } \Sigma \\ \varphi(x,0) = &\ 0 \quad \text{en } \Omega. \end{aligned} \tag{2.8}$$

para cada $\phi \in L^2(Q)$. El sistema anterior tiene una solución única $\varphi \in H^{2,1}(Q)$ ([Lions, 1971], pp.182). Por tanto la aplicación que asigna a cada $\phi \in L^2(Q)$ un $\varphi \in H^{2,1}(Q)$ solución de (2.8)

$$\begin{aligned} \mathcal{G} : L^2(Q) &\to H^{2,1}(Q) \\ \phi &\mapsto \mathcal{G}(\phi) = \varphi \end{aligned}$$

es un isomorfismo. Notemos además que G es biyectiva, de donde su inversa también lo será, es decir

$$\begin{aligned} \mathcal{G}^{-1} : H^{2,1}(Q) &\to L^2(Q) \\ \varphi &\mapsto \mathcal{G}^{-1}(\varphi) = \varphi_t - \Delta\varphi = \phi, \end{aligned}$$

constituye un isomorfismo definido de $H^{2,1}(Q)$ en $L^2(Q)$. Por transposición del isomorfismo biyectivo G^{-1} tenemos

$$\left\langle \mathcal{G}^{-1}(\varphi), v \right\rangle_{L^2(Q),L^2(Q)} = \langle \varphi, \left(\mathcal{G}^{-1}\right)^* v \rangle_{H^{2,1},(H^{2,1})^*}, \qquad \forall \varphi \in H^{2,1}(Q), \forall v \in L^2(Q), \tag{2.9}$$

donde

$$\begin{aligned} \left(\mathcal{G}^{-1}\right)^* : L^2(Q) &\to (H^{2,1}(Q))^* \\ v &\mapsto \left(\mathcal{G}^{-1}\right)^* v = -v_t - \Delta v. \end{aligned}$$

En particular tomando $p \in L^2(\Omega)$ y notando que la formulación variacional (2.7) se verifica para $\varphi \in H^{2,1}(Q)$ solución de (2.8), podemos expresar (2.9) como sigue:

$$\langle \phi, p \rangle_{L^2(Q),L^2(Q)} = \sum_k \sum_i w_k \left[z_o(x_k, t_i) - y(x_k, t_i) \right] \varphi(x_k, t_i). \tag{2.10}$$

Definamos ahora el siguiente operador

$$\begin{aligned} \sigma : H^{2,1}(Q) &\to \mathbb{R} \\ \varphi &\mapsto \sigma(\varphi) = \sum_{i,k} w_k \left[z_o(x_k, t_i) - y(x_k, t_i) \right] \varphi(x_k, t_i), \end{aligned}$$

como $H^{2,1}(Q)$ es un espacio de Hilbert ([Lions, 1971], pp 119), si $\sigma \in \left(H^{2,1}\right)^*$ entonces existe un único $p \in L^2(\Omega)$ solución de (2.10). Este resultado se verifica mediante la utilización del Teorema de representación de Riez. Notemos además que al hallar $p \in L^2(\Omega)$ solución de (2.10), hemos hallado $p \in L^2(\Omega)$ solución muy débil de (2.5). Para concluir la demostración resta probar que en efecto σ así definido es un elemento de $\left(H^{2,1}\right)^*$, es decir, que el operador σ sea lineal y continuo. Consideremos φ_1, $\varphi_2 \in H^{2,1}(Q)$ y $\lambda \in \mathbb{R}$, desarrollando $\sigma(\varphi_1 + \lambda\varphi_2)$ se sigue que

$$\begin{aligned}
\sigma(\varphi_1 + \lambda\varphi_2) &= \sum_k \sum_i w_k \left[z_o(x_k, t_i) - y(x_k, t_i)\right] (\varphi_1 + \lambda\varphi_2)(x_k, t_i) \\
&= \sum_k \sum_i w_k \left[z_o(x_k, t_i) - y(x_k, t_i)\right] (\varphi_1(x_k, t_i) + \lambda\varphi_2(x_k, t_i)) \\
&= \sum_k \sum_i w_k \left[z_o(x_k, t_i) - y(x_k, t_i)\right] \varphi_1(x_k, t_i) \\
&+ \lambda \sum_k \sum_i w_k \left[z_o(x_k, t_i) - y(x_k, t_i)\right] \varphi_2(x_k, t_i) \\
&= \sigma(\varphi_1) + \lambda\sigma(\varphi_2).
\end{aligned}$$

Lo que prueba la linealidad del operador. Gracias a la linealidad para probar la continuidad, basta demostrar que σ es acotado, es decir que existe una constante $C > 0$ tal que

$$|\sigma(\varphi)| \leq C \parallel \varphi \parallel_{H^{2,1}(Q)} .$$

En efecto,

$$\begin{aligned}
|\sigma(\varphi)| &= |\sum_k \sum_i w_k \left[z_o(x_k, t_i) - y(x_k, t_i)\right] \varphi(x_k, t_i)| \\
&= |\sum_k \sum_i \int_0^T w_k \left[z_o(x_k, t) - y(x_k, t)\right] \varphi(x_k, t)\delta(t - t_i)dt| \\
&\leq |\sum_k \sum_i \int_0^T w_k \left[z_o(x_k, t) - y(x_k, t)\right] \varphi(x_k, t)dt| \\
&\leq \sum_i \left(\sum_k \int_0^T |w_k \left[z_o(x_k, t) - y(x_k, t)\right] \varphi(x_k, t)dt| \right) \\
&\leq \sum_i \left(\sum_k \parallel w_k \left[z_o(x_k) - y(x_k)\right] \varphi(x_k) \parallel_{L^1(0,T)} \right)
\end{aligned}$$

con

$$\varphi(x_k) : [0,T] \to \mathbb{R}$$
$$t \mapsto \varphi(x_k)t = \varphi(x_k,t) \in L^2(0,T).$$

De aquí y aplicando la desigualdad de Hölder se sigue que

$$\begin{aligned}
|\sigma(\varphi)| &\leq \sum_i \left(\sum_k \| \, w_k \left[z_o(x_k) - y(x_k) \right] \|_{L^2(0,T)} \| \, \varphi(x_k) \, \|_{L^2(0,T)} \right) \\
&\leq \sum_i \left(\sum_k \| \, w_k \left[z_o(x_k) - y(x_k) \right] \|_{L^2(0,T)} \max_k \| \, \varphi(x_k) \, \|_{L^2(0,T)} \right) \\
&= \underbrace{\max_k \| \, \varphi(x_k) \, \|_{L^2(0,T)}}_{\|\varphi\|_{L^2(0,T;C(\bar{\Omega}))}} \sum_i \underbrace{\left(\sum_k \| \, w_k \left[z_o(x_k) - y(x_k) \right] \|_{L^2(0,T)} \right)}_{c_1 > 0} \\
&\leq \mu_T c_1 \| \, \varphi \, \|_{L^2(0,T;C(\bar{\Omega}))},
\end{aligned}$$

donde $\mu_T > 0$ representa el número de observaciones en $[0,T]$ de las que se dispone. Así,

$$|\sigma(\varphi)| \leq c_1 \mu_T \| \, \varphi \, \|_{L^2(0,T;C(\bar{\Omega}))} . \tag{2.11}$$

De las hipótesis del teorema tenemos que $n \leq 3$, por lo tanto se verifica la inmersión continua $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$ ([Quarteroni, 2010], pp 25), de donde existe $c_2 > 0$ tal que:

$$\| \, \varphi \, \|_{L^2(0,T;C(\bar{\Omega}))} \leq c_2 \| \, \varphi \, \|_{L^2(0,T;H^2(\Omega))}, \tag{2.12}$$

donde

$$H^2(\Omega) = \left\{ g, \frac{\partial g}{\partial x_i}, \frac{\partial^2 g}{\partial x_i x_j} \in L^2(Q) \right\},$$

con norma asociada

$$\| \, g \, \|^2_{H^2(\Omega)} = \int_\Omega \left(|g|^2 + \sum_i \left(\frac{\partial g}{\partial x_i} \right)^2 + \sum_{i,j} \left(\frac{\partial^2 g}{\partial x_i \partial x_j} \right)^2 \right),$$

de donde se sigue inmediatamente que

$$\| \, \varphi \, \|_{L^2(0,T;H^2(\Omega))} \leq \| \, \varphi \, \|_{H^{2,1}(Q)} . \tag{2.13}$$

En efecto,

$$\begin{aligned}
\| \varphi \|^2_{L^2(0,T;H^2(\Omega))} &= \int_0^T \| \varphi(t) \|^2_{H^2(\Omega)} \, dt \\
&= \int_0^T \left(\int_\Omega |\varphi|^2 + \sum_i \left(\frac{\partial \varphi}{\partial x_i} \right)^2 + \sum_{i,j} \left(\frac{\partial^2 \varphi}{\partial x_i \partial x_j} \right)^2 dx \right) dt \\
&\leq \int_0^T \int_\Omega \left(|\varphi|^2 + \sum_i \left(\frac{\partial \varphi}{\partial x_i} \right)^2 + \sum_{i,j} \left(\frac{\partial^2 \varphi}{\partial x_i \partial x_j} \right)^2 + \left(\frac{\partial \varphi}{\partial t} \right)^2 \right) dx dt \\
&= \| \varphi \|^2_{H^{2,1}(Q)} \, .
\end{aligned}$$

Finalmente de (2.11) - (2.13) se sigue que

$$|\sigma(\varphi)| \leq c_1 c_2 \mu_T \| \varphi \|_{H^{2,1}(Q)}, \tag{2.14}$$

tomando $C = c_1.c_2.\mu_T > 0$ se verifica el resultado. □

Como se había establecido, en el lado derecho de la ecuación adjunta los términos δ_{x_k} y δ_{t_i} representan para cada k e i la masa de Dirac concentrada en los puntos $x_k \in \Omega$ y $t_i \in]0, T[$ respectivamente. Al tratarse de medidas de probabilidad podemos definir una nueva medida en el espacio producto $Q = \Omega \times]0, T[$ como el producto de las medidas en cada espacio, es decir:

$$\begin{aligned}
\delta((x,t) - (x_k, t_i)) &= \delta(x - x_k)\delta(t - t_i) \\
&= \begin{cases} +\infty & \text{, si } (x_k, t_i) = (x, t) \\ 0 & \text{, caso contrario.} \end{cases}
\end{aligned}$$

Nótese que $\delta((x,t) - (x_k, t_i))$ así definida representa la masa de Dirac concentrada en el punto $(x_k, t_i) \in Q$. Una propiedad importante de las medidas de Dirac es que pertenecen al espacio de las medidas regulares de Borel [Bauer, 2001]. Así,

$$\delta((x,t) - (x_k, t_i)) \in \mathcal{M}(Q), \qquad \forall i = \{1, \ldots, \mu_T\} \text{ y } k = \{1, \ldots, \mu_E\},$$

donde $\mathcal{M}(Q)$ representa al conjunto de las medidas regulares de Borel en Q. Definiendo $\delta \in \mathcal{M}(Q)$ como la combinación lineal de medidas de Dirac dada por

$$\delta = \sum_k \sum_i w_k [z_o - y] \otimes \delta((x,t) - (x_k, t_i)), \tag{2.15}$$

se puede reescribir (2.5) como sigue

$$\begin{aligned} -p_t - \Delta p = &\quad \delta \quad \text{en } \Omega\times]0,T[\\ p = &\quad 0 \quad \text{en } \Gamma\times]0,T[\\ p(x,T) = &\quad 0 \quad \text{en } \Omega. \end{aligned}$$

Del Teorema 2.1 sabemos que para el sistema (2.5) existe una única solución muy débil$p \in L^2(Q)$, además como $\delta \in \mathcal{M}(Q)$ se puede concluir que $p \in L^2\left(0,T;W_0^{1,r}\right)$ con $r \in \left[1,\frac{n}{n-1}\right[$, con $r \in \left[1,\frac{n}{n-1}\right[$ ([Casas et al., 2013], pp 31). Es decir , podemos concluir que $p \in X$.

Ya que se ha mostrado la existencia de un único $p \in L^2\left(0,T;W_0^{1,r}\right)$, con $r \in \left[1,\frac{n}{n-1}\right[$, solución muy débil de la ecuación adjunta, el sistema de optimalidad del problema de asimilación de datos queda plenamente determinado y está expresado por:

Ecuación de estado:

$$\begin{aligned} \frac{\partial y}{\partial t} - \Delta y &= 0 \quad \text{en } Q = \Omega\times]0,T[\\ y &= 0 \quad \text{en } \Sigma = \Gamma\times]0,T[\\ y(0) &= u \quad \text{en } \Omega. \end{aligned}$$

Ecuación adjunta:

$$\begin{aligned} -\frac{\partial p}{\partial t} - \Delta p = &\quad \sum_k \sum_i w_k \left[z_o(x,t) - y(x,t)\right] \otimes \delta(x - x_k) \otimes \delta(t - t_i) \quad \text{en } Q = \Omega\times]0,T[\\ p = &\quad 0 \quad \text{en } \Sigma = \Gamma\times]0,T[\\ p(T) = &\quad 0 \quad \text{en } \Omega. \end{aligned}$$

Ecuación del gradiente:

$$\alpha u - p(0) = 0 \quad \text{en } L^2(\Omega).$$

4. Métodos Quasi-Newton

La idea principal de los métodos de descenso es encontrar en una iteración u_k, una dirección de descenso d_k tal que

$$f(u_k + \hat{\alpha}_k d_k) < f(u_k), \quad \hat{\alpha}_k > 0,$$

donde $\hat{\alpha}_k$ es un parámetro de búsqueda lineal que puede ser hallado con estrategias de búsqueda como la de Armijo o Wolfe. La desigualdad (2.16) representa la regla de Armijo, mientras que (2.16) y (2.17) las condiciones de Wolfe.

$$f(u_k + \alpha_k d_k) - f(u_k) \quad \leq \gamma \alpha_k \nabla f(u_k).d_k \tag{2.16}$$

$$\nabla f(u_k + \alpha_k d_k).d_k \quad \geq \beta \nabla f(u_k).d_k \tag{2.17}$$

con $0 < \gamma < \beta < 1$.

En los métodos Quasi-Newton la dirección de descenso se obtiene mediante

$$d_k = -(H_k)^{-1} \nabla f(u_k),$$

donde H_k representa una aproximación de la Hessiana mediante ecuaciones del tipo secante. Nótese que si se toma $H_k = I$, la dirección que se obtiene es la del máximo descenso. Por otro lado si H_k representa a la matriz Hessiana, d_k es la dirección obtenida mediante el método de Newton. De manera general las ecuaciones del tipo secante son del tipo

$$H_{k+1} \underbrace{(u_{k+1} - u_k)}_{s_k} = \underbrace{\nabla f(u_{k+1}) - \nabla f(u_k)}_{z_k},$$

ya que la matriz H_{k+1} no está unívocamente determinada se exigen criterios adicionales para la construcción de las matrices, uno de ellos consiste en actualizar H_{k+1} a partir de modificaciones sencillas de H_k de tal manera que se preserve la simetría de la matriz. Otra alternativa es escoger H_{k+1} como la solución de un problema de optimización, por ejemplo:

$$\begin{cases} \min\limits_{B} \parallel W \quad \left(B^{-1} - B_k^{-1}\right) W \parallel_F \\ \text{sujeto a:} \\ \qquad B \quad = B^T \\ \qquad Bs_k \quad = z_k, \end{cases} \tag{2.18}$$

donde $\parallel . \parallel_F$ representa la norma de Frobenius y W es una matriz definida positiva tal que $W^2 s_k = z_k$.

La solución del sistema (2.18) está dada por

$$B_{k+1} = B_k + \frac{(s_k - B_k z_k) s_k^T + s_k (s_k - B_k z_k)^T}{s_k^T z_k} - \frac{(s_k - B_k z_k)^T z_k}{(s_k^T z_k)^2} s_k s_k^T, \tag{2.19}$$

y corresponde a la actualización del método BFGS (Broyden-Flitcher-Goldfarb-Shanno).

Asumiendo que se verifica la condición de curvatura, es decir, $z_k^T s_k > 0$ y empleando las fórmulas de Sherman-Morrison-Woodbury se obtiene la actualización de H_k, donde $H_k = B_k^{-1}$.

$$H_{k+1} = H_k - \frac{H_k s_k s_k^T H_k}{s_k^T H_k s_k} + \frac{z_k z_k^T}{z_k^T s_k}. \tag{2.20}$$

Bajo ciertas hipótesis se puede demostrar que el algoritmo del BFGS está bien definido y genera una sucesión $\{u_k\}$ que converge hacia $\bar{u}$ superlinealmente. El siguiente teorema formaliza este resultado ([Geiger and Kanzow, 2013], [Kelley, 1999]).

Teorema 3. *Sea f dos veces continuamente diferenciable con Hessiana localmente Lipschitz continua y $\bar{u}$ tal que $\nabla^2 f(\bar{u})$ sea simétrica y definida positiva. Entonces existen $\epsilon > 0$ y $\delta > 0$ tal que si para u_0 se verifica que*

$$\| u_0 - \bar{u} \| < \epsilon \quad y \quad \| H_0 - \nabla^2 f(\bar{u}) \| < \delta,$$

entonces las iteraciones del BFGS están bien definidas y generan una sucesión $\{u_k\}$ que converge superlinealmente hacia $\bar{u}$.

Para poder justificar la utilización del Teorema 3 en el problema de asimilación de datos notemos lo siguiente. El funcional reducido, expresado por

$$f(u) = \frac{1}{2} \sum_k \sum_i w_k \left(Su(x_k, t_i) - z_o(x_k, t_i) \right)^2 + \frac{\alpha}{2} \| u \|^2_{L^2(\Omega)}$$

es dos veces continuamente diferenciable. En efecto, de lo tratado anteriormente, sabemos que el operador control-estado S es lineal y continuo respecto a $u \in L^2(\Omega)$, y es por tanto dos veces continuamente diferenciable. Como la norma al cuadrado también verifica esta propiedad se concluye el resultado.

Es importante notar que las hipótesis del teorema se aplicarán al funcional reducido discretizado y que mediante el proceso de discretización la regularidad del funcional no se ve afectada, con lo cual quedaría verificada la primera hipótesis del teorema. Por otro lado, tanto el funcional reducido como su versión discretizada dan lugar a una forma cuadrática en $L^2(\Omega)$ y $\mathbb{R}^{\mu_E}$ respectivamente. De aquí, se concluye inmediatamente la Lipschitz continuidad de la Hessiana. Finalmente, la matriz asociada a la forma cuadrática del funcional reducido discretizado es simétrica y definida positiva con lo cual se verifican todas las hipótesis para la utilización del teorema.

Es importante recordar que todo el sistema de optimalidad del problema de asimilación de datos será la restricción a la que esté sujeta el problema de localización óptima, en este sentido, para la resolución numérica del problema de asimilación de datos se utilizará el método BFGS debido a su velocidad de convergencia superlineal.

5. Implementación numérica

En está sección se mostrarán los resultados obtenidos con la implementación del BFGS para la resolución del problema de asimilación de datos cuando se trabaja con observaciones puntuales tanto en el tiempo como en el espacio. Lo que se busca es obtener un control y estado óptimos que aproximen al control deseado y al estado observado, proceso que recibirá el nombre de reconstrucción.

El dominio espacial en el que se va a trabajar es $\Omega = (0,1) \times (0,1)$, con un mallado de $m \times n$, $h_e = 1/(m-1)$ y $h_t = 1/(n+1)$ los pasos de discretización espacial y temporal respectivamente. Aquí, $\mu_E = m^2$ representa el total de puntos en el espacio y $\mu_T = n + 2$, los instantes en el tiempo. Para la resolución numérica del problema de asimilación de datos mediante el método del BFGS se utiliza la matriz de inicialización $B_0 = \frac{1}{2}I$, donde I representa la matriz identidad.

Primer experimento

El objetivo de este experimento es el de observar la reconstrucción de la condición inicial deseada cuando se trabaja con observaciones puntuales y se varía el número de las mismas. Intuitivamente mientras menos observaciones se tomen, la reconstrucción tanto del estado observado como del control deseados serán menos precisas. En este primer experimento se fijarán $m \times n = 10 \times 12$, el parámetro de regularización de Tikhonov en $\alpha = 10^{-4}$ y se trabajará con una tolerancia de 10^{-4}. La condición inicial a reconstruir estará dada por

$$\begin{aligned} u^{Train} : \Omega \subset \mathbb{R}^2 &\to \mathbb{R} \\ (x,y) &\mapsto u^{Train}(x,y) = \sin(x) + y. \end{aligned}$$

En la figura (1) se observa gráficamente el control que se desea reconstruir, mientras que las figuras (2) y (3) muestra la reconstrucción de u^{Train} cuando se considera una y dos observaciones, y cinco y treinta observaciones respectivamente.

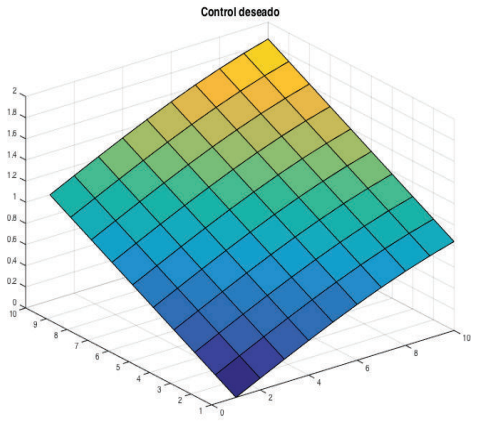

Figura 1: Control deseado

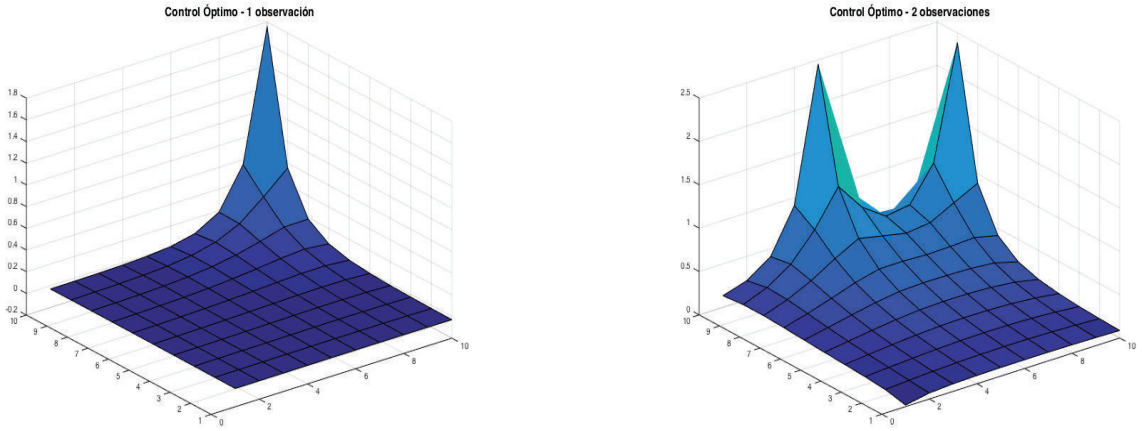

Figura 2: Reconstrucción u^{Train}: 1 y 2 observaciones

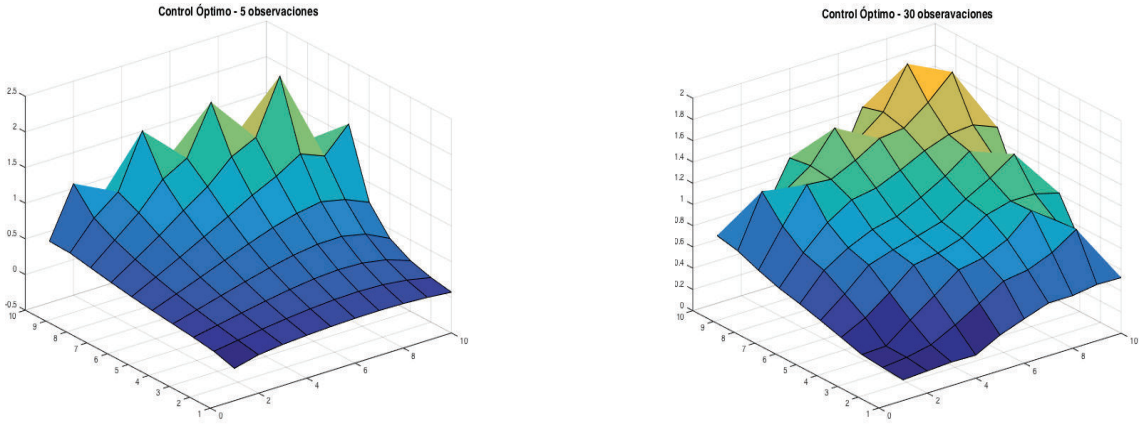

Figura 3: Reconstrucción u^{Train}: 5 y 30 observaciones

Gráficamente se observa que mientras menos datos son tomados en cuenta, la reconstrucción del control deseado es menos exacta. El Cuadro 1 muestra los resultados

obtenidos al comparar los errores absolutos y relativos entre la condición inicial deseada u^{Train} y la condición inicial óptima obtenida u, además del número de iteraciones en cada caso. Aquí,

$$\text{error abs.} = \| u^{Train} - u \|_2$$

y

$$\text{error rel.} = \frac{\| u^{Train} - u \|_2}{\| u^{Train} \|_2}.$$

Nº Obs	iteraciones (s)	error abs.	error rel.
1	21	9,95	0,95
2	38	6,94	0,67
5	56	5,34	0,51
30	71	2,34	0,22

Cuadro 1: Iteraciones y error reconstrucción - Observaciones puntuales

Segundo experimento

Es importante recordar que todo el sistema de optimalidad del problema de asimilación de datos será la restricción a la que esté sujeta el problema de localización óptima. En este sentido, el tiempo de ejecución del algoritmo debe ser relativamente corto. Sin embargo, dependiendo del paso en la discretización que se tome, el tiempo de ejecución puede variar.

El objetivo de este experimento es justamente el de comparar los tiempos de ejecución y número de iteraciones con distintos tamaños de mallas. Al igual que en el primer experimento se fijará $\alpha = 10^{-4}$ y se trabajará con una tolerancia de 10^{-4}. El Cuadro 2 muestra los resultados obtenidos al trabajar con los parámetros señalados y todos los puntos en las discretizaciones espacial y temporal, como se mencionó μ_E indica el número total de puntos en la discretización espacial con los cuales estamos trabajando en cada caso.

Malla	iteraciones (s)	tiempo (s)	h_e
5×6	40	0,23	0,25
10×12	66	1,87	0,11
20×24	97	9,66	0,05
50×50	143	764,39	0,02

Cuadro 2: Tiempo de ejecución e iteraciones

La figura (4) muestra gráficamente el control óptimo obtenido con diferentes pasos en la discretización.

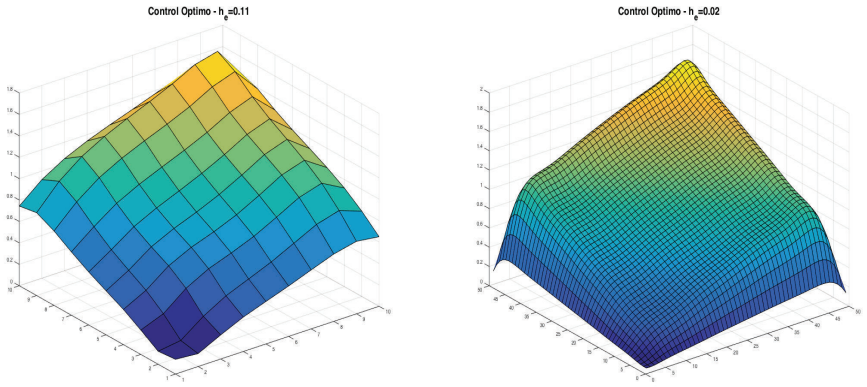

Figura 4: Reconstrucción u^{Train} diferentes pasos de discretización.

Como se puede observar en la figura 4, la reconstrucción de la condición inicial deseada es más precisa cuando el paso de discretización es más pequeño. Sin embargo, debido a que el tiempo de ejecución es significativamente mayor y a que en la resolución numérica del problema a dos niveles se deberá resolver una gran cantidad de problemas de reconstrucción como los tratados, se trabajará en la parte experimental del problema de asimilación con un mallado no tan fino de $m \times n = 10 \times 12$.

Capítulo 3

Problema de Localización Óptima de Observaciones

1. Planteamiento del problema

Dependiendo de la calidad de la información, así como la cantidad de observaciones de las que se disponga, el proceso de asimilación de datos será más o menos útil para la reconstrucción de una condición inicial deseada. En el contexto meteorológico, los datos necesarios para la asimilación pueden ser recopilados de estaciones de medición, radio sondeos, entre otros. Sin embargo, debido al costo o a la dificultad en la instalación de los equipos que permiten obtener estos datos se busca poder localizarlos de manera óptima, de tal forma que la información obtenida sea útil para el proceso de asimilación.

Motivándonos en la resolución de este problema común en la predicción numérica del tiempo y del clima, lo que se busca es resolver un problema de localización óptima de observaciones de tal manera que se logre la reconstrucción de la condición inicial cuando se trabaja con un problema que evoluciona en el tiempo, teniendo en cuenta además que el número de observaciones que pueden ser localizadas es limitado. Para ello, se considerará un modelo de optimización a dos niveles donde el nivel inferior se encargará de la reconstrucción de la condición inicial del estado del sistema, es decir, de encontrar la solución a un problema de asimilación de datos, mientras que el nivel superior resolverá el problema de localización óptima de observaciones. Usando la notación

$$\mathbf{y} = (y(w), p(w), u(w)),$$

la formulación matemática del problema es la siguiente:

$$\min_{w\in\{0,1\}} J(\mathbf{y}, w) = \sum_{j=1}^{N} \parallel y_j^{Train} - y_j \parallel^2 + \beta \sum_{j=1}^{N} \parallel u_j^{Train} - u_j \parallel^2 + \gamma \parallel w \parallel_{l_0}$$

$$\text{s.a} \begin{cases} \min\limits_{u} J(y,u) & = \dfrac{1}{2}\sum\limits_{k}\sum\limits_{i} w_k \left(y_j(x_k,t_i) - y_j^{Train}(x_k,t_i)\right)^2 + \dfrac{\alpha}{2}\int_{\Omega} |u_j|^2 \\ \text{sujeto a:} & \\ \dfrac{\partial y_j}{\partial t} - \Delta y_j & = 0 \text{ en } \Omega\times]0,T[\\ y_j & = 0 \text{ en } \Gamma\times]0,T[\\ y_j(0) & = u_j \text{ en } \Omega. \end{cases}$$

$\forall j = 1, \ldots, N$.

Debido a la dificultad que implica trabajar con la "norma" l_0 o de conteo, que devuelve el número total de elementos diferentes de cero de un vector, se realiza una primera aproximación del problema utilizando $\| \ . \ \|_{l_1}$ en lugar de $\| \ . \ \|_{l_0}$, de donde el nuevo funcional objetivo a resolver sería

$$\min_{w} J(\mathbf{y}, w) = \sum_{j=1}^{N} \parallel y_j^{Train} - y_j \parallel^2 + \beta \sum_{j=1}^{N} \parallel u_j^{Train} - u_j \parallel^2 + \gamma \underbrace{\parallel w \parallel_{l_1}}_{\sum_k |w_k|}$$

Aquí y^{Train} y u^{Train} toman el lugar del estado observado y control deseado respectivamente. Se puede observar que la restricción del problema a dos niveles corresponde a N problemas de asimilación de datos de los estudiados en el capítulo anterior. $\gamma > 0$ es tal que a medida que incremente su valor, el último termino del funcional objetivo también lo hará. Al tratarse de un problema de minimización, mientras más grande sea el valor de γ, w tomará valores cada vez más pequeños. Con la utilización del parámetro γ conjuntamente con la norma l_1 se busca obtener un vector solución w con más entradas nulas. De aquí en adelante γ será referido como el parámetro de penalización de w. Aquí, el término $\beta > 0$ representa el parámetro de regularización de Tikhonov del problema de localización óptima.

Notemos que debido a la presencia de la norma l_1, el funcional objetivo es no diferenciable por lo tanto no se pueden usar las técnicas usuales de optimización que asumen diferenciabilidad de la función objetivo. Se considera entonces una relajación del problema, donde las entradas de w toman valores entre 0 y 1. Con este cambio se consigue la diferenciabilidad del funcional objetivo, pero el problema se convierte en

uno con restricciones tipo caja.

$$\min_{0\leq w\leq 1} J(\mathbf{y}, w) = \sum_{j=1}^{N} \parallel y_j^{Train} - y_j \parallel^2 + \beta \sum_{j=1}^{N} \parallel u_j^{Train} - u_j \parallel^2 + \gamma \sum_k w_k \tag{3.1}$$

$$\text{s.a} \begin{cases} \min\limits_u J(y,u) & = \dfrac{1}{2}\sum\limits_k \sum\limits_i w_k \left(y_j(x_k,t_i) - y_j^{Train}(x_k,t_i) \right)^2 + \dfrac{\alpha}{2}\displaystyle\int_\Omega |u_j|^2 \\ \text{sujeto a:} & \\ \dfrac{\partial y_j}{\partial t} - \Delta y_j & = 0, \text{ en } \Omega \times]0,T[\\ y_j & = 0, \text{ en } \Gamma \times]0,T[\\ y_j(0) & = u_j, \text{ en } \Omega. \end{cases}$$

$\forall j = 1, \ldots, N$.

El índice j recorre el conjunto de entrenamiento o training set y está dado por las duplas:

$$\begin{gathered} \left(u_1^{Train}, y_1^{Train}\right) \\ \vdots \\ \left(u_j^{Train}, y_j^{Train}\right) \\ \vdots \\ \left(u_N^{Train}, y_N^{Train}\right), \end{gathered}$$

que representan simulaciones de la condición inicial y el estado observado. La idea de trabajar con conjuntos de entrenamiento proviene del Machine Learning y consiste en aprender del conjunto dado y aplicar la información aprendida para predecir alguna característica en un dato desconocido. En el contexto en el que estamos trabajando lo que nos interesa aprender del conjunto de entrenamiento es precisamente el vector de localizaciones w, de tal manera que las ubicaciones dadas por él sean en promedio óptimas para todos los pares considerados.

Como una primera aproximación al problema de localización óptima se considerará un caso particular del mismo tomando $j = 1$, es decir, se trabajará solamente con un elemento del conjunto de entrenamiento, ya que los resultados obtenidos pueden extenderse directamente al caso general, es decir, cuando $j > 1$. Con esta consideración, (3.1) puede ser reescrito como

$$\min_{0\leq w\leq 1} J(\mathbf{y}, w) = \parallel y^{Train} - y \parallel^2 + \beta \parallel u^{Train} - u \parallel^2 + \gamma \sum_k w_k, \tag{3.2}$$

donde $(\mathbf{y}, w)$ es la solución de:

$$\begin{aligned}
\frac{\partial y}{\partial t} - \Delta y &= 0, && \text{en } \Omega\times]0,T[\\
y &= 0, && \text{en } \Gamma\times]0,T[\\
y(0) &= u, && \text{en } \Omega \\
-\frac{\partial p}{\partial t} - \Delta p &= \sum_k \sum_i w_k \left[y^{Train}(x,t) - y(x,t) \right] \otimes \delta(x - x_k) \otimes \delta(t - t_i), && \text{en } \Omega\times]0,T[\\
p &= 0, && \text{en } \Gamma\times]0,T[\\
p(T) &= 0, && \text{en } \Omega \\
\alpha u - p(0) &= 0, && \text{en } \Omega.
\end{aligned} \tag{3.3}$$

Sea μ_E el número total de observaciones en el espacio, notemos que para cada $w \in \mathbb{R}^{\mu_E}$, la ecuación de estado (3.3) tiene solución única. Este resultado se formaliza en el siguiente teorema.

Teorema 4. *Si $n \leq 3$ y $\alpha > 0$ con α el parámetro de regularización de Tikhonov del problema de asimilación de datos, entonces para cada $w \in \mathbb{R}^{\mu_E}$ la ecuación de estado* (3.3) *tiene solución única determinada por w,*

$$(y, p, u) \in W(0,T) \times L^2\left(0, T; W_0^{1,r}\right) \times L^2(\Omega)$$

con $r \in \left[1, \frac{n}{n-1}\right[$.

Demostración. En efecto, el sistema (3.3) corresponde también al sistema de optimalidad del problema de asimilación de datos dados por los sistemas (2.3), (2.5) y (2.6) tratados en el capítulo anterior, donde se estableció para cada $u \in L^2(\Omega)$ la existencia y unicidad de $y \in W(0,T)$ solución de (2.3) y de $p \in L^2\left(0, T; W_0^{1,r}\right)$ con $r \in \left[1, \frac{n}{n-1}\right[$ solución muy débil de (2.5) tal que verifican la ecuación del gradiente (2.6), con lo cual se concluye el resultado. □

Gracias al Teorema 4 sabemos que para cualquier control $w \in \mathbb{R}^{\mu_E}$ existe una única solución $(y, p, u) \in W(0,T) \times L^2\left(0, T; W_0^{1,r}\right) \times L^2(\Omega)$, con $r \in \left[1, \frac{n}{n-1}\right[$, para la ecuación de estado (3.3). En adelante, asumiremos la existencia de un control óptimo único $\bar{w} \in \mathbb{R}^{\mu_E}$ para el problema de optimización binivel, con $(\bar{y}, \bar{p}, \bar{u}) = (y(\bar{w}), p(\bar{w}), u(\bar{w}))$ su estado óptimo asociado.

2. Sistema de optimalidad

En el capítulo anterior definimos al espacio X como $X = L^2\left(0,T;W_0^{1,r}\right)$ con $r \in \left[1, \frac{n}{n-1}\right[$. Notemos como $\mathbf{Z}$ y $\mathbf{Z}^*$ al siguiente espacio y su dual respectivamente

$$\mathbf{Z} = X \times W(0,T) \times L^2(\Omega)$$
$$\mathbf{Z}^* = X^* \times W(0,T)^* \times L^2(\Omega).$$

Así mismo, bajo el supuesto de que existe un control óptimo único $\bar{w} \in \mathbb{R}^{\mu_E}$ del problema (3.2) con $\bar{\mathbf{y}} := (y(\bar{w}), p(\bar{w}), u(\bar{w})) \in \mathbf{Y} := W(0,T) \times X \times L^2(\Omega)$ su estado óptimo asociado. Notando $\varepsilon(\mathbf{y}, w) = 0$, la restricción del problema de localización, donde

$$\begin{aligned} \varepsilon : \mathbf{Y} \times \mathbb{R}^{\mu_E} &\to \mathbf{Z}^* \\ (\mathbf{y}, w) &\mapsto \varepsilon(\mathbf{y}, w) = (\varepsilon_1(\mathbf{y}, w), \varepsilon_2(\mathbf{y}, w), \varepsilon_3(\mathbf{y}, w)) \end{aligned}$$

y donde $\varepsilon_1, \varepsilon_2, \varepsilon_3$ verifican lo siguiente:

$$\begin{aligned} \varepsilon_1 : \mathbf{Y} \times \mathbb{R}^{\mu_E} &\to X^* \\ (\mathbf{y}, w) &\mapsto \varepsilon_1(\mathbf{y}, w), \end{aligned}$$

es tal que

$$\langle \varepsilon_1(\mathbf{y}, w), \eta \rangle_{X^*,X} = \int_\Omega (y(T)\eta(T) - u\eta(0))\, dx - \int_\Omega \int_0^T (\eta_t y + \nabla\eta\nabla y) dxdt - \int_\Gamma \int_0^T \eta \partial_\nu y dxdt$$

$\forall \eta \in X$.

$$\begin{aligned} \varepsilon_2 : \mathbf{Y} \times \mathbb{R}^{\mu_E} &\to W^* \\ (\mathbf{y}, w) &\mapsto \varepsilon_2(\mathbf{y}, w), \end{aligned}$$

tal que

$$\langle \varepsilon_2(\mathbf{y}, w), \sigma \rangle_{W^*,W} = \int_\Omega (p(0)\sigma(0))\, dx + \int_\Omega \int_0^T (\sigma_t p + \nabla\sigma\nabla p) dxdt - \int_\Gamma \int_0^T \sigma \partial_\nu p dxdt$$
$$- \int_\Omega \int_0^T \sigma \left(\sum_k \sum_i w_k \left[y^{Train}(x,t) - y(x,t) \right] \delta(x - x_k)\delta(t - t_i) \right) dxdt$$

$\forall \sigma \in W := W(0,T)$.

Sea $\delta \in \mathcal{M}(Q)$ definida como en (2.15), con está notación la última integral del producto anterior quedaría expresada como

$$\int_\Omega \int_0^T \sigma \delta dxdt,$$

que está bien definida ya que $\sigma \in W(0,T)$ y este espacio se inyecta continuamente en el espacio de las funciones continuas ([Tröltzsch, 2010], pp.148).

$$\begin{aligned} \varepsilon_3 : \mathbf{Y} \times \mathbb{R}^{\mu_E} &\to L^2(\Omega) \\ (\mathbf{y}, w) &\mapsto \varepsilon_3(\mathbf{y}, w), \end{aligned}$$

tal que

$$\langle \varepsilon_3(\mathbf{y}, w), \theta \rangle_{L^2(\Omega),L^2(\Omega)} = \int_\Omega \theta\, (\alpha u - p(0))\, dx$$
$$\forall \theta \in L^2(\Omega).$$

Para derivar el sistema de optimalidad del problema de localización óptima se utilizará nuevamente el enfoque Lagrangiano. Sea $\mathbf{p}$ el multiplicador de Lagrange, con

$$\mathbf{p} := (\eta, \sigma, \theta),$$

asociado a la restricción $\varepsilon(\mathbf{y}, w) = 0$, el operador Lagrangiano queda expresado por

$$\begin{aligned} \mathcal{L} : \mathbf{Y} \times \mathbb{R}^{\mu_E} \times \mathbf{Z} &\to \mathbb{R} \\ (\mathbf{y}, w, \mathbf{p}) &\mapsto \mathcal{L}(\mathbf{y}, w, \mathbf{p}) = J(\mathbf{y}, w) + \langle \varepsilon(\mathbf{y}, w), \mathbf{p} \rangle_{Z^*,Z} \end{aligned}$$

Es decir,

$$\begin{aligned}\mathcal{L}(\mathbf{y},w,\mathbf{p}) &= J(\mathbf{y},w)+\int_\Omega (y(T)\eta(T)-u\eta(0))\,dx-\int_\Omega\int_0^T(\eta_t y+\nabla\eta\nabla y)dxdt-\int_\Gamma\int_0^T\eta\partial_\nu y dxdt\\ &+\int_\Omega (p(0)\sigma(0))\,dx+\int_\Omega\int_0^T(\sigma_t p+\nabla\sigma\nabla p)dxdt-\int_\Gamma\int_0^T\sigma\partial_\nu p dxdt+\int_\Omega\theta\,(\alpha u-p(0))\,dx\\ &+\int_\Omega\int_0^T\sigma\left(-\sum_k\sum_i w_k\left[y^{Train}(x,t)-y(x,t)\right]\delta(x-x_k)\delta(t-t_i)\right)dxdt\end{aligned}$$

o equivalentemente

$$\begin{aligned}\mathcal{L}(\mathbf{y},w,\mathbf{p}) &= J(\mathbf{y},w)+\int_\Omega (y(T)\eta(T)-u\eta(0))\,dxdt-\int_\Omega\int_0^T(\eta_t+\Delta\eta)\,y dxdt-\int_\Gamma\int_0^T\eta\partial_\nu y dxdt\\ &+\int_\Omega (p(0)\sigma(0))\,dx+\int_\Omega\int_0^T(\sigma_t-\Delta\sigma)p dxdt-\int_\Gamma\int_0^T\sigma\partial_\nu p dxdt+\int_\Omega\theta\,(\alpha u-p(0))\,dx\\ &+\int_\Omega\int_0^T\sigma\left(-\sum_k\sum_i w_k\left[y^{Train}(x,t)-y(x,t)\right]\delta(x-x_k)\delta(t-t_i)\right)dxdt\end{aligned}\tag{3.4}$$

La ecuación adjunta resulta de fijar en cero la derivada de $\mathcal{L}(\mathbf{y},w)$ respecto a $\mathbf{y}=(y,p,u)$ en alguna dirección $v=(v_1,v_2,v_3)$, es decir,

$$\mathcal{L}_{\mathbf{y}}(v)=\mathcal{L}_y(v_1)+\mathcal{L}_p(v_2)+\mathcal{L}_u(v_3)=0.\tag{3.5}$$

Para hacerlo, vamos a considerar los sistemas de manera separada. Así,

$$\begin{aligned}\mathcal{L}_y(v_1) &= \int_\Omega\int_0^T\left(-\frac{\partial\eta}{\partial t}-\Delta\eta-2(y^{Train}-y)+\sum_k\sum_i w_k\sigma_1(x,y)\delta(x-x_k)\delta(t-t_i)\right)v_1 dxdt\\ &-\int_\Gamma\int_0^T\eta\partial_\nu v_1 dxdt+\int_\Omega(\eta(T)v_1(T))\,dx=0.\end{aligned}$$

De aquí,

(i) $\forall v_1\in X^*$ tal que $v_1(T)=0$, $\partial_\nu v_1=0$ se sigue que

$$\int_\Omega\int_0^T\left(-\frac{\partial\eta}{\partial t}-\Delta\eta\right)v_1 dxdt=\int_\Omega\int_0^T\left(2(y^{Train}-y)-\sum_k\sum_i w_k\sigma(x,y)\delta(x-x_k)\delta(t-t_i)\right)v_1 dxdt$$

(ii) Si $\partial_v v_1 = 0$, entonces $\eta(x,T) = 0$

(iii) Si $v_1(T) = 0$, entonces $\eta|_\Sigma = 0$.

Así, la formulación fuerte de este primer sistema está dada por

$$\begin{aligned} -\frac{\partial \eta}{\partial t} - \Delta\eta = &\ 2(y^{Train} - y) - \sum_k \sum_i w_k \sigma(x,t) \otimes \delta(x - x_k) \otimes \delta(t - t_i) && \text{en } \Omega\times]0,T[\\ \eta = &\ 0 && \text{en } \Gamma\times]0,T[\\ \eta(T) = &\ 0 && \text{en } \Omega. \end{aligned} \tag{3.6}$$

Ahora,

$$\mathcal{L}_p(v_2) = \int_\Omega \int_0^T \left(\frac{\partial \sigma}{\partial t} - \Delta\sigma\right) v_2 dx dt + \int_\Gamma \int_0^T \sigma \partial_v v_2 dx dt + \int_\Omega (v_2(0)[\sigma(0) - \theta])\, dx$$

De aquí,

(i) $\forall v_2 \in W^*$ tal que $v_2(0) = 0$, $\partial_v v2 = 0$ se sigue que

$$\int_\Omega \int_0^T \left(\frac{\partial \sigma}{\partial t} - \Delta\sigma\right) v_2 dx dt = 0$$

(ii) Si $\partial_v v_2 = 0$, entonces $\sigma(x,0) = \theta$

(iii) Si $v_2(0) = 0$, entonces $\sigma|_\Sigma = 0$.

Y su formulación fuerte es

$$\begin{aligned} \frac{\partial \sigma}{\partial t} - \Delta\sigma = &\ 0 && \text{en } \Omega\times]0,T[\\ \sigma = &\ 0 && \text{en } \Gamma\times]0,T[\\ \sigma(0) = &\ \theta && \text{en } \Omega. \end{aligned} \tag{3.7}$$

Finalmente,

$$\begin{aligned} \mathcal{L}_u(v_3) &= -2\beta \int_\Omega \left(u^{Train} - u\right) v_3 dx - \int_\Omega \eta(0) v_3 dx + \int_\Omega \theta\alpha v_3 dx \\ &= \int_\Omega \left(-2\beta(u^{Train} - u) - \eta(0) + \theta\alpha\right) v_3 dx \end{aligned}$$

de donde

$$\theta = \frac{1}{\alpha}\left[2\beta(u^{Train} - u) + \eta(0)\right]. \tag{3.8}$$

La solución de (3.5) se da al resolver (3.6), (3.7) y (3.8) simultáneamente, así la ecuación adjunta del problema de localización está dada por

$$\begin{aligned}
-\frac{\partial \eta}{\partial t} - \Delta\eta = \;& 2(y^{Train} - y) - \sum_k \sum_i w_k \sigma(x,t) \otimes \delta(x - x_k) \otimes \delta(t - t_i) && \text{en } \Omega\times]0,T[\\
\eta = \;& 0 && \text{en } \Gamma\times]0,T[\\
\eta(T) = \;& 0 && \text{en } \Omega \\
\frac{\partial \sigma}{\partial t} - \Delta\sigma = \;& 0 && \text{en } \Omega\times]0,T[\\
\sigma = \;& 0 && \text{en } \Gamma\times]0,T[\\
\sigma(0) = \;& \tfrac{1}{\alpha}\left[2\beta(u^{Train} - u) + \eta(0)\right] && \text{en } \Omega.
\end{aligned} \tag{3.9}$$

La ecuación del gradiente del problema de localización es el resultado de fijar en cero la derivada de $\mathcal{L}(\mathbf{y}, w)$ respecto a w en alguna dirección $h = (h_1, \ldots, h_{\mu_E})$, con $\mu_E = m^2$, el número de puntos en el espacio.

Así,

$$\begin{aligned}
\mathcal{L}_{w_k}(h_k) &= \gamma h_k - \sum_i h_k \sigma(x_k, t_i)\left(y^{Train}(x_k, t_i) - y(x_k, t_i)\right) \\
&= \left(\gamma - \sum_i \sigma(x_k, t_i)\left(y^{Train}(x_k, t_i) - y(x_k, t_i)\right)\right).h_k, \quad \forall k = 1, \ldots, \mu_E
\end{aligned}$$

de donde, la $k-$ésima componente de la ecuación del gradiente del problema a dos niveles está dada por:

$$\nabla f(w)_k = \gamma - \sum_i \sigma(x_k, t_i)\left(y^{Train}(x_k, t_i) - y(x_k, t_i)\right), \quad \forall k = 1, \ldots, \mu_E, \tag{3.10}$$

donde (3.10) satisface la desigualdad variacional

$$\nabla f(w)(v - w) \geq 0, \quad \forall v \in U_{ad} \tag{3.11}$$

con

$$U_{ad} = \{w \in \mathbb{R}^{\mu_E} : 0 \leq w \leq 1\}.$$

Si definimos,

$$\begin{aligned}
\lambda_a &= \max\{0, \nabla f(w)\} \\
\lambda_b &= |\min\{0, \nabla f(w)\}|,
\end{aligned}$$

la desigualdad variacional (3.11) es equivalente al siguiente sistema

$$\begin{cases} (\lambda_a)_k \geq 0, (\lambda_b)_k \geq 0 & ,\forall k = 1, \ldots, \mu_E \\ (\lambda_a)_k(0 - w_k) = (\lambda_b)_k(w_k - 1) = 0 & ,\forall k = 1, \ldots, \mu_E \\ 0 \leq w_k \leq 1 & ,\forall k = 1, \ldots, \mu_E \end{cases} \tag{3.12}$$

(3.12) es conocido como sistema de complementariedad.

Aquí, asumiremos la existencia de solución de la ecuación adjunta (3.9), con lo cual el sistema de optimalidad del problema de optimización a dos niveles quedaría correctamente definido.

3. Función de penalización para inducir dispersión

La solución que se espera del problema sin relajar es un vector binario de localizaciones que establezca claramente donde ubicar las observaciones. Debido a que lo que se está resolviendo es una relajación del modelo, la solución que se obtendrá es un vector w cuyas entradas serán valores entre cero y uno. Así, si una entrada del vector toma valores intermedios no queda plenamente determinado si se debe tomar o no dicha ubicación. Para obtener un vector w con su mayoría de entradas binarias, cuando se resuelva el problema relajado, se empleará una secuencia de funciones de penalización que sucesivamente aproximen a la "norma" de conteo. Mediante este procedimiento se espera obtener un vector de localizaciones óptimas con su mayoría de entradas binarias. Para utilizar este tipo de funciones que inducen a la obtención de una solución con más entradas nulas es necesario modificar la función objetivo. Así,

$$\min_{0 \leq w \leq 1} J(\mathbf{y}, w) = \sum_{j=1}^{N} \parallel y_j^{Train} - y_j \parallel^2 + \beta \sum_{j=1}^{N} \parallel u_j^{Train} - u_j \parallel^2 + \gamma \Phi_\epsilon(w),$$

donde $\Phi_\epsilon(.)$ es una función que favorece la dispersión en la solución. Una opción para una familia de funciones de penalización es la dada en ([Alexanderian et al., 2014], pp.A 2135), donde

$$\Phi_\epsilon(w) = \sum_k \varphi_\epsilon(w_k)$$

y

$$\varphi_\epsilon(w_k) = \begin{cases} \frac{w_k}{\epsilon} & ,0 \leq w_k < \frac{1}{2}\epsilon \\ p_\epsilon(w_k) & ,\frac{1}{2}\epsilon < w_k \leq 2\epsilon \\ 1 & ,2\epsilon < w_k \leq 1 \end{cases}$$

con $p_\epsilon(.)$ un polinomio de tercer grado únicamente determinado de tal manera que $\varphi_\epsilon : [0,1] \longrightarrow [0,1]$ sea continuamente diferenciable.

Sea $p_\epsilon(w) = aw^3 + bw^2 + cw + e$, sus coeficientes pueden ser obtenidos para cada valor de $\epsilon > 0$ mediante la resolución del siguiente sistema:

$$\begin{pmatrix} \frac{\epsilon^3}{8} & \frac{\epsilon^2}{4} & \frac{\epsilon}{2} & 1 \\ 8\epsilon^3 & 4\epsilon^2 & 2\epsilon & 1 \\ \frac{3\epsilon^2}{4} & \epsilon & 1 & 0 \\ 12\epsilon^2 & 4\epsilon & 1 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ e \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 1 \\ \frac{1}{\epsilon} \\ 0 \end{pmatrix} \tag{3.13}$$

La Figura 1 muestra gráficamente la función de penalización con diferentes valores de ϵ. Como se observa, cuando $\epsilon \to 0$, $\varphi_\epsilon(.)$ aproxima mejor a la "norma l_0" o de conteo.

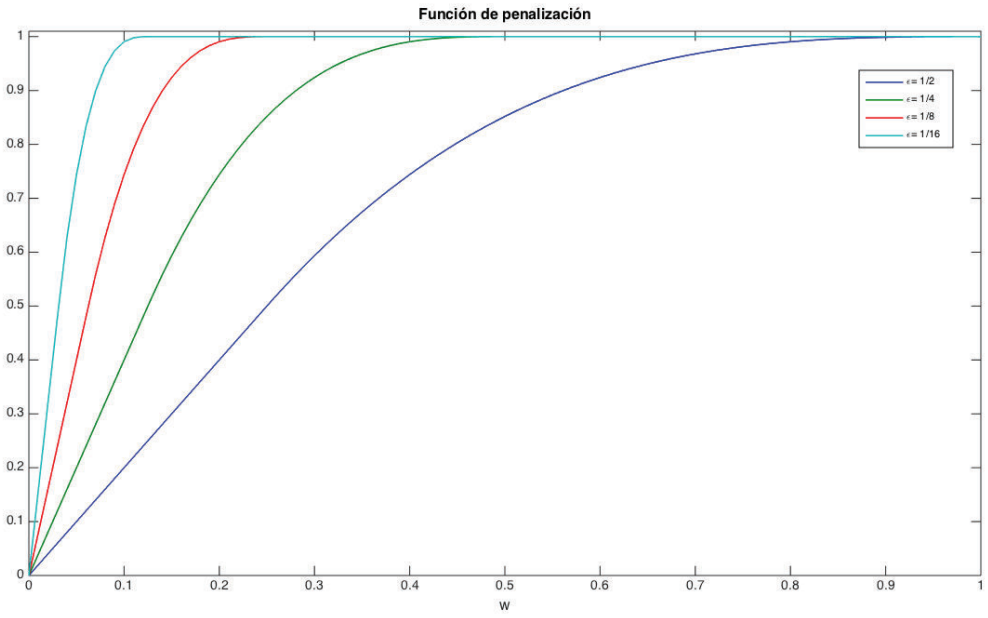

Figura 1: Gráfico función de penalización φ_ϵ

Al obtener el sistema de optimalidad con el nuevo funcional objetivo, el único cambio que se observa es en el cálculo de la ecuación del gradiente. De la misma manera que en el caso anterior, la nueva ecuación del gradiente se obtiene fijando en cero la

derivada del nuevo Lagrangiano respecto a w en alguna dirección h. Así,

$$\begin{aligned}\mathcal{L}_{w_k}(h_k) &= \gamma\varphi'_\epsilon(w_k).h_k - \sum_i h_k\sigma_1(x_k,t_i)\left(y^{Train}(x_k,t_i) - y(x_k,t_i)\right)\\ &= h_k.\left(\gamma\varphi'_\epsilon(w_k) - \sum_i \sigma_1(x_k,t_i)\left(y^{Train}(x_k,t_i) - y(x_k,t_i)\right)\right), \quad \forall k = 1,\ldots,\mu_E\end{aligned}$$

de donde, la $k-$ésima componente de la ecuación del gradiente del problema a dos niveles que considera la función de penalización para inducir dispersión es:

$$\nabla f(w)_k = \gamma\varphi'_\epsilon(w_k) - \sum_i \sigma_1(x_k,t_i)\left(y^{Train}(x_k,t_i) - y(x_k,t_i)\right), \quad \forall k = 1\ldots\mu_E, \quad (3.14)$$

con

$$\varphi'_\epsilon(w_k) = \begin{cases} \frac{1}{\epsilon} & ,0 \le w_k < \frac{1}{2}\epsilon \\ p'_\epsilon(w_k) & ,\frac{1}{2}\epsilon < w_k \le 2\epsilon \\ 0 & ,2\epsilon < w_k \le 1. \end{cases}$$

Similar al caso anterior, (3.14) satisface la desigualdad variacional siguiente

$$\nabla f(w)(v-w) \ge 0, \quad \forall v \in U_{ad}$$

la cual es equivalente a verificar el sistema de complementariedad

$$\begin{cases} (\lambda_a)_k \ge 0, (\lambda_b)_k \ge 0 & ,\forall k = 1,\ldots,\mu_E \\ (\lambda_a)_k(0-w_k) = (\lambda_b)_k(w_k-1) = 0 & ,\forall k = 1,\ldots,\mu_E \\ 0 \le w_k \le 1 & ,\forall k = 1,\ldots,\mu_E \end{cases}$$

donde

$$\begin{aligned}\lambda_a &= \text{máx}\{0, \nabla f(w)\}\\ \lambda_b &= |\,\text{mín}\{0, \nabla f(w)\}|.\end{aligned}$$

Como se mencionó, esta familia de funciones de penalización busca aproximar a la "norma" l_0. Sin embargo, existen otras maneras de conseguir este objetivo, entre ellas se puede mencionar la utilización de las "normas" $\|\,.\,\|_q$ con $q \in (0,1)$. Notemos que cuando se toma $q \in (0,1)$ la función resultante no es una norma, ya que al ser concava, no verifica la desigualdad triangular. En efecto, mientras más cercano a cero se tome q, la "norma" obtenida será mas cercana a la de conteo, lo cual se observa gráficamente en la figura (2). Sin embargo, es importante notar que al trabajar con $\|\,.\,\|_q$ con $q \in (0,1)$ se está realizando una aproximación a través de funciones continuas no convexas.

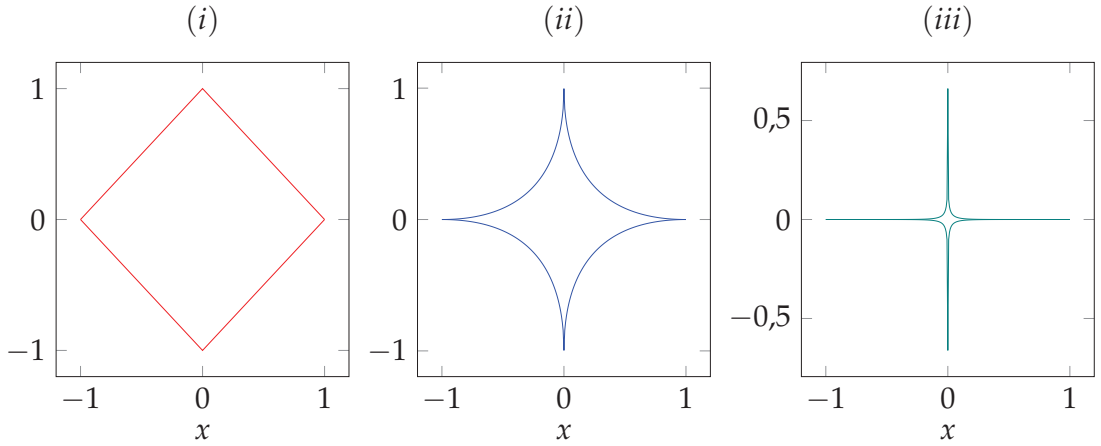

Figura 2: Gráfico $\| x \|_q$ distintos valores de q. $(i)q = 1$, $(ii)q = 0{,}5$, $(iii)q = 0{,}2$.

4. Métodos Quasi-Newton proyectados

4.1. Aspectos generales

De manera general los métodos de proyección usan la dirección de descenso de un problema sin restricciones y proyectan la nueva iteración dentro del conjunto admisible ([De los Reyes, 2015], pp.75). Así, las actualizaciones de w_k estarán dadas por

$$w_{k+1} = P_{U_{ad}}(w_k + \alpha_k d_k),$$

donde U_{ad} es el conjunto de soluciones admisibles definido por a_i y b_i, es decir,

$$U_{ad} = \{w : a_i \leq w_i \leq b_i\},$$

y donde P representa el operador proyección, definido de la siguiente manera:

$$(P_{U_{ad}}(w))_i = \begin{cases} a_i, w_i < a_i \\ w_i, a_i \leq w_i \leq b_i \\ b_i, w_i > b_i. \end{cases}$$

d_k representa alguna dirección de descenso y $\alpha_k \in (0, 1)$ un parámetro de búsqueda lineal. Es importante mencionar que α_k no es calculado de la misma manera que en los métodos no proyectados. Así por ejemplo, una regla de Armijo modificada puede ser usada para hallar el valor más grande de α tal que

$$f\left(P_{U_{ad}}(w_k + \alpha_k d_k)\right) - f(w_k) \leq -\frac{\hat{\gamma}}{\alpha_k} \| P_{U_{ad}}(w_k + \alpha_k d_k) - w_k \|^2 \qquad (3.15)$$

donde $\hat{\gamma} \in (0,1)$ es una constante típicamente fijada en 10^{-4}. Además del criterio de búsqueda para el parámetro α_k, el criterio de parada para los métodos de proyección también difiere en relación con los métodos no proyectados, podemos escoger por ejemplo el siguiente criterio de parada que considera pasos completos:

$$\| w_k - P_{U_{ad}}(w_k - \nabla f(w_k)) \| < \hat{\epsilon}$$

para algún $0 < \hat{\epsilon} \ll 1$ ([De los Reyes, 2015], pp.77).

Usando la dirección de descenso dada por los métodos Quasi-Newton, d_k quedaría expresada por

$$d_k = -(H_k)^{-1}\nabla f(w_k),$$

donde H_k representa a la matriz Hessiana o una aproximación de ella. Siguiendo la lógica anterior se podría pensar que para obtener una dirección de descenso mediante algún método Quasi-Newton proyectado solo bastaría reemplazar d_k por las direcciones dadas por los métodos de segundo orden, es decir

$$w_{k+1} = P_{U_{ad}}(w_k - \alpha_k(H_k)^{-1}\nabla f(w_k)). \tag{3.16}$$

Sin embargo, se pueden construir contraejemplos donde se muestre que esta afirmación es falsa y que usar (3.16) puede llevar a soluciones incorrectas ([Kelley, 1999], pp.98). Esto se debe a que la matriz Hessiana o sus aproximaciones no generan necesariamente direcciones de descenso para el problema con restricciones. En este sentido, en lugar de utilizar la información de la Hessiana se utiliza la información proporcionada por la Hessiana reducida, basada en la estimación de conjuntos $\epsilon-$activos, los cuales están dados por

$$A^{\epsilon}(w) = \{i : a_i \leq w_i \leq a_i + \epsilon \text{ o } b_i \geq w_i \geq b_i - \epsilon\}.$$

Si se considera el caso cuando $\epsilon = 0$, el conjunto recibe el nombre de conjunto activo, y se define como sigue

$$A(w) = \{i : a_i = w_i \text{ o } b_i = w_i\}.$$

Por otro lado, $I^{\epsilon}(w)$ e $I(w)$ representan el conjunto $\epsilon-$inactivo e inactivo, y son los complementos del primer conjunto y del segundo conjunto, respectivamente.

De manera general, si S representa un conjunto de índices cualquiera, entonces R_S denotará la matriz

$$R_S = (\delta_{ij}), \text{ si } i \in S \text{ o } j \in S$$

con

$$\delta_{ij} = \begin{cases} 1, \text{si } i = j, \\ 0, \text{si } i \neq j. \end{cases}$$

Así, la matriz Hessiana reducida para la iteración w_k se define como

$$\begin{aligned} \tilde{R}(w_k, \epsilon_k, H_k) &= R_{A^{\epsilon_k}(w_k)} + R_{I^{\epsilon_k}(w_k)} H_k R_{I^{\epsilon_k}(w_k)} \\ &= \begin{cases} \delta_{ij}, \text{si } i \in A^{\epsilon_k}(w_k) \text{ o } j \in A^{\epsilon_k}(w_k); \\ (H_k)_{ij}, \text{ caso contrario.} \end{cases} \end{aligned}$$

y las iteraciones del método estarán dadas por

$$w_{k+1} = P_{U_{ad}}(w_k - \alpha_k \tilde{R}(w_k, \epsilon_k, H_k)^{-1} \nabla f(w_k)). \tag{3.17}$$

Si la matriz H_k simboliza a la matriz Hessiana, (3.17) representa las iteraciones del método de Newton proyectado. Si en cambio H_k representa la aproximación de la matriz Hessiana dada por las actualizaciones del método BFGS, el método implementado será el BFGS proyectado.

Al igual que su versión no proyectada, el método de Newton proyectado tiene una velocidad de convergencia cuadrática. El Teorema 5 formaliza este resultado (ver [Kelley, 1999], pp.101).

Teorema 5. *Sea f dos veces continuamente diferenciable, $\nabla^2 f$ localmente Lipschitz continua y sea $\bar{w}$ un mínimo local no degenerado. Si w_0 está suficientemente cerca a $\bar{w}$ y $A(w_0) = A(\bar{w})$, entonces las iteraciones del método de Newton proyectado con $\epsilon_k = \| w_k - w_k(1) \|$ convergen q-cuadráticamente a $\bar{w}$.*

4.2. BFGS Proyectado

Sea H_k la aproximación de la matriz Hessiana dada por el método del BFGS. Por simplicidad en la notación de aquí en adelante, notaremos de igual manera a su matriz reducida, es decir

$$H_{k+1} = R_{I^{\epsilon_k}(w_k)} H_k R_{I^{\epsilon_k}(w_k)} - R_{I^{\epsilon_k}(w_k)} \frac{H_k s_k s_k^T H_k}{s_k^T H_k s_k} R_{I^{\epsilon_k}(w_k)} + \frac{y_k^{\#} (y_k^{\#})^T}{s_k^T y_k^{\#}},$$

que representa la matriz de iteraciones del método BFGS proyectado ([Kelley, 1999], pp.102), con

$$y_k^{\#} = R_{I^{\epsilon_k(w_k)}} \left(\nabla f(w_{k+1}) - \nabla f(w_k) \right). \tag{3.18}$$

De aquí, las direcciones de descenso generadas por el método vienen dadas por

$$d_k = -(H_k)^{-1} \nabla f(w_k),$$

y las iteraciones se siguen como

$$w_{k+1} = P_{U_{ad}}(w_k + \alpha_k d_k). \tag{3.19}$$

Nótese que debido a (3.18) la condición de curvatura del método proyectado cambia en relación a la condición de curvatura del método no proyectado.

De manera similar al caso no proyectado existe una actualización de la matriz inversa para el método del BFGS proyectado. Sea $B_k = H_k^{-1}$, si los conjuntos activos en las iteraciones w_k y w_{k+1} no cambian, es decir, $A(w_k) = A(w_{k+1})$. Entonces se puede actualizar a la matriz B_{k+1} como sigue

$$B_{k+1} = \left(I - \frac{s_k (y_k^{\#})^T}{s_k^T y_k^{\#}} \right) B_k \left(I - \frac{y_k^{\#} s_k^T}{s_k^T y_k^{\#}} \right) + \frac{s_k s_k^T}{s_k^T y_k^{\#}}, \tag{3.20}$$

con

$$s_k^{\#} = R_{I^{\epsilon_k(w_k)}}(w_{k+1} - w_k).$$

Por otro lado, si los conjuntos activos cambian de iteración en iteración, (3.20) ya no se verifica. En su lugar, se considera una aproximación basada en la forma recursiva de las actualizaciones de la matriz BFGS ([Kelley, 1999], pp.103).

$$B_{k+1} = \left(I - \frac{s_k^{\#} (y_k^{\#})^T}{(y_k^{\#})^T s_k^{\#}} \right) R_I B_k R_I \left(I - \frac{y_k^{\#} (s_k^{\#})^T}{(y_k^{\#})^T s_k^{\#}} \right) + \frac{s_k^{\#} (s_k^{\#})^T}{(y_k^{\#})^T s_k^{\#}}, \tag{3.21}$$

donde $R_I = R_{I^{\epsilon_k(w_k)}}$.

Para encontrar direcciones de descenso para resolver el problema de localización óptima de observaciones se utilizará las actualizaciones de la matriz inversa del BFGS proyectado conjuntamente con una regla de búsqueda lineal. El conjunto sobre el cual se buscará el parámetro α_k mediante la regla de Armijo modificada es el siguiente

$$\left\{ \frac{1}{2^i \parallel \nabla f(w_0) \parallel} \text{ con } i = \{0, 1, 2, \ldots\} \right\},$$

donde $\nabla f(w_0)$ representa el gradiente del problema evaluado en un vector de localizaciones iniciales w_0 dado. Notemos que este es un conjunto factible para la búsqueda del parámetro si se toma w_0 tal que

$$0 < \frac{1}{2^k \parallel \nabla f(w_0) \parallel} < 1, \qquad \forall k > 1.$$

El criterio de parada que se utilizará para el método del BFGS proyectado es uno con pasos completos, viene dado por

$$\parallel w_k - P_{U_{ad}}(w_k + d_k) \parallel < \hat{\epsilon}, \tag{3.22}$$

donde

$$d_k = -B_k \nabla f(w_k),$$

para algún $0 < \hat{\epsilon} \ll 1$.

4.3. Convergencia del BFGS proyectado

Intuitivamente, si las iteraciones del BFGS proyectado comienzan cerca de un mínimo local no degenerado, con una buena aproximación de la matriz Hessiana y basados en el resultado del Teorema 5, se esperaría que las iteraciones del método tengan una velocidad de convergencia superlineal.

Antes de establecer este resultado, se enunciarán sin demostración algunos teorema necesarios para la identificación de los conjuntos activos y acotación del error. La demostración de los mismos puede ser encontrada en [Kelley, 1999].

Teorema 6. *Sea f dos veces continuamente diferenciable sobre el conjunto factible U_{ad}, $\bar{w}$ un punto estacionario no degenerado del problema de minimización con restricciones*

$$\min_{w \in U_{ad}} f(w),$$

y $\alpha \in]0, 1]$. Entonces para w suficientemente cercano a $\bar{w}$ se tiene que:

1. $A(w) \subseteq A(\bar{w})$ *y* $w_i = \bar{w}_i, \forall i \in A(w)$.
2. $A(w(\alpha)) = A(\bar{w})$ *y* $w(\alpha)_i = \bar{w}_i, \forall i \in A(\bar{w})$.

Donde $w(\alpha) = P_{U_{ad}}(w + \alpha d)$.

Teorema 7. *Sea* f *dos veces continuamente diferenciable sobre el conjunto factible* U_{ad}, $\bar{w}$ *un punto estacionario no degenerado. Asumiendo que las condiciones suficientes se verifican en* $\bar{w}$, *entonces existen constantes* $\delta > 0$ *y* $M > 0$ *tal que si:*

$$\| w - \bar{w} \| < \delta \quad y\ A(w) = A(\bar{w}),$$

entonces

$$\frac{\| w - \bar{w} \|}{M} \leq \| w - w(1) \| \leq M \| w - \bar{w} \| .$$

El Teorema 8 formaliza el resultado de convergencia del método quasi-Newton implementado, quedando así sentado que al igual que el método BFGS, el método del BFGS proyectado tiene una velocidad de convergencia superlineal ([Kelley, 1999], pp.104).

Teorema 8. *Sea* f *dos veces continuamente diferenciable,* $\nabla^2 f$ *localmente Lipschitz continua y sea* $\bar{w}$ *un mínimo local no degenerado. Si* w_0 *está suficientemente cerca a* $\bar{w}$, $A(w_0) = A(\bar{w})$ *y* $R_{I^{\epsilon_0}(w_0)}$ *es tomada lo suficientemente cerca a* $R_{I(\bar{w})}\nabla^2 f(\bar{w})R_{I(\bar{w})}$, *entonces las iteraciones del método BFGS proyectado, con* $\epsilon_k = \| w_k - w_k(1) \|$, *convergen q-superlinealmente a* $\bar{w}$.

Demostración. Ya que $\bar{w}$ es un mínimo local no degenerado y se toma w_0 suficientemente cerca a $\bar{w}$, del Teorema 6 tenemos que el conjunto activo está identificado, es decir

$$A(w_k) = A(w_{k+1}) = A(\bar{w}),$$

de donde se sigue también que

$$R_{A(w_k)}(w_k - \bar{w}) = R_{A(w_k)}(w_{k+1} - \bar{w}) = 0.$$

Sea ahora

$$\bar{\zeta} = \min_{i \in I(\bar{w})} \{|w_i - a_i|, |w_i - b_i|\} > 0,$$

tomando $\| w_0 - \bar{w} \| < \frac{\bar{\zeta}}{M}$, con $M > 0$ la constante del Teorema 7. Aplicando dicho resultado, se sigue que

$$\epsilon_k = \| w_k - w_k(1) \| \leq \bar{\zeta}$$

y

$$\| w_k - \bar{w} \| \leq \bar{\zeta}, \tag{3.23}$$

de donde cualquier índice $i \in A^{\epsilon_k}(w_k)$ está también en $A(w_k) = A(\bar{w})$, es decir

$$A^{\epsilon_k}(w_k) = A(w_k) = A(\bar{w}). \tag{3.24}$$

Por otro lado al tomar $R_{I^{\epsilon_0}(w_0)}$ lo suficientemente cerca a $R_{I(\bar{w})}\nabla^2 f(\bar{w})R_{I(\bar{w})}$ y al haber identificado los conjuntos activos $A^{\epsilon_0}(w_0) = A(w_0)$, se sigue inmediatamente que la aproximación inicial de la Hessiana reducida es buena, es decir

$$\| R_{I(w_0)} - R_{I(\bar{w})}\nabla^2 f(\bar{w})R_{I(\bar{w})} \| \leq \bar{\zeta}. \tag{3.25}$$

Finalmente de (3.23) y (3.25), y utilizando el Teorema 3 se concluye el resultado. □

Es importante mencionar que para poder aplicar directamente el resultado de convergencia del Teorema 8 en nuestro caso, es decir cuando se trabaja con problemas de optimización a dos niveles, donde la restricción del problema en el nivel superior es el sistema de optimalidad del nivel inferior, se debe asumir que la resolución de este sistema de optimalidad se realiza de manera exacta. En la práctica es posible hacer esta suposición si la tolerancia con la que se trabaja en el nivel inferior es mucho menor a la tolerancia exigida para el nivel superior.

5. Implementación numérica

Para la implementación numérica del problema de localización en primer lugar se debe resolver la ecuación de estado, lo que significa resolver todo el sistema de optimalidad del problema de asimilación de datos. Gracias a la velocidad de convergencia superlineal del método BFGS, se utilizará este para calcular la solución del problema de asimilación. Después, con está información se calculará el sistema adjunto, que corresponde a resolver un sistema acoplado. Luego, con los valores de los sistemas estado y adjunto se calculará la ecuación del gradiente del problema de localización y finalmente con el método del BFGS proyectado se obtendrá una dirección de descenso, necesaria para calcular el vector de localizaciones óptimas. El Algoritmo 1, muestra los pasos implementados necesarios para la resolución del problema de localización óptima.

Para todos los experimentos computacionales desarrollados en esta sección, se consideran $\Omega = (0,1) \times (0,1)$ el dominio espacial, un mallado de $m \times n = 10 \times 12$ con $h_e = 1/(m-1)$ y $h_t = 1/(n+1)$ los pasos de discretización espacial y temporal respectivamente. Para la resolución numérica del problema de asimilación de datos mediante el método del BFGS se utiliza la matriz de inicialización $B_0 = \frac{1}{2}I$, donde I representa a la matriz identidad y se toma $\alpha = 10^{-4}$ como parámetro de regularización de Tikhonov del problema en el nivel inferior. Para la resolución numérica del problema de localización óptima de observaciones mediante el método del BFGS proyectado se considera al igual que para el problema interior una matriz de inicializa-

Algoritmo 1

1. Fijar los valores de m, n, w_0, α, β, γ y $k = 0$.
2. Calcular $\nabla f(w_0)$.
3. **Repetir**
4. Hallar (y_k, p_k, u_k) solución de (3.3) utilizando el método del BFGS.
5. Hallar (η_k, σ_k) solución del sistema acoplado (3.9).
6. Calcular $\nabla f(w_k)$ como en (3.10).
7. Estimar el conjunto de índices activos de w_k.
8. Calcular B_{k+1}, actualización de la matriz inversa del BFGS proyectado, según (3.21).
9. Hallar dirección de descenso $d_k = -B_k \nabla f(w_k)$.
10. Hallar α_k, parámetro de búsqueda lineal tal que verifique (3.15), tomando los α_k en el conjunto

$$\left\{ \frac{1}{2^i \parallel \nabla f(w_0) \parallel} \text{ con } i = \{0, 1, 2, \ldots\} \right\}$$

11. Actualizar w_{k+1} según (3.19).
12. Hacer k=k+1.
13. **Hasta**. Criterio de parada dado por (3.22).

ción $B_{b0} = \frac{1}{2} I$ y una tolerancia de 10^{-2}. La Figura 3 muestra gráficamente la condición inicial que se desea reconstruir y su expresión analítica esta dada por

$$u^{Train}(x, y) = \sin(x) + y.$$

El vector de localizaciones inicial w_0 fue obtenido aleatoriamente y tiene la siguiente estructura:

Intervalo	**0's**	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	**1's**
Nº entradas $\mathbf{w_0}$	1	25	12	62	0

Cuadro 1: Estructura vector de localizaciones inicial w_0.

Donde, **0's** representa el número de entradas nulas, **1's** contará las entradas del vector de localizaciones con valor 1, $\mathbf{I_1}$, $\mathbf{I_2}$, y $\mathbf{I_3}$ representan a los intervalos $]0, 0{,}005]$, $]0{,}005, 0{,}75]$ y $]0{,}75, 1[$. Es importante señalar que una vez obtenido w_0 se fijó para todos los experimentos, es decir, todos los experimentos desarrollados trabajan con el mismo vector de localizaciones inicial.

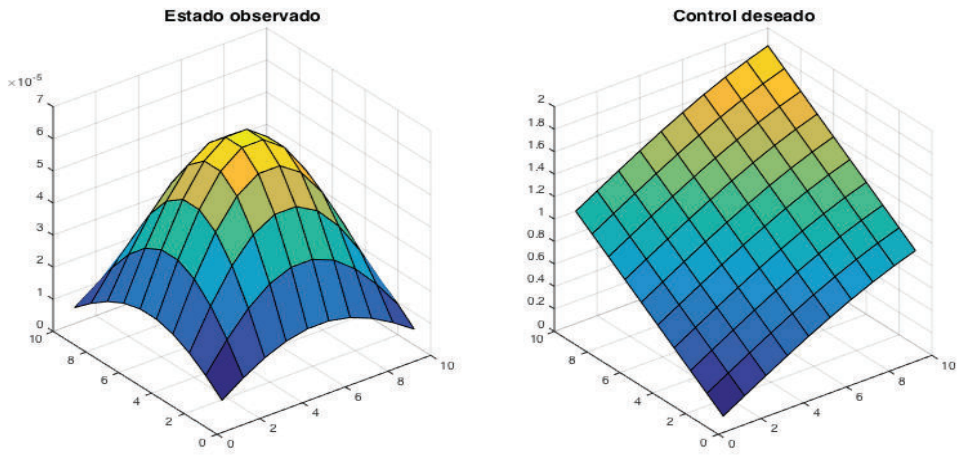

Figura 3: Control deseado y estado observado

Primer experimento

El objetivo principal de este primer experimento es observar cómo varía la estructura del vector de localizaciones óptimo al trabajar con distintos valores de los parámetros γ y β, presentes en el funcional objetivo del problema en el nivel superior. Donde γ representa un parámetro que controla en cierta manera el grado de dispersión del vector de localizaciones y β el parámetro de regularización de Tikhonov del problema de localización óptima de ubicaciones.

El segundo objetivo de este experimento es comparar los resultados que se obtienen cuando se trabaja con y sin función de penalización para inducir dispersión en la solución del problema de localización, es decir, en el vector de localizaciones óptimas w. Una característica importante de este experimento es que considera todos los puntos de la discretización espacial como posibles ubicaciones.

Tanto en este como en los siguientes experimentos se utilizó valores de γ y β comprendidos entre $10^{-4} - 1000$ y $10^{-4} - 1$, respectivamente. En el Cuadro 2 se presentan los resultados obtenidos con algunos de estos valores de los parámetros. De igual manera, para este y los siguientes experimentos se usará la siguiente notación: $\mathbf{J_0}$ y $\mathbf{J}_{\text{end}}$ representarán los valores inicial y final del funcional objetivo, mientras que **0's**, $\mathbf{I_1}$, $\mathbf{I_2}$, y $\mathbf{I_3}$, **1's** representaran los intervalos indicados anteriormente. En la tabla, los experimentos de la sección *Caso I* muestran los resultados obtenidos sin considerar la función de penalización para inducir dispersión en el vector de localizaciones óptimas y los de la sección *Caso II* muestra los resultados en los cuales se considera dicha función fijando $\epsilon = \frac{1}{8}$.

A continuación se presentan los resultados gráficos obtenidos en este experimento,

las Figuras (4) y (5) muestran el decrecimiento de la función objetivo para diferentes valores de γ con $\beta = 10^{-4}$ fijo para el Caso I y $\beta = 10^{-2}$ fijo para el Caso II. Las Figuras (6) y (7) muestran la reconstrucción de la condición inicial deseada cuando se varía el valor de γ y se fija β en los valores señalados para el Caso I y el Caso II. Finalmente, las Figuras (8) y (9) muestra la estructura del vector de localizaciones óptimas obtenido para diferentes valores de γ, tanto para el Caso I como para el Caso II. De aquí en adelante, un indicador morado representa una componente de w que pertenece al intervalo I_1, uno verde a una componente de w que está en I_2, uno azul si el valor pertenece al intervalo I_3, y uno rojo si el punto representa una entrada del vector de localizaciones que toma el valor de 1.

	γ	β	**0's**	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	**1's**	$\mathbf{J_0}$	$\mathbf{J_{end}}$	**iter**	$\| \mathbf{w} \|$
Caso I	10^{-3}	10^{-4}	68	0	0	1	31	0,063	0,033	6	31,88
	10^{-2}		67	0	0	0	33	0,62	0,33	21	33
	0,1		57	0	0	0	43	6,19	4,30	21	43
	1		95	0	0	0	5	61,79	5,01	9	5
	10		62	0	1	0	37	617,98	375,37	6	37,54
	100		75	25	0	0	0	$6{,}17 \times 10^3$	0,021	17	$7{,}33 \times 10^{-12}$
Caso I	10^{-3}	10^{-3}	60	0	0	0	40	0,063	0,041	9	40
	10^{-2}		59	0	0	1	40	0,62	0,41	7	40,95
	0,1		62	0	0	0	38	6,18	3,80	9	38
	1		47	0	0	0	53	61,79	53,01	7	53
	10		95	0	0	0	5	617,98	50,00	27	5
	100		74	26	0	0	0	$6{,}18 \times 10^3$	0,076	44	$5{,}47 \times 10^{-4}$
	γ	β	**0's**	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	**1's**	$\mathbf{J_0}$	$\mathbf{J_{end}}$	**iter**	$\| \mathbf{w} \|$
Caso II	10^{-3}	10^{-4}	45	1	0	0	54	0,074	0,055	8	54
	10^{-2}		62	0	0	0	38	0,73	0,38	10	38
	0,1		62	0	0	0	38	7.32	3,80	7	38
	1		72	0	1	0	27	73,21	28,00	7	27,24
	10^*		27	0	11	62	0	732,05	730,00	3	61,80
	100^*		27	0	11	62	0	$7{,}32 \times 10^3$	$7{,}30 \times 10^3$	3	61,78
Caso II	10^{-3}	10^{-2}	61	0	0	0	39	0,074	0,041	9	39
	10^{-2}		62	0	0	0	38	0,73	0,38	11	38
	0,1		45	0	0	1	54	7,32	5,50	7	54,89
	1		54	0	1	0	45	73,21	46,00	7	45,49
	10		64	0	0	0	36	732,04	360,00	7	36
	100^*		27	0	11	62	0	$7{,}32 \times 10^3$	$7{,}30 \times 10^3$	3	61,84

Cuadro 2: Experimento 1 - Diferentes valores de γ y β.

Los valores atípicos obtenidos para el *Caso II*, indicados en la tabla con el superíndice *, se explican ya que para dicha elección de parámetros la tolerancia con la que se trabaja es muy poco restrictiva y el algoritmo converge casi sin haber realizado iteraciones, por lo cual la estructura de w es similar a la de w_0.

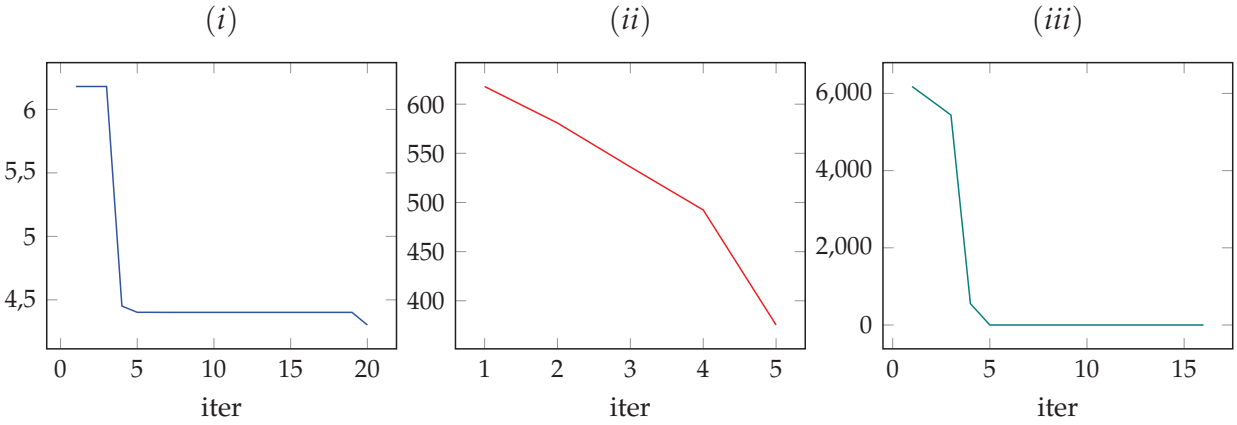

Figura 4: Decrecimiento de la función objetivo *Caso I* - diferentes valores de γ. $(i)\gamma = 0{,}1$, $(ii)\gamma = 10$, $(iii)\gamma = 100$.

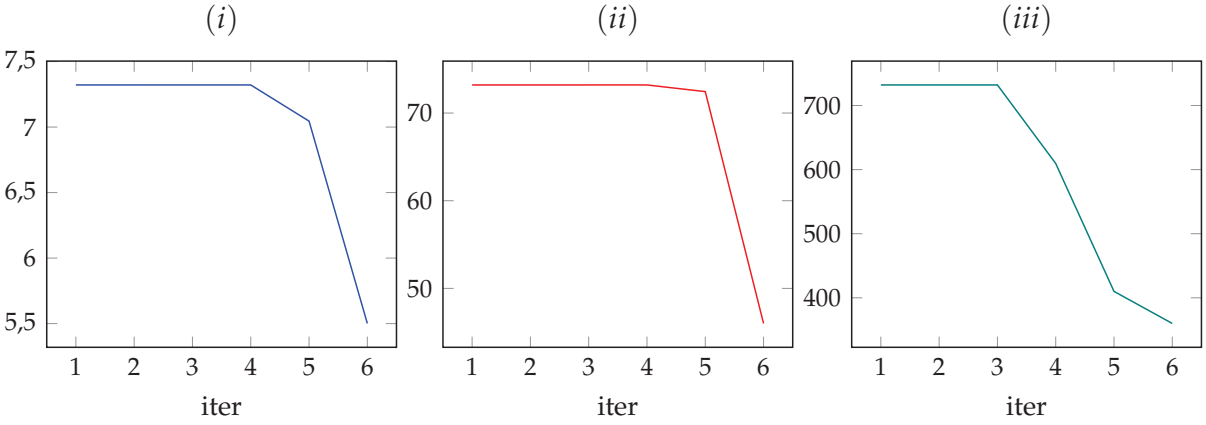

Figura 5: Decrecimiento de la función objetivo *Caso II* - diferentes valores de γ. Función de penalización. $(i)\gamma = 0{,}1$, $(ii)\gamma = 1$, $(iii)\gamma = 10$.

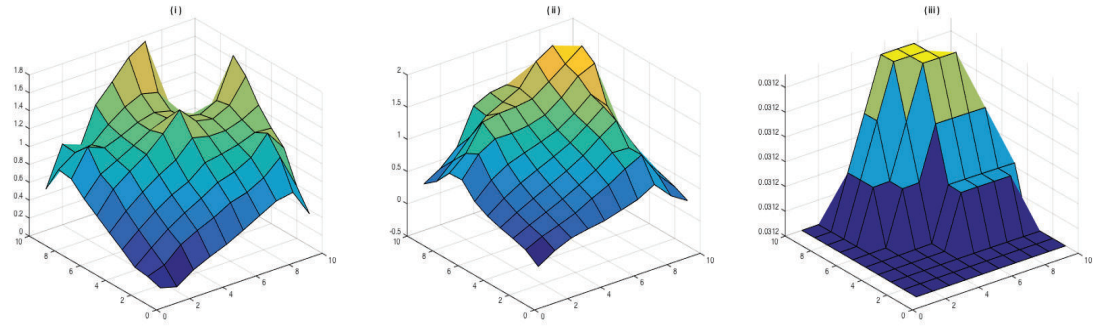

Figura 6: Control óptimo *Caso I* - diferentes valores de γ. $(i)\gamma = 0{,}1$, $(ii)\gamma = 10$, $(iii)\gamma = 100$.

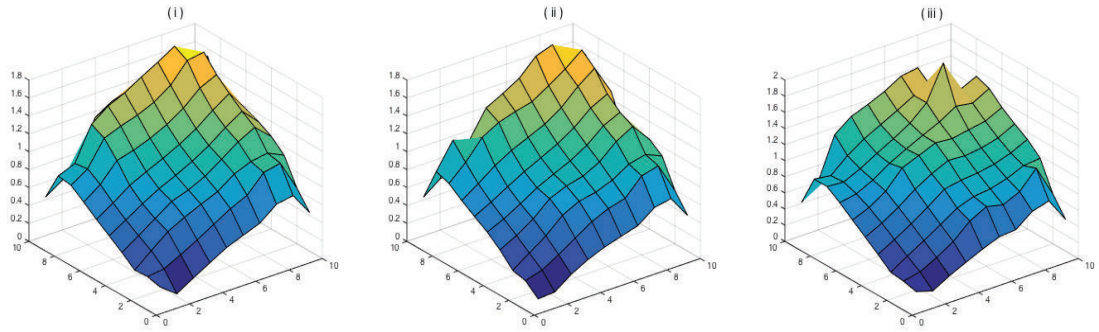

Figura 7: Control óptimo *Caso II* - diferentes valores de γ. Función de penalización. $(i)\gamma = 0{,}1$, $(ii)\gamma = 1$, $(iii)\gamma = 10$.

Tanto en el Caso I y el Caso II, se puede observar que la reconstrucción de la condición inicial deseada es más exacta cuando el vector de localizaciones w tiene menos entradas nulas.

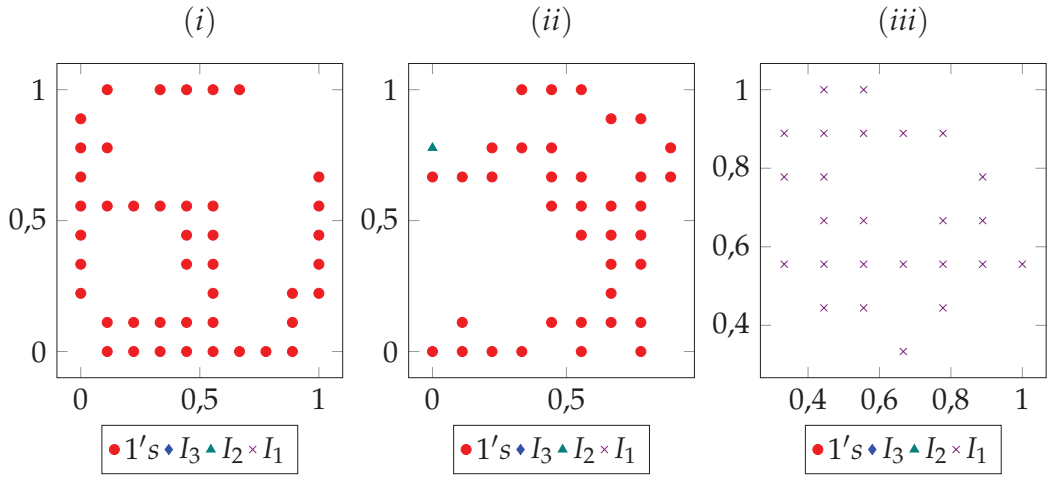

Figura 8: Estructura del vector de localizaciones *Caso I* - diferentes valores de γ. $(i)\gamma = 0{,}1$, $(ii)\gamma = 10$, $(iii)\gamma = 100$.

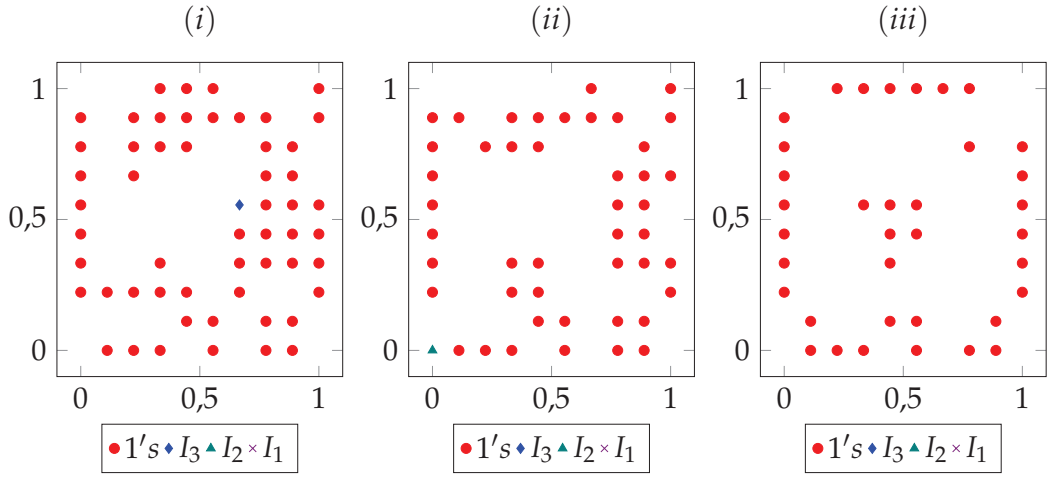

Figura 9: Estructura del vector de localizaciones *Caso II* - diferentes valores de γ. Función de penalización. $(i)\gamma = 0{,}1$, $(ii)\gamma = 1$, $(iii)\gamma = 10$.

Gráficamente se observa que la estructura del vector de localizaciones óptimas obtenida con los diferentes valores de los parámetros γ y β para los casos I y II es distinta. Se realizó además una comparación más detallada midiendo el tiempo de ejecución del algoritmo en cada caso, además de una comparación de los errores absoluto y relativo entre la condición inicial deseada u^{Train} y la condición inicial óptima obtenida u, los resultados obtenidos se presentan en el Cuadro 3, en donde:

$$\text{error abs.} = \| u^{Train} - u \|_2$$

y

$$\text{error rel.} = \frac{\| u^{Train} - u \|_2}{\| u^{Train} \|_2}.$$

Caso	γ	β	tiempo (s)	error abs.	error rel.
	0,1		1375,49	3,38	0,32
Caso I	10	10^{-4}	129,76	3,73	0,36
	100		241,65	10,15	0,97
	0,1		161,79	2,08	0,19
Caso II	1	10^{-2}	162,06	2,24	0,21
	10		153,36	2,37	0,23

Cuadro 3: Experimento 1 - Comparación *Caso I* y *Caso II*

De la realización de este experimento podemos concluir que aunque con la utilización de la función de penalización para inducir dispersión no se obtenga un vector

de localización óptima w con una visible mayor cantidad de entradas nulas, los tiempos de ejecución son menores y además la reconstrucción de la condición inicial es más precisa lo cual se puede observar además de gráficamente, en los valores de los errores absoluto y relativo obtenidos.

Segundo experimento

Siempre hay lugares donde la colocación de un sensor puede ser difícil o costosa. Para considerar esta variable y obtener un experimento más cercano a la realidad se fija un subconjunto de localizaciones como posibles ubicaciones. Para este experimento vamos a considerar seis puntos dentro del cuadrado unidad, que es el dominio en el que estamos trabajando. Teniendo en cuenta que los puntos dados no tienen porque coincidir con los puntos de la malla se tomará el punto del mallado más cercano a la ubicación dada. Los puntos que se van a tomar son los siguientes:

$$\begin{aligned} x_1 &= (0{,}1, 0{,}3) & x_4 &= (0{,}7, 0{,}2) \\ x_2 &= (0{,}2, 0{,}2) & x_5 &= (0{,}8, 0{,}8) \\ x_3 &= (0{,}5, 0{,}5) & x_6 &= (0{,}6, 0{,}9), \end{aligned} \tag{3.26}$$

su ubicación en la malla, considerando los mismos parámetros que en el primer experimento se puede observar gráficamente en la Figura 10.

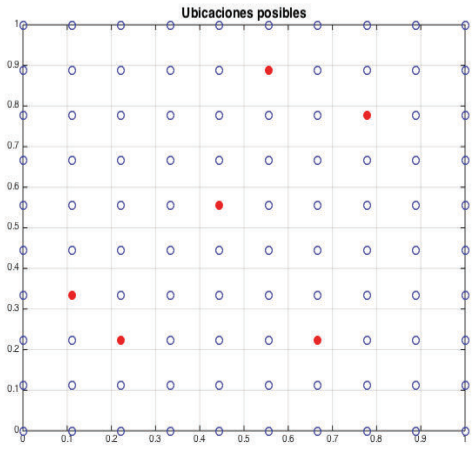

Figura 10: Subconjunto de posibles ubicaciones.

Al igual que el primero, en este segundo experimento se busca observar la variación de la estructura de localizaciones óptimas cuando se trabaja con distintos valores de γ y β. Además de comparar los resultados obtenidos cuando se trabaja sin y con

función de penalización para inducir dispersión en el vector de localizaciones resultante (*CasoI* y *Caso II* respectivamente) cuando se considera un subconjunto reducido de posibles localizaciones.

El Cuadro 4 muestra los resultados obtenidos con diferentes valores de γ y β para los casos estudiados. Notemos que como solo estamos considerando seis puntos como posibles ubicaciones, el resultado será un vector de localizaciones w con muchas menos entradas iguales a uno, si lo comparamos con el obtenido en el experimento anterior. De igual manera y por la misma razón, la reconstrucción de la condición inicial será menos exacta. Al igual que en el primer experimento para el segundo caso se fija $\epsilon = \frac{1}{8}$.

	γ	β	**0's**	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	**1's**	$\mathbf{J_0}$	$\mathbf{J_{end}}$	**iter**	$\parallel \mathbf{w} \parallel$
Caso I	10^{-3}	10^{-4}	3	0	0	0	3	0,011	0,009	7	3
	10^{-2}		0	0	1	5	0	0,065	0,056	6	5,07
	0,1		4	0	0	0	2	0,605	0,208	5	2
	1		2	0	0	1	3	6,01	3,82	6	3,80
	10		2	0	2	1	1	60,01	25,63	9	2,55
	100		3	1	0	1	1	600,01	200,01	9	2,00
Caso I	10^{-3}	10^{-3}	0	0	0	0	6	0,011	0,011	2	6
	10^{-2}		0	0	0	0	6	0,065	0,065	2	6
	0,1		0	0	0	0	6	0,605	0,605	2	6
	1		0	0	0	2	4	6,01	5,95	3	5,94
	10		2	1	0	2	1	60,01	29,85	11	2,98
	100		3	0	0	1	2	600,01	281,65	8	2,82
	γ	β	**0's**	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	**1's**	$\mathbf{J_0}$	$\mathbf{J_{end}}$	**iter**	$\parallel \mathbf{w} \parallel$
Caso II	10^{-3}	10^{-4}	4	0	1	0	1	0,011	0,009	18	1,63
	10^{-2}		4	0	0	0	2	0,065	0,027	15	2
	0,1		4	0	0	0	2	0,605	0,207	15	2
	1		4	0	0	0	2	6,01	2,01	15	2
	10		2	2	0	2	0	60,01	20,02	17	2
	100		2	1	2	0	1	600,01	202,64	22	1,17
Caso II	10^{-3}	10^{-3}	0	0	0	0	6	0,001	0,001	2	6
	10^{-2}		0	0	0	0	6	0,065	0,065	2	6
	0,1		0	0	0	0	6	0,605	0,605	2	6
	1		0	0	0	0	6	6,01	6,01	2	6
	10		0	0	0	0	6	60,01	60,01	2	6
	100		0	0	0	0	6	600,01	600,01	2	6

Cuadro 4: Experimento 2 - Puntos dados en la malla.

Los resultados gráficos que se presentan a continuación fueron obtenidos al trabajar con distintos valores de γ y fijando $\beta = 10^{-4}$ tanto para el caso I como para el caso

II. Las Figuras (11) y (12) muestran el decrecimiento de la función objetivo al trabajar con los parámetros indicados. Las Figuras (13) y (14) muestran la reconstrucción de la condición inicial deseada cuando se varía el valor de γ y se fija β en los valores señalados para el Caso I y el Caso II, respectivamente. Notemos que al trabajar con tan pocos puntos no se puede esperar una reconstrucción precisa de la condición inicial deseada y esto es justamente lo que se refleja gráficamente.

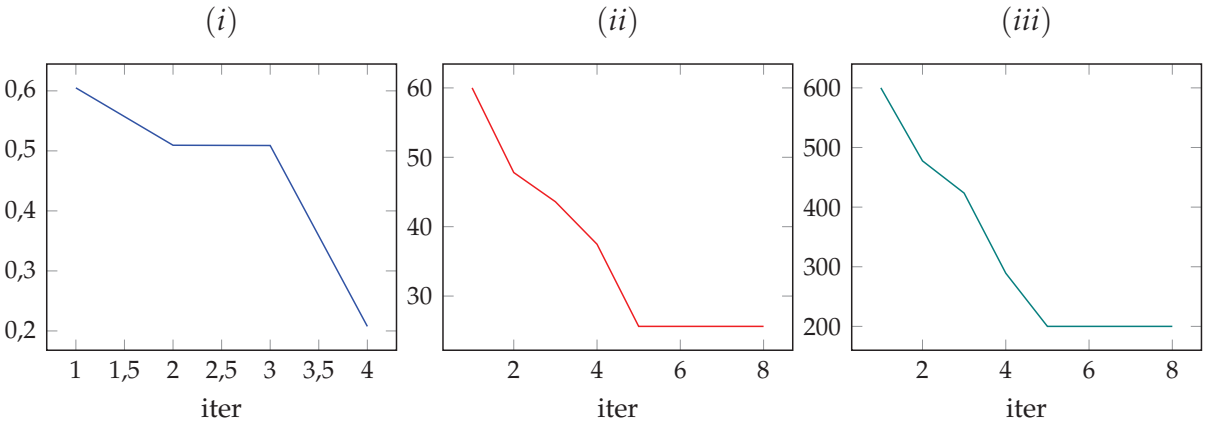

Figura 11: Decrecimiento de la función objetivo *Caso I* - diferentes valores de γ. Puntos dados en la malla. $(i)\gamma = 0{,}1$, $(ii)\gamma = 10$, $(iii)\gamma = 100$.

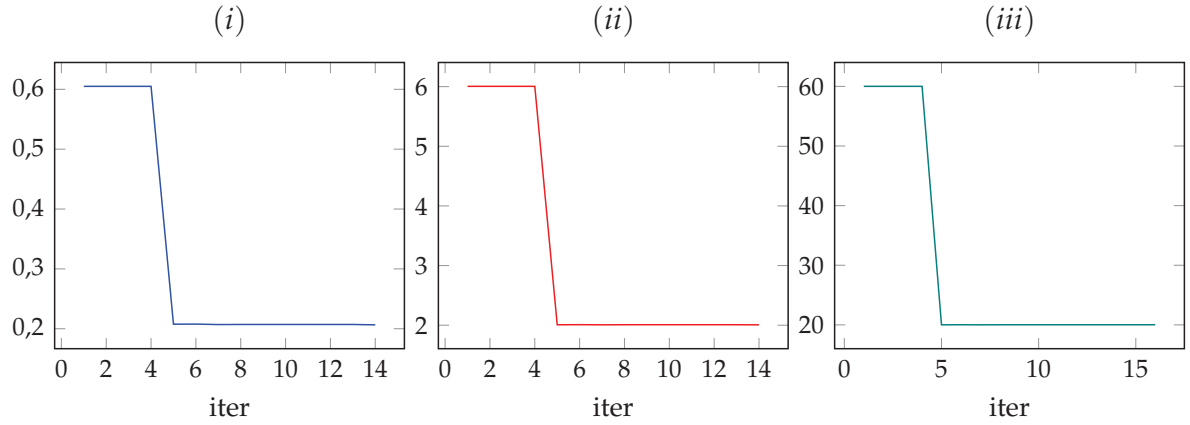

Figura 12: Decrecimiento de la función objetivo *Caso II* - diferentes valores de γ. Puntos dados en la malla. $(i)\gamma = 0{,}1$, $(ii)\gamma = 1$, $(iii)\gamma = 10$.

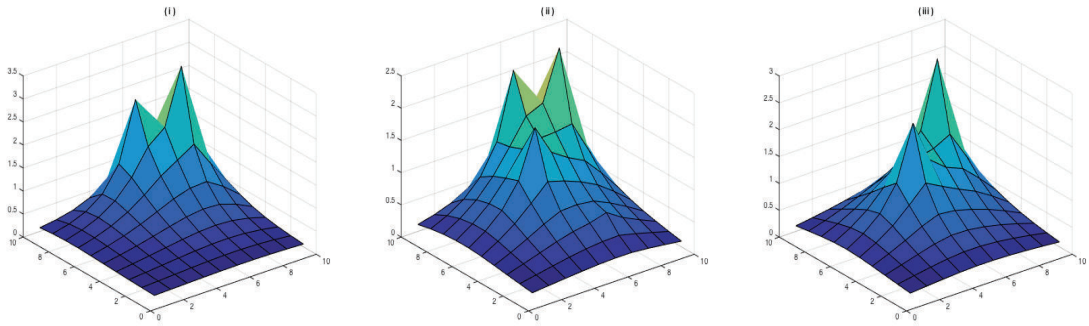

Figura 13: Control óptimo *Caso I* - diferentes valores de γ. Puntos dados en la malla.$(i)\gamma = 0{,}1$, $(ii)\gamma = 10$, $(iii)\gamma = 100$.

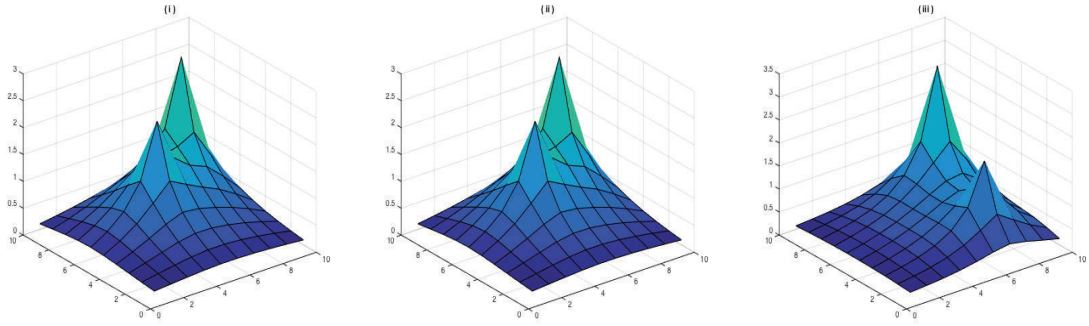

Figura 14: Control óptimo *Caso II* - Puntos dados en la malla. $(i)\gamma = 0{,}1$, $(ii)\gamma = 1$, $(iii)\gamma = 10$.

Las Figuras (15) y (16) muestran la estructura del vector de localizaciones óptimas obtenido para diferentes valores de γ, tanto para el Caso I como para el Caso II.

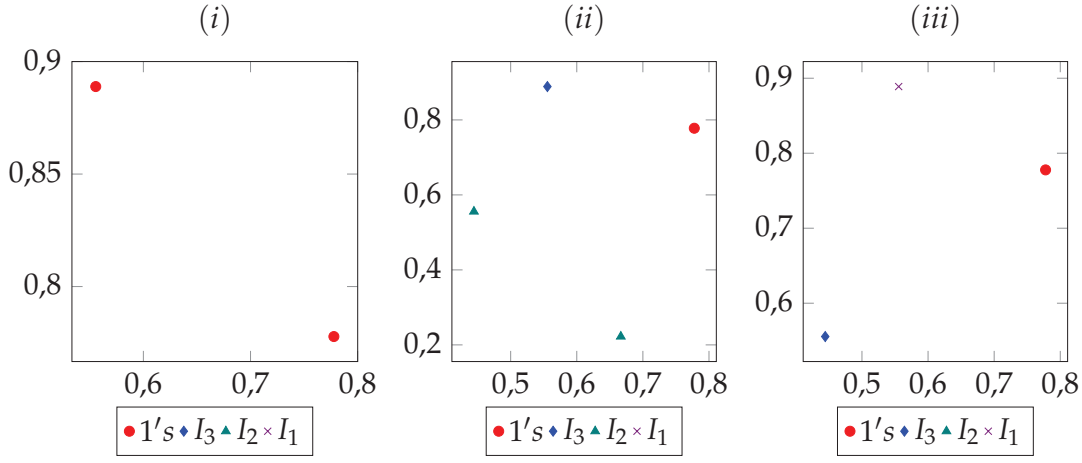

Figura 15: Estructura del vector de localizaciones *Caso I* - diferentes valores de γ. $(i)\gamma = 0{,}1$, $(ii)\gamma = 10$, $(iii)\gamma = 100$.

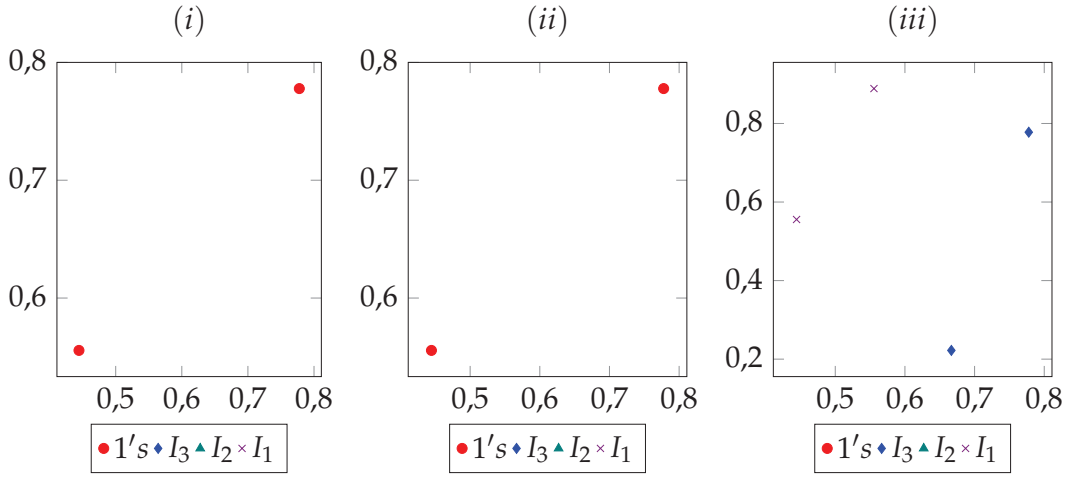

Figura 16: Estructura del vector de localizaciones *Caso II* - diferentes valores de γ. $(i)\gamma = 0{,}1$, $(ii)\gamma = 1$, $(iii)\gamma = 10$.

A diferencia del experimento anterior, para este caso con la utilización de la función de penalización se obtiene un vector solución con más entradas nulas. En el Cuadro 5 se presentan los resultados obtenidos al comparar los tiempos de ejecución y los errores absoluto y relativo entre la condición inicial deseada u^{Train} y la condición inicial óptima obtenida u en cada caso.

Caso	γ	β	tiempo (s)	error abs.	error rel.
	0,1		11,55	6,48	0,62
Caso I	10	10^{-4}	64,67	5,74	0,55
	100		73,81	6,18	0,59
	0,1		223,35	6,18	0,59
Caso II	1	10^{-4}	223,51	6,18	0,59
	10		300,44	6,30	0,60

Cuadro 5: Experimento 2 - Comparación puntos dados en la malla *Caso I* y *Caso II*

En este caso debido a que la cantidad de ubicaciones que son tomadas como posibles es muy reducida, la reconstrucción es menos precisa, debido a lo cual se obtiene un valor más elevado del error tanto absoluto como relativo.

Una variante de este experimento es fijar el número máximo de sensores que se pueden instalar dentro del pequeño subconjunto de ubicaciones preestablecidas, dadas por (3.26). Vamos a fijar en 3 la cantidad de sensores a ser instalados. Para obtener la ubicación óptima de estos tres sensores se halló experimentalmente los parámetros

con los que se debía trabajar, fijándolos en $\gamma = 7$ y $\beta = 10^{-3}$ respectivamente. El Cuadro 6 muestra los resultados obtenidos con estos parámetros.

	γ	β	**0's**	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	**1's**	$\mathbf{J_0}$	$\mathbf{J_{end}}$	**iter**	$\| \mathbf{w} \|$
Caso I	7	10^{-3}	3	0	0	0	3	42,01	21,01	7	3

Cuadro 6: Experimento 2 - Puntos dados en la malla con máximo fijo.

Es importante señalar que el control que se tienen sobre el número de sensores a instalarse, preestablecido como 3 en este experimento, es indirecto y depende de la elección de parámetros que se realice. Gráficamente la Figura 17 muestra la reconstrucción de la condición inicial cuando solo se considera la información de tres sensores y la Figura 18 muestra donde se encuentran ubicados.

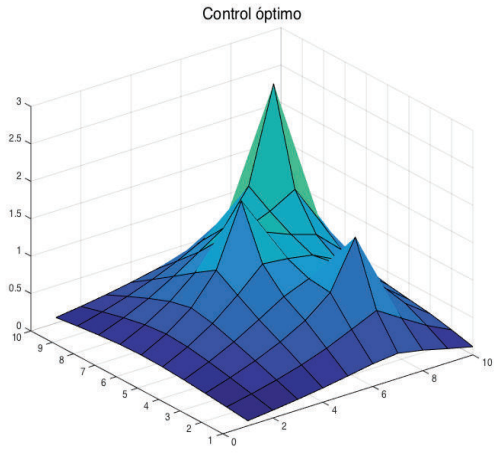

Figura 17: Reconstrucción u^{Train} - Puntos dados en la malla con máximo fijo.

Como hemos podido apreciar a través de los experimentos realizados, la convergencia del algoritmo, la reconstrucción de la condición inicial y el hecho de que la función objetivo decrezca su valor en cada iteración, dependen de los parámetros con los que se trabajen.

En el primer experimento, aunque la utilización de la función de penalización no resulte en un vector de localizaciones óptimas con más entradas nulas al compararlo con aquellos casos en los que no se utiliza dicha función, su utilización permite una mejor reconstrucción de la condición inicial deseada lo cual se verifica al comparar el valor del error en cada caso, resultado que se intuía también de la comparación gráfica realizada.

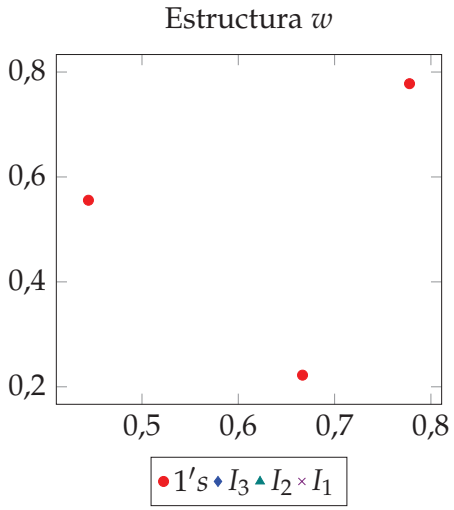

Figura 18: Estructura w óptimo - Puntos dados en la malla con máximo fijo.

El segundo experimento considera un subconjunto pequeño de puntos como posibles ubicaciones, este caso se apega más a la realidad ya que en la práctica la mayoría de las veces las observaciones no pueden ser recolectadas en cualquier punto del dominio, limitando de esta manera el número de lugares donde ciertamente estas pueden ser tomadas. En este caso, debido a la pequeña cantidad de puntos con los que se cuenta, el vector óptimo tendrá muchas más entradas nulas, haciendo que la reconstrucción de la condición inicial sea menos precisa que la obtenida en el experimento anterior. Resultado que se visualiza gráficamente y se comprueba al analizar los errores obtenidos en cada experimento.

6. Conjunto de Entrenamiento - Training Set

Los problemas de aprendizaje en el aprendizaje de máquinas o Machine Learning consideran una muestra de N datos, que recibe el nombre de conjunto de entrenamiento o Training Set, y aplicando la información aprendida de esta muestra tratan de predecir propiedades de un dato desconocido. En nuestro caso, el conjunto de entrenamiento estará formado por

$$\begin{gathered}\left(u_1^{Train}, y_1^{Train}\right)\\ \left(u_2^{Train}, y_2^{Train}\right)\\ \vdots\\ \left(u_N^{Train}, y_N^{Train}\right)\end{gathered}$$

donde $\left(u_j^{Train}, y_j^{Train}\right), \forall j = 1, \ldots, N$ representan simulaciones de la condición inicial real y el estado observado resultante con dicha condición inicial, respectivamente.

En las secciones anteriores, se planteó y resolvió numéricamente el problema de localización de observaciones con $j = 1$, es decir, solo con un elemento del conjunto de entrenamiento, se obtuvo un vector de localizaciones w tal que el control óptimo obtenido aproxime al control o condición inicial simulada y el estado óptimo al estado observado, es decir se obtuvo una aproximación o reconstrucción de $\left(u_1^{Train}, y_1^{Train}\right)$. La idea de usar un conjunto de entrenamiento con $j > 1$ es encontrar un vector solución w tal que las localizaciones dadas por él sean en promedio óptimas para todos los pares de entrenamiento considerados. En otras palabras lo que se busca aprender al usar el conjunto de entrenamiento es precisamente el vector de localizaciones para con esto obtener un resultado de ubicaciones más robusto.

Si se toma más de un par de entrenamiento el problema a resolver está dado por

$$\min_{0\leq w\leq 1} J(y,p,u,w) = \sum_{j=1}^{N} \parallel y_j^{Train} - y_j \parallel^2 + \beta \sum_{j=1}^{N} \parallel u_j^{Train} - u_j \parallel^2 + \gamma \sum_k w_k \qquad (3.27)$$

donde (y,p,u,w) es la solución de:

$$\begin{aligned}
\frac{\partial y_j}{\partial t} - \Delta y_j &= 0 && \text{en } \Omega\times]0,T[\\
y_j &= 0 && \text{en } \Gamma\times]0,T[\\
y_j(0) &= u_j && \text{en } \Omega \\
-\frac{\partial p_j}{\partial t} - \Delta p_j &= \sum_k \sum_i w_k \left[y_j^{Train}(x,t) - y_j(x,t)\right] \otimes \delta(x-x_k) \otimes \delta(t-t_i) && \text{en } \Omega\times]0,T[\\
p_j &= 0 && \text{en } \Gamma\times]0,T[\\
p_j(T) &= 0 && \text{en } \Omega \\
\alpha u_j - p_j(0) &= 0 && \text{en } \Omega.
\end{aligned}$$

$\forall j = 1,\ldots,N.$

6.1. Sistema de optimalidad - Training Set

Al igual que cuando considerabamos $j = 1$, el sistema de optimalidad se obtiene utilizando el enfoque Lagrangiano y es similar al dado por (3.3) y (3.9) - (3.12), pero esta vez el número de sistemas a resolver es mayor.

Ecuación de estado: $\forall j = 1, \ldots, N$

$$\begin{aligned}
\frac{\partial y_j}{\partial t} - \Delta y_j &= 0 && \text{en } \Omega\times]0,T[\\
y_j &= 0 && \text{en } \Gamma\times]0,T[\\
y_j(0) &= u_j && \text{en } \Omega \\
-\frac{\partial p_j}{\partial t} - \Delta p_j &= \sum_k \sum_i w_k \left[y_j^{Train}(x,t) - y_j(x,t) \right] \otimes \delta(x - x_k) \otimes \delta(t - t_i) && \text{en } \Omega\times]0,T[\\
p_j &= 0 && \text{en } \Gamma\times]0,T[\\
p_j(T) &= 0 && \text{en } \Omega \\
\alpha u_j - p_j(0) &= 0 && \text{en } \Omega.
\end{aligned} \tag{3.28}$$

Ecuación adjunta: $\forall j = 1, \ldots, N$

$$\begin{aligned}
-\frac{\partial \eta_j}{\partial t} - \Delta \eta_j &= 2(y_j^{Train} - y_j) - \sum_k \sum_i w_k \sigma_j(x,t) \otimes \delta(x - x_k) \otimes \delta(t - t_i) && \text{en } \Omega\times]0,T[\\
\eta_j &= 0 && \text{en } \Gamma\times]0,T[\\
\eta_j(T) &= 0 && \text{en } \Omega \\
\frac{\partial \sigma_j}{\partial t} - \Delta \sigma_j &= 0 && \text{en } \Omega\times]0,T[\\
\sigma_j &= 0 && \text{en } \Gamma\times]0,T[\\
\sigma_j(0) &= \tfrac{1}{\alpha} \left[2\beta(u_j^{Train} - u_j) + \eta_j(0) \right] && \text{en } \Omega.
\end{aligned} \tag{3.29}$$

Ecuación del gradiente:

$$\nabla f(w)_k = \gamma - \sum_{j=1}^{N} \sum_{i}^{\mu_T} \sigma_j(x_k, t_i) \left(y_j^{Train}(x_k, t_i) - y_j(x_k, t_i) \right), \quad \forall k = 1, \ldots, \mu_E, \tag{3.30}$$

donde (3.30) satisface el sistema de complementariedad dado por (3.31)

$$\begin{cases}
(\lambda_a)_k \geq 0, (\lambda_b)_k \geq 0 & , \forall k = 1, \ldots, \mu_E \\
(\lambda_a)_k (0 - w_k) = (\lambda_b)_k (w_k - 1) = 0 & , \forall k = 1, \ldots, \mu_E \\
0 \leq w_k \leq 1 & , \forall k = 1, \ldots, \mu_E
\end{cases} \tag{3.31}$$

con

$$\begin{aligned}
\lambda_a &= \max\{0, \nabla f(w)\} \\
\lambda_b &= |\min\{0, \nabla f(w)\}|.
\end{aligned}$$

6.2. Implementación numérica - Training Set

Para esta sección, en primer lugar se establecerán los pares de entrenamiento con los que se va a trabajar. Como el estado observado y^{Train} se construye a partir de la condición inicial simulada u^{Train}, solo se debe fijar el conjunto de controles deseados $u_j^{Train}, \forall j = 1, \ldots, N$. Para la implementación numérica se consideraron $N = 10$ pares de entrenamiento, dados por:

$$\begin{aligned}
u_1^{Train} &= \sin(x) + y & u_6^{Train} &= \sin(x)\cos(y) \\
u_2^{Train} &= x^3 + xy - y & u_7^{Train} &= \sin(2\pi x)\sin(2\pi y) \\
u_3^{Train} &= \exp(y) + cos(x) & u_8^{Train} &= \log(y\cos(x) + 1) \\
u_4^{Train} &= \exp(\sin(xy)) & u_9^{Train} &= x^2\cos(y + 1) \\
u_5^{Train} &= \log(xy + 1) & u_{10}^{Train} &= x^2y^3 - 1.
\end{aligned}$$

El algoritmo utilizado para la resolución del problema de localización óptima considerando varios pares de entrenamiento es similar al desarrollado en la sección anterior cuando se consideraba $j = 1$, pero esta vez tenemos que resolver N sistemas de EDPs para encontrar la solución de la ecuación de estado (3.28) y N sistemas de EDPs para resolver la ecuación adjunta (3.29). Notemos sin embargo que, para cada $j = 1, \ldots, N$ los sistemas de la ecuación de estado y la ecuación adjunta son independientes el uno del otro. Tomando ventaja de esta estructura se puede resolver cada sistema de forma paralela. Luego con esta información se calculará la ecuación del gradiente del problema de localización y finalmente usando el BFGS proyectado se obtendrá la dirección de descenso, necesaria para calcular el vector de localizaciones óptimas. Esta vez el vector w obtenido será en promedio óptimo para cada $j = 1, \ldots, N$ de tal forma que en cada caso el par $\left(u_j^{Train}, y_j^{Train}\right)$ será reconstruido por el control y estado óptimos (u_j, y_j) obtenidos.

Siguiendo el esquema de la sección anterior, los experimentos numéricos de esta utilizarán los mismos parámetros de inicialización, es decir, se tomará $\Omega = (0,1) \times (0,1)$ como dominio espacial, un mallado de $m \times n = 10 \times 12$ con $h_e = 1/(m-1)$ y $h_t = 1/(n+1)$ los pasos de discretización espacial y temporal respectivamente. La matriz $B_0 = \frac{1}{2}I$ como matriz de inicialización para el BFGS utilizado en la resolución del problema en el nivel interior y $\alpha = 10^{-4}$ como parámetro de regularización de Tikhonov. Para la inicialización del BFGS proyectado se toma $B_{b0} = \frac{1}{2}I$ y se trabaja con una tolerancia de 10^{-2}.

Primer experimento - Training Set

El objetivo principal de este experimento es observar la estructura del vector de localizaciones, el número de iteraciones que realiza el algoritmo, y el valor en norma del vector w obtenido al trabajar con distintos valores de γ y β. Se compararon los resultados obtenidos cuando se trabaja con y sin función de penalización para inducir dispersión en la solución del problema de localización. Nuevamente, para este primer experimento se consideran todos los puntos del dominio como posibles ubicaciones.

El Cuadro 7 muestra los resultados obtenidos con diferentes valores de los parámetros γ y β para los casos estudiados. En los experimentos realizados se trabaja sin y con función de penalización para inducir dispersión en el vector de localizaciones (*Caso I* y *Caso II* respectivamente), fijando para el caso II $\epsilon = \frac{1}{8}$.

	γ	β	**0's**	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	**1's**	$\mathbf{J_0}$	$\mathbf{J}_{\text{end}}$	**iter**	$\|\mathbf{w}\|$
Caso I	0,1	10^{-4}	67	0	0	0	33	6,187	3,314	7	33
	10		89	0	0	0	11	617,99	110,06	11	11
Caso I	10^{-2}	10^{-3}	76	0	0	0	24	0,625	0,256	7	24
	10		85	0	0	0	15	617,99	150,05	7	15
	γ	β	**0's**	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	**1's**	$\mathbf{J_0}$	$\mathbf{J}_{\text{end}}$	**iter**	$\|\mathbf{w}\|$
Caso II	10^{-2}	10^{-4}	68	0	0	0	32	0,739	0,334	7	32
	1		67	0	0	0	33	73,21	33,01	7	33
Caso II	10^{-2}	10^{-3}	75	0	0	0	25	0,739	0,265	7	25
	10		85	0	0	0	15	732,05	150,03	7	15

Cuadro 7: Training set. Experimento 1 - Distintos valores de γ y β.

De los resultados obtenidos en el Cuadro 7 se observa nuevamente que mientras más alto es el valor de γ el vector de localizaciones óptimas promedio w tiene más entradas nulas.

A continuación se presentan los resultados gráficos obtenidos en este experimento, las Figuras (19) y (20) muestran el decrecimiento de la función objetivo para los casos I y II respectivamente, cuando se trabaja con dos valores de γ y fijando $\beta = 10^{-3}$ en ambos casos.

Las Figuras (21) y (22) muestras la estructura del vector de localizaciones óptimas promedio para todos los pares considerados, cuando se trabaja sin función de penalización para inducir dispersión y con dicha función respectivamente. Fijando $\beta = 10^{-3}$ en ambos casos y trabajando con dos valores de γ.

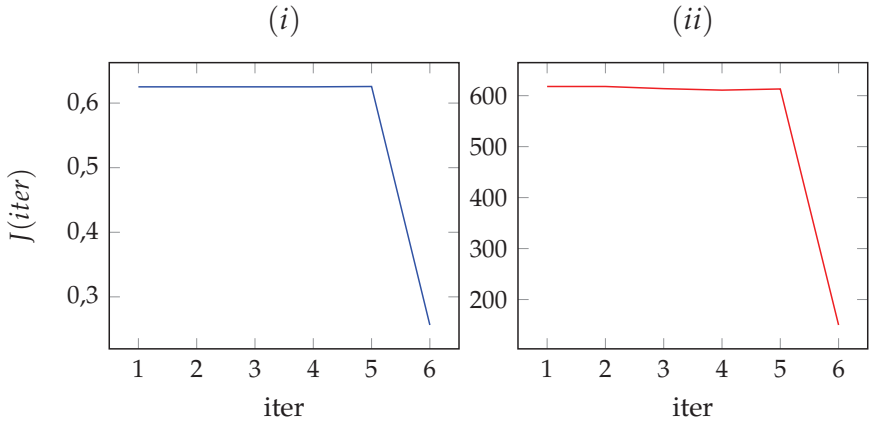

Figura 19: Decrecimiento de la función objetivo Training Set *Caso I* - diferentes valores de γ. $(i)\gamma = 10^{-2}$, $(ii)\gamma = 10$.

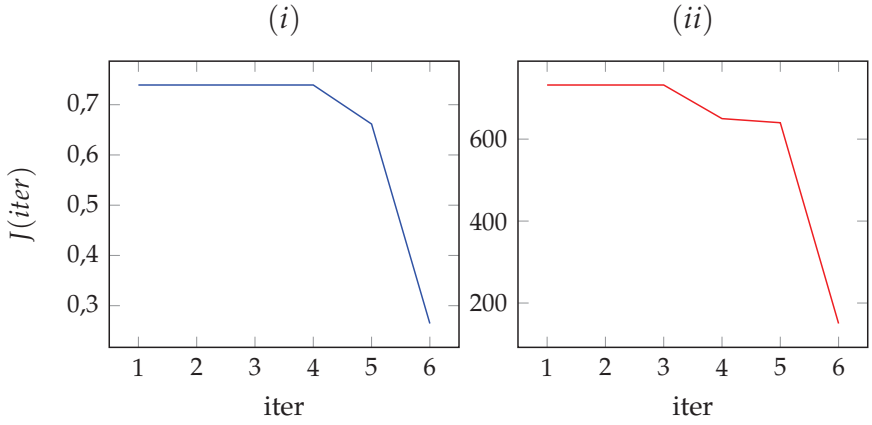

Figura 20: Decrecimiento de la función objetivo Training Set *Caso II* - diferentes valores de γ. Función de penalización. $(i)\gamma = 10^{-2}$, $(ii)\gamma = 10$.

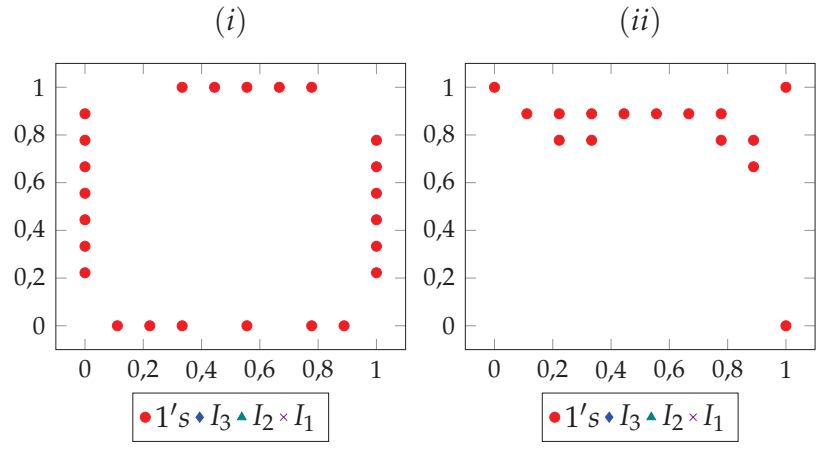

Figura 21: Estructura del vector de localizaciones Training Set *Caso I* - diferentes valores de γ. $(i)\gamma = 10^{-2}$, $(ii)\gamma = 10$.

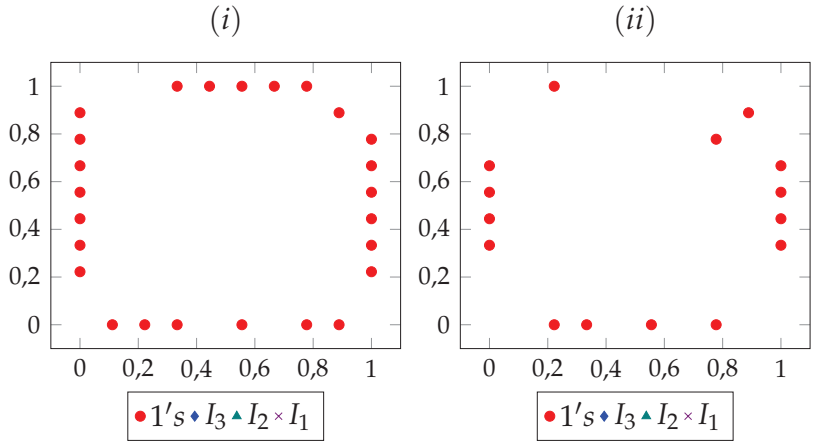

Figura 22: Estructura del vector de localizaciones Training Set *Caso II* - diferentes valores de γ. Función de penalización. $(i)\gamma = 10^{-2}$, $(ii)\gamma = 10$.

Gráficamente se observa que la estructura del vector de localizaciones óptimas obtenida con los parámetros indicados es distinta para los casos I y II. Se realizó además una comparación más detallada midiendo el tiempo de ejecución del algoritmo en cada caso, además de una comparación de los errores absoluto y relativo entre las condiciones iniciales deseadas u_j^{Train} y las condiciones óptimas obtenidas u_j para cada $j = 1, \ldots, N$. Los resultados obtenidos se presentan en el Cuadro 8, donde:

$$\text{error abs.} = \frac{\sum_{j=1}^{N} \| u_j^{Train} - u_j \|_2}{N}$$

y

$$\text{error rel.} = \frac{\sum_{j=1}^{N} \| u_j^{Train} - u_j \|_2}{\sum_{j=1}^{N} \| u_j^{Train} \|_2}.$$

Caso	γ	β	tiempo (s)	error abs.	error rel.
Caso I	10^{-2}	10^{-3}	646,89	2,48	0,30
	10		351,34	3,99	0,49
Caso II	10^{-2}	10^{-3}	452,85	2,33	0,28
	10		290,39	3,38	0,41

Cuadro 8: Training Set. Comparación *Caso I* y *Caso II*

De la realización de este experimento podemos concluir que aunque con utilización de la función de penalización para inducir dispersión no se obtenga un vector de localización óptima con una mayor cantidad de entradas nulas, el tiempo de ejecución en el segundo caso es menor y la reconstrucción de la condición inicial es más precisa, lo cual se deduce de los valores de los errores absoluto y relativo obtenidos.

Segundo experimento - Training Set

Al igual que el experimento anterior, aquí se busca observar la variación de la estructura de localizaciones óptimas cuando se trabaja con distintos valores de γ y β. Además de comparar los resultados obtenidos cuando se trabaja con y sin función de penalización para inducir dispersión en el vector de localizaciones resultante cuando se considera el subconjunto reducido de posibles localizaciones, seis puntos, dados por (3.26) como posibles ubicaciones. El Cuadro 9 muestra los resultados obtenidos con diferentes valores de γ y β para los casos estudiados. Debido al reducido número de puntos con los que se trabaja, la reconstrucción de la condición inicial será menos exacta. Los dos primeros bloques de la tabla muestran los resultados obtenidos sin considerar la función de penalización *(Caso I)* y el segundo bloque muestra los resultados en los cuales si considera dicha función fijando $\epsilon = \frac{1}{8}$ *(Caso II)*.

	γ	β	**0's**	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	**1's**	$\mathbf{J_0}$	$\mathbf{J_{end}}$	**iter**	$\parallel \mathbf{w} \parallel$
Caso I	0,1	10^{-4}	2	0	4	0	0	0,656	0,194	13	0,97
Caso I	0,1	10^{-3}	0	0	0	0	6	0,659	0,659	2	6
	1		0	0	0	0	6	6,059	6,059	2	6
	10		0	0	0	0	6	60,06	60,06	2	6
	100		3	0	0	0	3	600,06	300,12	5	3
	γ	β	**0's**	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	**1's**	$\mathbf{J_0}$	$\mathbf{J_{end}}$	**iter**	$\parallel \mathbf{w} \parallel$
Caso II	10^{-2}	10^{-4}	1	0	0	0	5	0,116	0,108	9	5
	0,1		0	0	1	0	5	0,656	0,649	7	5,13
	1		0	0	1	0	5	6,056	5,943	7	5,13
	10		0	0	1	0	5	60,06	58,88	7	5,13
Caso II	0,1	10^{-3}	0	0	0	0	6	0,659	0,659	2	6

Cuadro 9: Training set. Experimento 2 - Puntos dados en la malla.

Los resultados gráficos que se presentan a continuación, las Figuras (23) y (24) muestran el decrecimiento de la función objetivo para los casos I y II respectivamente, cuando se trabaja con distintos valores en los parámetros γ y β.

Las Figuras (25) y (26) muestran la estructura del vector de localización, que en promedio es óptimo para los pares de entrenamiento considerados en los casos I y II

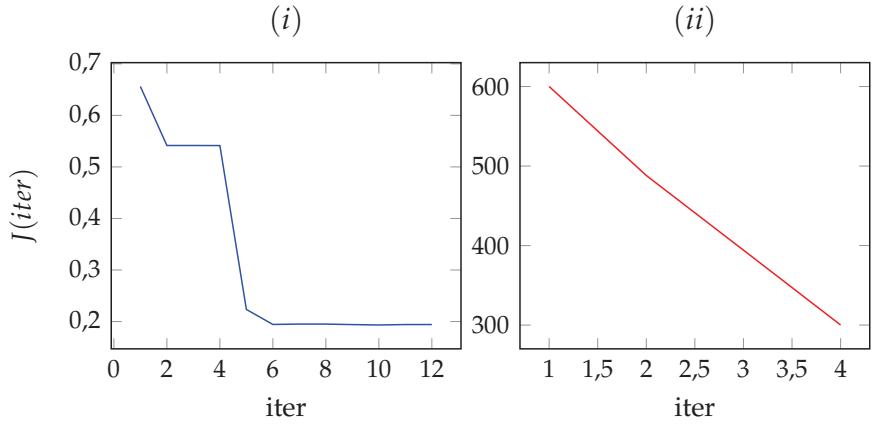

Figura 23: Decrecimiento de la función objetivo Training Set *Caso I* - $(i)\gamma = 0{,}1, \beta = 10^{-4}$, $(ii)\gamma = 100, \beta = 10^{-3}$.

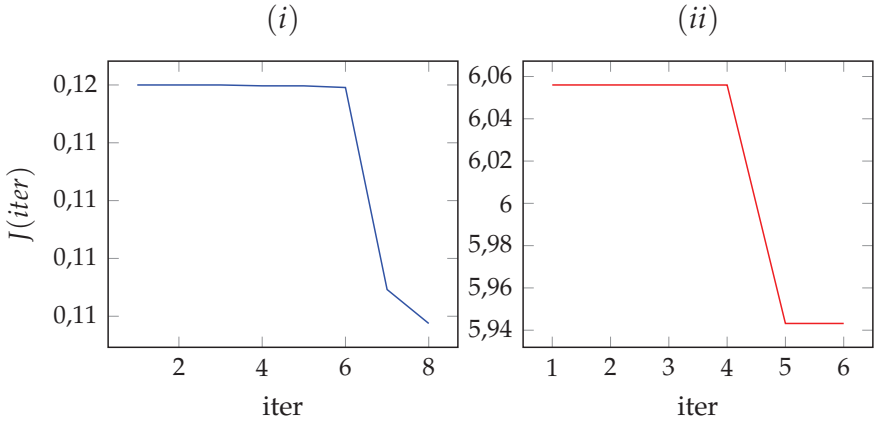

Figura 24: Decrecimiento de la función objetivo Training Set *Caso II* - Función de penalización. $(i)\gamma = 10^{-2}, \beta = 10^{-4}$, $(ii)\gamma = 1, \beta = 10^{-4}$.

respectivamente, cuando se trabaja con distintos valores en los parámetros γ y β.

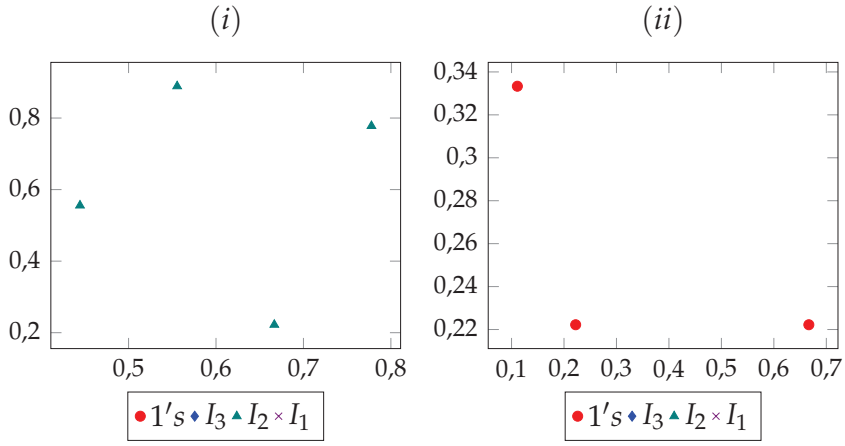

Figura 25: Estructura del vector de localizaciones Training Set *Caso I* - $(i)\gamma = 0{,}1, \beta = 10^{-4}$, $(ii)\gamma = 100, \beta = 10^{-3}$.

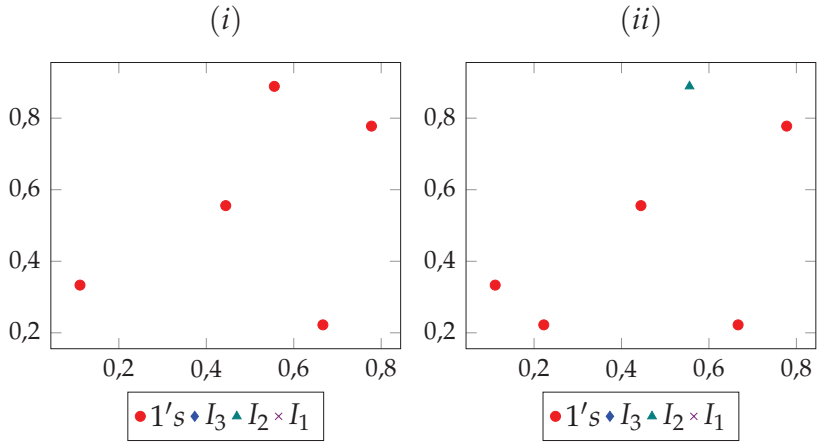

Figura 26: Estructura del vector de localizaciones Training Set *Caso II* - Función de penalización. $(i)\gamma = 10^{-2}, \beta = 10^{-4}$, $(ii)\gamma = 1, \beta = 10^{-4}$.

En el Cuadro 10 se presentan los resultados obtenidos al comparar los tiempos de ejecución y los errores absoluto y relativo entre las condiciones iniciales deseadas u_j^{Train} y las condiciones óptimas obtenidas u_j para cada $j = 1, \ldots, N$.

Caso	γ	β	tiempo (s)	error abs.	error rel.
Caso I	0,1	10^{-4}	259,11	5,66	0,69
	100	10^{-3}	28,35	6,09	0,75
Caso II	10^{-2}	10^{-4}	157,28	4,69	0,58
	1		157,03	4,76	0,58

Cuadro 10: Training Set. Comparación puntos dados en la malla *Caso I* y *Caso II*

En este caso debido a que la cantidad de ubicaciones que son tomadas como posibles es muy reducida se obtiene un valor más elevado del error tanto absoluto como relativo si lo comparamos con el experimento anterior. Así mismo debido a que los valores de los errores absoluto y relativo son menores para el caso II, se concluye que la reconstrucción de la condición inicial es más precisa en este caso.

En los dos experimentos de esta sección, aunque la utilización de la función de penalización no resulte en un vector de localizaciones óptimas con más entradas nulas, mediante su utilización se consigue una mejor reconstrucción de las condiciones iniciales deseadas, conclusión a la que se llega el comparar el valor de los errores obtenidos en cada caso.

Finalmente, para concluir esta sección se mostrará para cada $j = 1, \ldots, 10$, los pares de entrenamiento $\left(u_j^{Train}, y_j^{Train}\right)$ con los que se trabajó y los (u_j, y_j) obtenidos en el primer experimento, es decir, cuando se consideró todos los puntos en el espacio como posibles ubicaciones y se fijó $\gamma = 10$ y $\beta = 10^{-3}$. Con los parámetros escogidos, el vector de localizaciones óptimas tiene 15 entradas no nulas. La reconstrucción de los pares de entrenamiento se observa en las Figuras (27) - (36).

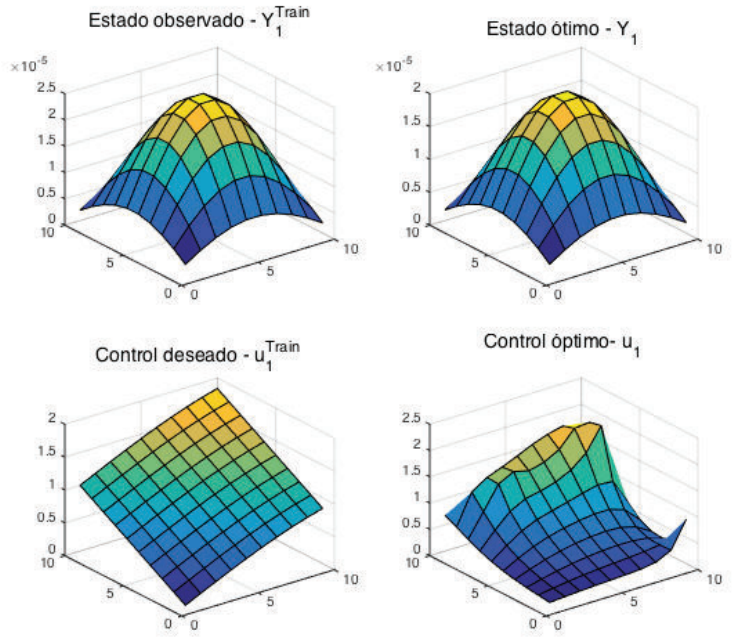

Figura 27: Reconstrucción pares de entrenamiento. $\left(u_1^{Train}, y_1^{Train}\right)$ - (u_1, y_1) .

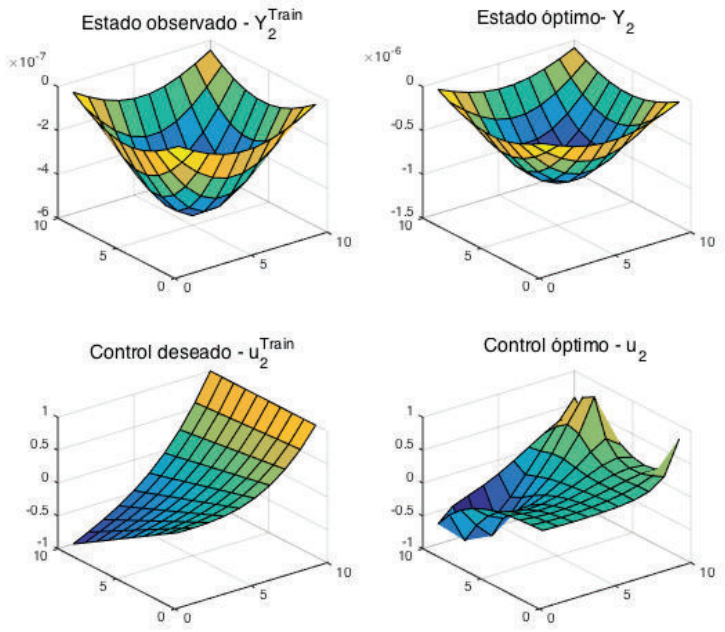

Figura 28: Reconstrucción pares de entrenamiento. $\left(u_2^{Train}, y_2^{Train}\right)$ - (u_2, y_2) .

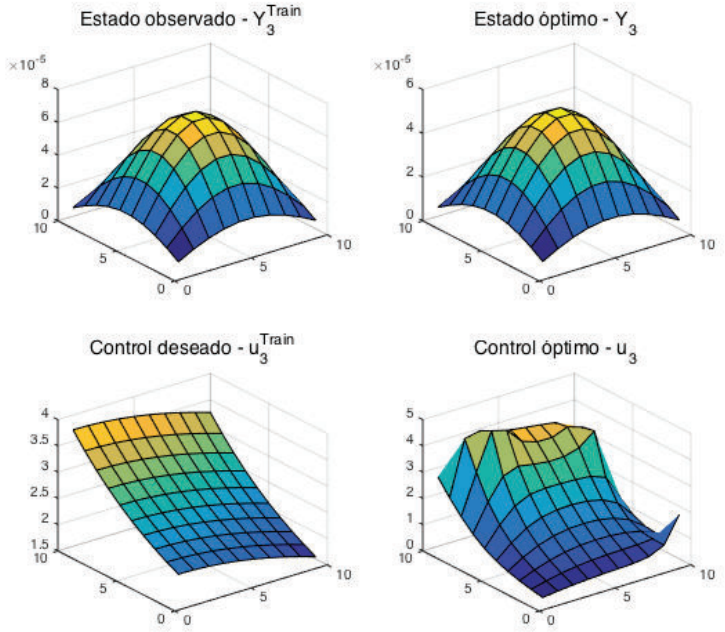

Figura 29: Reconstrucción pares de entrenamiento. $\left(u_3^{Train}, y_3^{Train}\right)$ - (u_3, y_3) .

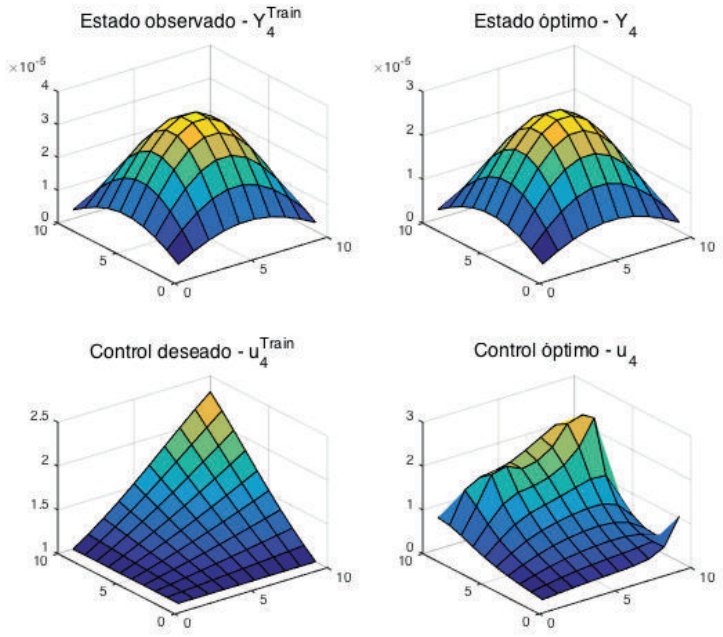

Figura 30: Reconstrucción pares de entrenamiento. $\left(u_4^{Train}, y_4^{Train}\right)$ - (u_4, y_4) .

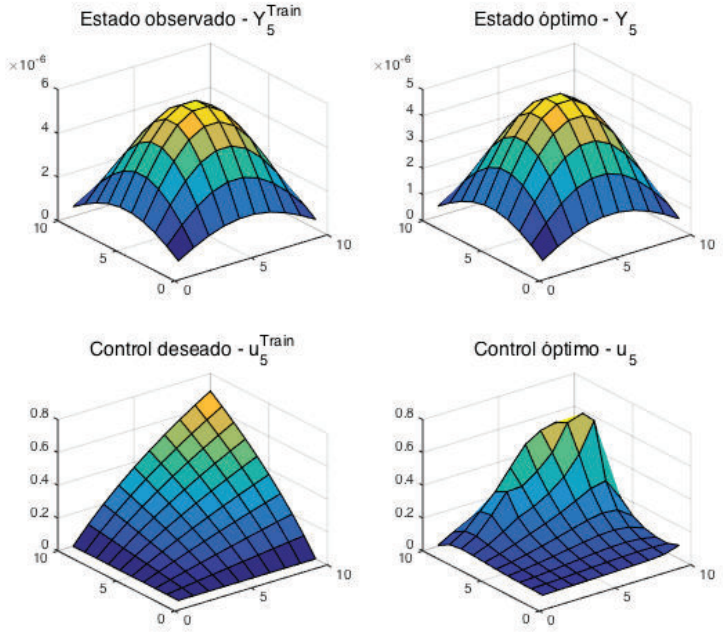

Figura 31: Reconstrucción pares de entrenamiento. $\left(u_5^{Train}, y_5^{Train}\right)$ - (u_5, y_5) .

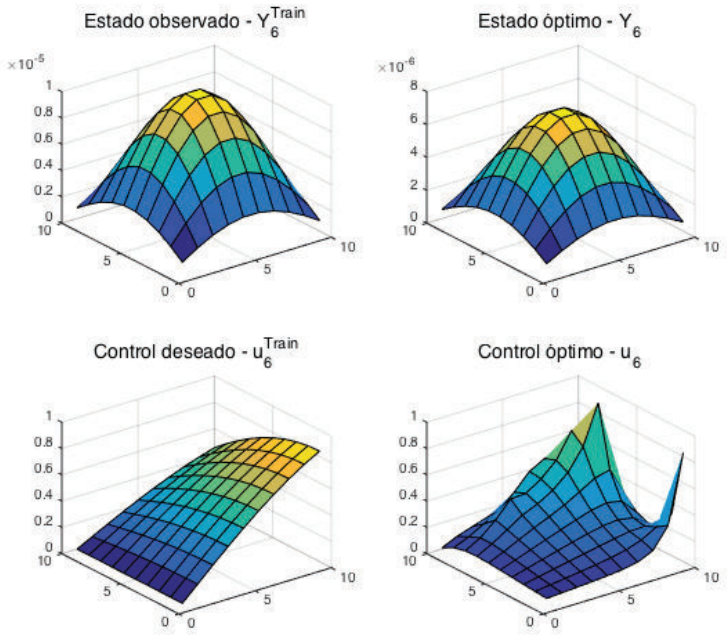

Figura 32: Reconstrucción pares de entrenamiento. $\left(u_6^{Train}, y_6^{Train}\right)$ - (u_6, y_6) .

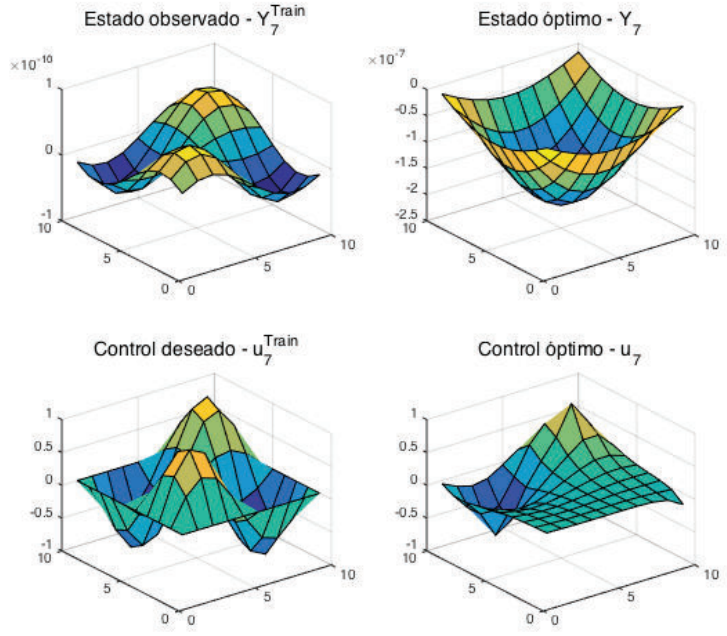

Figura 33: Reconstrucción pares de entrenamiento. $\left(u_7^{Train}, y_7^{Train}\right)$ - (u_7, y_7) .

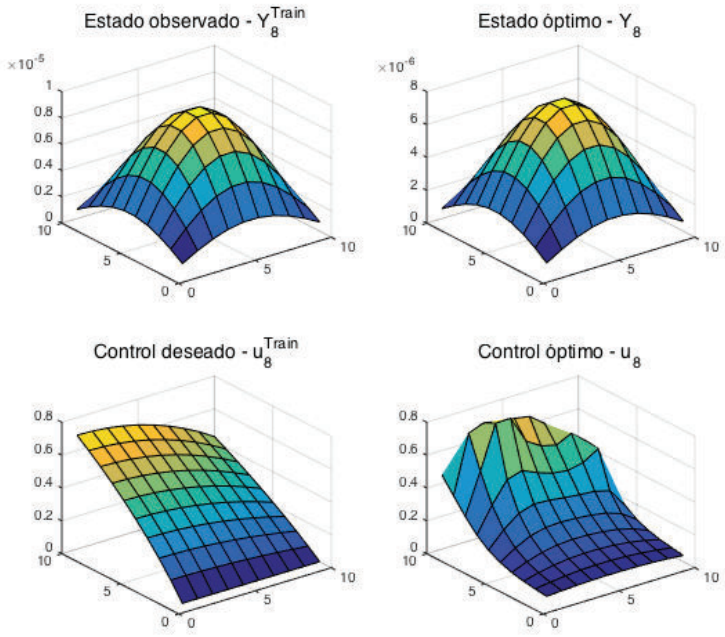

Figura 34: Reconstrucción pares de entrenamiento. $\left(u_8^{Train}, y_8^{Train}\right)$ - (u_8, y_8) .

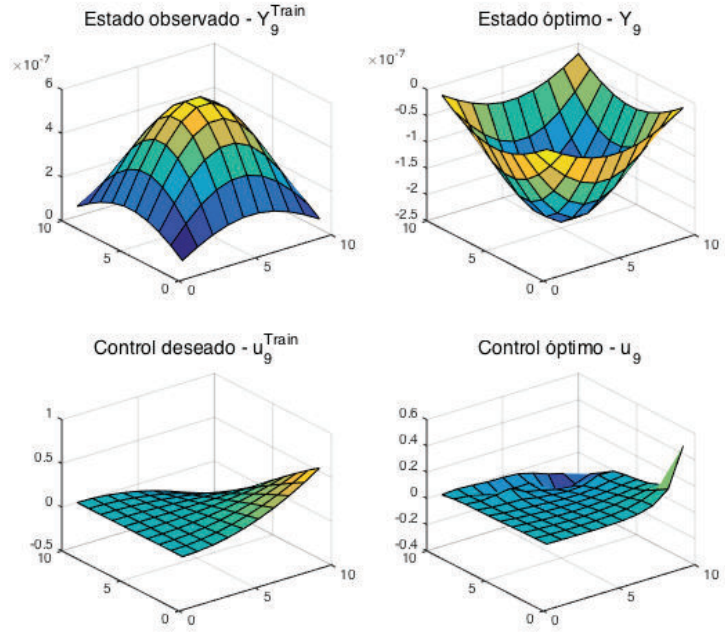

Figura 35: Reconstrucción pares de entrenamiento. $\left(u_9^{Train}, y_9^{Train}\right)$ - (u_9, y_9) .

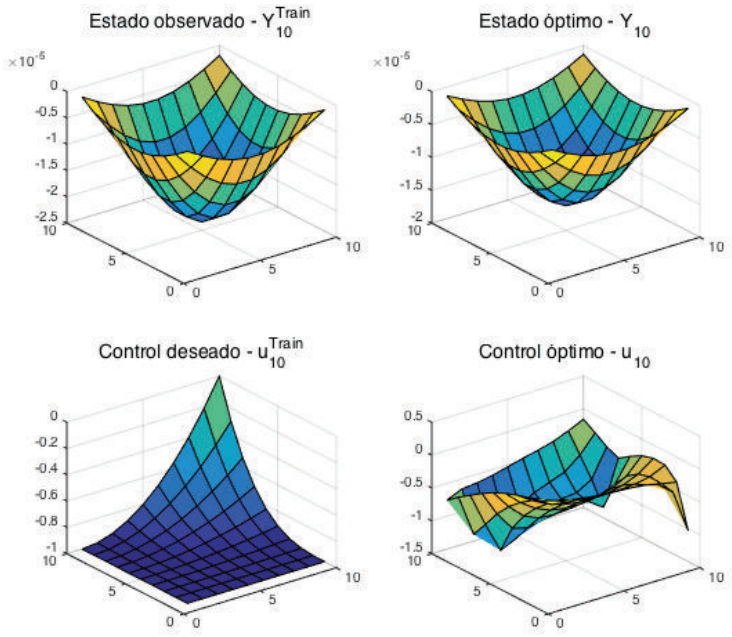

Figura 36: Reconstrucción pares de entrenamiento. $\left(u_{10}^{Train}, y_{10}^{Train}\right)$ - $\left(u_{10}, y_{10}\right)$.

Capítulo 4

Conclusiones

En el presente trabajo se ha tratado el problema de localización óptima de observaciones, de tal manera que se logre la reconstrucción de la condición inicial de un problema parabólico, como un problema de optimización continua a dos niveles. El enfoque utilizado fue el de los conjuntos de entrenamiento con la finalidad de obtener un vector de localizaciones óptimas robusto.

Para resolver el problema de la reconstrucción de la condición inicial del estado del sistema, es decir, para encontrar la solución del problema de asimilación de datos, que corresponde al problema interior del problema binivel, se utilizó el enfoque variacional, específicamente la técnica del $4D-$VAR que considera observaciones distribuidas en un intervalo de tiempo determinado. Mientras que la solución obtenida para el problema de localización, que corresponde al nivel superior del problema binivel, es un vector de localizaciones óptimas de tal manera que con las ubicaciones obtenidas se consiga una mejor reconstrucción de la condición inicial deseada y del estado del sistema observado, respectivamente.

Considerando que el número de observaciones que se pueden obtener es limitado, se incluyó en el planteamiento del problema al parámetro de penalización $\gamma > 0$, que actúa sobre w de la siguiente forma: a medida que más observaciones son tomadas en cuenta el producto $\gamma \sum_k |w_k|$, presente en el funcional objetivo, será mayor. Al tratarse de un problema de minimización se favorecerá los casos en los cuales el vector w tenga más entradas nulas. Este parámetro será denominado como parámetro de penalización del vector w. Sin embargo, debido a la no diferenciabilidad de este funcional se resolvió una relajación de este problema, haciéndolo diferenciable pero aumentando una restricción tipo caja al sistema.

Sin embargo la solución de la relajación del modelo ya no es necesariamente un vector de entradas binarias y no establece claramente donde tomar o no una ubicación

si para esa entrada el vector de localizaciones resultante toma valores entre cero y uno. Para obtener un vector w con su mayoría de entradas binarias, aún cuando se resuelva el problema relajado se introdujo una función de penalización que induce a que el vector solución obtenido tenga un mayor número de entradas nulas.

La resolución numérica del problema también se la realizó a dos niveles. Debido a que todo el sistema de optimalidad del problema de asimilación es la restricción del problema en el nivel superior, de localización óptima, se utilizó un método con una rápida velocidad de convergencia para la resolución de este primer sistema, el BFGS. Para encontrar direcciones de descenso y resolver el problema de localización óptima se utilizaron las actualizaciones de la matriz inversa del BFGS proyectado conjuntamente con una regla de búsqueda lineal de Armijo modificada. Es importante notar que el criterio de parada que se utilizó en la implementación numérica considera pasos completos.

Se desarrolló un algoritmo que permitió obtener un vector de localizaciones, control y estado óptimos que aproximen a la condición inicial simulada y al estado del sistema observado. Como una primera aproximación se realizó la parte experimental considerando solo una condición inicial simulada y un estado observado, es decir, se trabajo únicamente con la dupla $\left(u_1^{Train}, y_1^{Train}\right)$.

Se realizaron varios experimentos numéricos, cuando se trabajó únicamente con un par de entrenamiento el algoritmo dio como resultado un vector de localizaciones óptimo, tal que la condición inicial simulada y el estado observado fueron reconstruidos, con la utilización de diferentes parámetros que intervienen en el modelo. Sin embargo, se observó que con ciertas elecciones de los parámetros el algoritmo converge casi sin realizar iteraciones y la estructura del vector óptimo w es similar al vector con el que se inicializó w_0. Esto se debe a que la tolerancia con la que se trabajó es muy poco restrictiva. Así también, cuando se trabajó con varios pares de entrenamiento se observó que existen parámetros con los cuales el funcional objetivo no disminuía su valor en cada iteración. Al tratarse de un método de descenso, la razón por la cual el funcional objetivo no decreciera es una mala elección del parámetro de búsqueda lineal. En efecto, al trabajar con pasos pequeños el funcional reducía su valor en cada iteración, sin embargo requiere un número excesivo de iteraciones. Además del parámetro de búsqueda lineal, la elección de un test de parada alternativo es importante para evitar este comportamiento en el algoritmo.

En uno de los experimentos implementados se consideró solo un subconjunto pequeño de locaciones como factibles. Se realizó esta distinción para acercarnos más a la realidad del problema, ya que en la práctica no siempre es posible tomar observaciones en todos los puntos del dominio debido a temas logísticos o de costos. Por tanto,

ya que se consideran menos puntos como factibles, el vector de localizaciones tiene muchas más entradas nulas, haciendo que la reconstrucción de la condición inicial sea menos exacta, pero a la vez menos costosa.

Es importante notar que al utilizar la función de penalización para inducir dispersión en el vector w, el funcional objetivo difiere en relación a cuando no se utiliza dicha función de penalización. Ya que los dos funcionales que se están tratando de minimizar tienen ciertas variaciones uno respecto del otro, mismos valores en los parámetros γ y β no nos conducirán a un único vector w óptimo para los dos casos.

Como se mencionó anteriormente, para obtener resultados de localizaciones más robustos se trabajó con conjuntos de entrenamiento, una técnica del Machine Learning que permite predecir información de un dato desconocido basándose en la información aprendida de una muestra de datos conocidos. En nuestro caso el conjunto de entrenamiento se formó por simulaciones de la condición inicial y observaciones del estado del sistema respectivamente, y lo que se obtuvo es un vector de localizaciones en promedio óptimo para todos los pares de entrenamiento considerados. El algoritmo utilizado para la resolución del problema de localización cuando se consideran varios pares de entrenamiento toma ventaja de la estructura del problema y resuelve cada sistema en forma paralela.

De los experimentos realizados pudimos observar que una potencial limitación del algoritmo implementado es que depende mucho de los parámetros con los cuales se trabaje. Así también, el algoritmo no proporciona un control directo sobre el número de observaciones que se obtienen, es decir para que el algoritmo dé como solución un número fijo pre establecido de observaciones o sensores a ser instalados se deben fijar los parámetros experimentalmente. En la práctica resolver el problema de localización óptima requerirá determinar de manera experimental los valores adecuados para los parámetros a utilizarse.

Bibliografía

[Alexanderian et al., 2014] 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. *SIAM Journal on Scientific Computing*, 36(5):A2122–A2148.

[Alexanderian et al., 2016] Alexanderian, A., Petra, N., Stadler, G., and Ghattas, O. (2016). A fast and scalable method for a-optimal design of experiments for infinite-dimensional bayesian nonlinear inverse problems. *SIAM Journal on Scientific Computing*, 38(1):A243–A272.

[Bauer, 2001] Bauer, H. (2001). *Measure and integration theory*, volume 26. Walter de Gruyter.

[Casas et al., 2013] Casas, E., Clason, C., and Kunisch, K. (2013). Parabolic control problems in measure spaces with sparse solutions. *SIAM Journal on Control and Optimization*, 51(1):28–63.

[Casas and Kunisch, 2016] Casas, E. and Kunisch, K. (2016). Parabolic control problems in space-time measure spaces. *ESAIM: Control, Optimisation and Calculus of Variations*, 22(2):355–370.

[De los Reyes, 2015] De los Reyes, J. (2015). *Numerical PDE-constrained optimization*. Springer.

[Evans, 1998] Evans, L. (1998). Partial differential equations (graduate studies in mathematics 19 american mathematical society).

[Geiger and Kanzow, 2013] Geiger, C. and Kanzow, C. (2013). *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. Springer-Verlag.

[Guan and Gray, 2013] Guan, W. and Gray, A. (2013). Sparse high-dimensional fractional-norm support vector machine via dc programming. *Computational Statistics & Data Analysis*, 67:136–148.

[Kalnay, 2003] Kalnay, E. (2003). *Atmospheric modeling, data assimilation and predictability*. Cambridge university press.

[Kang and Xu, 2012] Kang, W. and Xu, L. (2012). Optimal placement of mobile sensors for data assimilations. *Tellus A*, 64.

[Kelley, 1999] Kelley, C. (1999). *Iterative methods for optimization*, volume 18. Siam.

[Krause and Guestrin, 2007] Krause, A. and Guestrin, C. (2007). Near-optimal observation selection using submodular functions. In *AAAI*, volume 7, pages 1650–1654.

[Kubrusly, 2015] Kubrusly, C. S. (2015). *Essentials of Measure Theory*. Springer.

[Lions, 1971] Lions, J. (1971). *Optimal control of systems governed by partial differential equations*, volume 170. Springer Verlag.

[Lions, 1992] Lions, J.-L. (1992). Pointwise control for distributed systems. *Control and Estimation in Distributed Parameter Systems, Frontiers in Applied Mathematics*, 11:1–39.

[Lions and Magenes, 1972] Lions, J. L. and Magenes, E. (1972). *Non-homogeneous boundary value problems and applications*, volume 1. Springer Science & Business Media.

[Petra and Stadler, 2011] Petra, N. and Stadler, G. (2011). Model variational inverse problems governed by partial differential equations. Technical report, DTIC Document.

[Quarteroni, 2010] Quarteroni, A. (2010). *Numerical models for differential problems*, volume 2. Springer Science & Business Media.

[Raschke and Jacob, 2013] Raschke, E. and Jacob, D. (2013). *Energy and water cycles in the climate system*, volume 5. Springer Science & Business Media.

[Raymond, 1997] Raymond, J. (1997). Nonlinear boundary control of semilinear parabolic problems with pointwise state constraints. *Discrete and Continuous Dynamical Systems*, 3:341–370.

[Roubíček, 2013] Roubíček, T. (2013). *Nonlinear partial differential equations with applications*, volume 153. Springer Science & Business Media.

[Tröltzsch, 2010] Tröltzsch, F. (2010). Optimal control of partial differential equations. *Graduate studies in mathematics*, 112.

[Warner, 2010] Warner, T. T. (2010). *Numerical weather and climate prediction*. Cambridge University Press.