ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

REGULARIZACIÓN DE VARIACIÓN TOTAL GENERALIZADA PARA EL PROBLEMA MAL CONDICIONADO DE ASIMILACIÓN DE DATOS

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

TESIS

KAREN ESTEFANÍA LOAYZA ROMERO
eloayza16@gmail.com

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

QUITO, MAYO 2017

DECLARACIÓN

Yo KAREN ESTEFANÍA LOAYZA ROMERO, 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.

_____

Karen Estefanía Loayza Romero

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por KAREN ESTEFANÍA LOAYZA ROMERO, bajo mi supervisión.

Dr. Juan Carlos De los Reyes Bueno
Director del Trabajo

AGRADECIMIENTOS

A mis padres y hermanos por su apoyo incondicional y por siempre haber creído en mí. Al Dr. Juan Carlos De los Reyes por su guía a lo largo de este trabajo y por darme la confianza para desarrollar esta investigación. A Paula Castro por todos sus consejos y recomendaciones.

Esta investigación fue financiada por los proyectos: *"Sistema de Pronóstico del Tiempo para todo el Territorio Ecuatoriano: Modelización Numérica y Asimilación de Datos."*, PIC-15-INAMHI-001, un convenio entre el Centro de Modelización Matemática (MODEMAT) y el Instituto Nacional de Meteorología e Hidrología (INAMHI); el proyecto *"Elaboración de un plan de información y contingencia ante eventos oceánicos extremos para la municipalidad de San Cristóbal"*, PIJ-15-03, de la Escuela Politécnica Nacional; y el proyecto *"Sparse Optimal Control of Differential Equations: Algorithms and Applications"* de la red Math-AmSud.

DEDICATORIA

A mi familia y amigos.

Índice general

Resumen

En este trabajo estudiamos la regularización de variación total (TV) y la regularización de variación total generalizada de segundo orden (TGV) aplicada al conocido problema de asimilación de datos variacional usando la ecuación de Burgers como ecuación de estado. Estas regularizaciones son heredadas de los problemas de restauración de imágenes en los cuales se ha mostrado experimentalmente que la regularización TGV conserva los puntos de discontinuidad de las funciones y las reconstruye de mejor manera comparado con la regularización TV, debido a que elimina el efecto de escalonamiento. El trabajo se concentra en la resolución del problema de asimilación de datos en dimensión finita, asumiendo que los datos de entrada del problema son vectores. Realizamos además la derivación de las condiciones de optimalidad de primer orden. Para la resolución numérica de la ecuación de Burgers se realizó una discretización semi–implícita para las variables temporales y para las variables espaciales se usó el esquema de diferencias finitas con *Upwinding* para la discretización de la primera derivada espacial. La solución del problema se realizó con métodos iterativos de optimización: el método del descenso más profundo, BFGS y un método de Newton globalizado. Mostraremos además, para cada método presentado en este trabajo, resultados teóricos que garantizan la convergencia a puntos estacionarios del problema. El capítulo de experimentos numéricos está dedicado a mostrar el desempeño de los algoritmos iterativos de optimización con respecto al número de iteraciones y la manera en la que las diferentes regularizaciones reconstruyen las soluciones. En particular mostraremos el efecto de escalonamiento producido por la regularización TV y la manera en la que la regularización TGV elimina dicho efecto.

Abstract

In this work we study the total variation (TV) regularization and the second order total generalized variation (TGV) regularization applied to the well–known data assimilation problem using the Burgers equation as the state equation. These regularizations are inherited from image restoration problems in which it has been experimentally shown that the TGV regularization preserve the sharp fronts and recovers the solutions better compared to the TV regularization, mainly because it eliminates the staircase effect. The paper focuses on solving the finite dimension data assimilation problem, assuming that the input data of the problem are vectors. We also perform the derivation of the first order optimality conditions. For the numerical solution of the Burgers equation, a semi-implicit time discretization was performed and for the spatial variables the finite differences scheme with *upwinding* was used for the first order spatial derivative. The solution of the problem was made using iterative optimization methods: the steepest descent method, BFGS and globalized Newton methods. We will also show, for each method presented in this work, the theoretical results that guarantee the convergence to stationary points of the problem. The numerical experiments chapter is devoted to show the performance of the iterative optimization algorithms with respect to the number of iterations and the way each regularization recover the solution. In particular, we show the staircase effect produced by the TV regularization and the way how the TGV regularization eliminates it.

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

El problema de asimilación de datos puede ser descrito como el proceso mediante el cual se desea encontrar una aproximación para la condición inicial de un sistema dinámico usando observaciones e información previa, a la cual denominaremos *background*. Este proceso es ampliamente utilizado en la predicción numérica del tiempo. La solución de este problema puede ser aproximada a través de varios enfoques: interpolación óptima, estimación estadística o un enfoque variacional.

Una de las maneras de estimar estadísticamente la solución del problema de asimilación de datos consiste en usar los conocidos filtros de Kalman. La idea principal de este método es asumir que el background y las observaciones siguen una distribución normal, donde sus medias son el valor actual y las mediciones, respectivamente. Las matrices de covarianza asociadas a estas distribuciones representan a la incertidumbre en la solución y el ruido en las mediciones, respectivamente. En este contexto, el estado más probable, aquel que queremos aproximar, es alcanzado cuando ambas probabilidades son verdaderas. Este proceso se reduce a la estimación de la media y la varianza de la función de probabilidad conjunta. Además de este método, tenemos el método de máxima verosimilitud y el método de estimación bayesiana los cuales serán explicados de manera más extensa en el siguiente capítulo.

Por otro lado tenemos el enfoque variacional, el que será motivo de estudio a lo largo de este trabajo. La principal propiedad de este enfoque es la formulación del problema inverso como un problema de optimización no lineal. Se diferencian dos tipos de problemas: el 3D-VAR el cual considera varias observaciones distribuidas en el dominio de estudio, tomadas en un instante de tiempo. Este problema puede ser

representado matemáticamente como:

$$\begin{aligned} \min_{\mathbf{y},u} \quad & J(\mathbf{y},u) = \frac{1}{2}(z-\mathcal{H}(\mathbf{y}))^T R^{-1}(z-\mathcal{H}(\mathbf{y})) + \frac{1}{2}(u-u^b)^T B^{-1}(u-u^b) \\ \text{sujeto a:} \quad & \\ & y_i = u, \quad i = 1, \\ & y_{i+1} = \mathcal{M}_{i+1,i}(y_i), \quad \forall i = \{1,\ldots,N\}, \end{aligned} \tag{1.1}$$

donde u^b es la información previa o *background* que conocemos sobre la atmósfera, B es la matriz de covarianza de los errores del *background*. El vector z contiene información de las observaciones meteorológicas tomadas en ciertas ubicaciones del dominio para un instante de tiempo determinado. Ademas, tenemos la matriz de covarianza de los errores en las observaciones R y el operador de observaciones $\mathcal{H}(\cdot)$, el cual extrae los valores de la variable de estado de la malla espacial que coinciden con las observaciones y $\mathbf{y}$ es el arreglo que contiene todos los instantes y_i. En el caso del problema 3D–VAR, este operador además extrae el valor de la variable de estado en el instante que coincide con el de las observaciones. Finalmente, tenemos el operador no lineal control–estado $\mathcal{M}(\cdot)$. En el caso de la asimilación de datos meteorológicos, este operador consiste en la resolución del sistema de ecuaciones diferenciales parciales que describen la atmósfera, donde el índice i indica la evolución en el tiempo de la variable de estado.

El segundo problema considera observaciones distribuidas en el dominio de estudio tomadas en varios instantes sobre una ventana de tiempo. Este problema es denominado 4D–VAR. Matemáticamente está dado por la siguiente expresión:

$$\begin{aligned} \min_{\mathbf{y},u} \quad & J(\mathbf{y},u) = \frac{1}{2}\sum_{i=1}^{N}(z_i-\mathcal{H}_i(y_i))^T R_i^{-1}(z_i-\mathcal{H}_i(y_i)) + \frac{1}{2}(u-u^b)^T B^{-1}(u-u^b) \\ \text{sujeto a:} \quad & \\ & y_i = u, \quad i = 1, \\ & y_{i+1} = \mathcal{M}_{i+1,i}(y_i), \quad \forall i = \{1,\ldots,N\}. \end{aligned} \tag{1.2}$$

donde la principal diferencia con el problema anterior radica en la cantidad de observaciones que poseemos. En este caso contamos con N vectores de observaciones z_i ubicadas sobre el dominio de estudio, los cuales están distribuidos en una ventana de tiempo, donde cada índice i corresponde a un instante de tiempo determinado. Las matrices R_i son aquellas que contiene la información de la covarianza de los errores en las observaciones para cada instante de tiempo considerado. El operador $\mathcal{H}_i$ es el operador de observaciones para un tiempo determinado y extrae la información de la variable de estado en un instante i determinado para poder compararla con las observaciones. Como se mencionó anteriormente el índice i en la ecuación de estado indica la evolución en el tiempo de la variable de estado. El resto de parámetros no mencionados coinciden con los descritos en el problema 3D–VAR.

El problema de asimilación de datos es un problema inverso mal condicionado, ya que se busca la condición inicial de una ecuación diferencial parcial con información incompleta.

La gran mayoría de problemas inversos mal condicionados como por ejemplo en la restauración de imágenes pueden ser regularizados para obtener una mejor reconstrucción de la solución utilizando información que se conoce *a priori* sobre la solución. En el caso de que se conozca de antemano que las soluciones a un problema determinado son discontinuas se proponen funciones regularizadoras que conserven dichas características. Por ejemplo, en el procesamiento de imágenes en donde se espera que las soluciones sean constantes o lineales a trozos se utilizan regularizaciones como la de variación total o variación total generalizada [Bredies and Valkonen, 2011, Bredies et al., 2010, Knoll et al., 2011, Bredies et al., 2013].

En el problema de asimilación de datos el tipo de soluciones que esperamos obtener son constantes a trozos, lineales a trozos o en general discontinuas y por esta razón el uso de la regularización de variación total generalizada nos podría garantizar mejores resultados que los obtenidos con la regularización de variación total. Investigaciones previas proponen el uso de la regularización de variación total en el problema de asimiliación de datos [Freitag et al., 2010] con la finalidad de conservar los puntos de discontinuidad denominados en este contexto *sharp fronts*.

Como se mencionó anteriormente, en la asimilación de datos meteorológicos las restricciones del problema de optimización están dadas por el sistema de ecuaciones que describen la atmósfera. El costo computacional de la solución del sistema de ecuaciones diferenciales parciales que gobiernan la atmósfera es muy alto, por lo tanto, para simplificar este problema se propone usar una ecuación de Burgers como ecuación de estado. Esta ecuación fue desarrollada en 1939 por J.M. Burgers y es una simplificación de la ecuación de Navier–Stokes que modela la turbulencia de un fluido con un número de Reynolds alto. La ecuación de Burgers con viscosidad está dada por la siguiente expresión

$$\begin{aligned} \frac{\partial y(x,t)}{\partial t} - \nu \frac{\partial^2 y(x,t)}{\partial x^2} + \frac{\partial y(x,t)}{\partial x} \cdot y(x,t) &= f(x,t) && \text{en } Q = \Omega \times [0,T], \\ y(x,t) &= 0 && \text{sobre } \Gamma \times [0,T], \\ y(x,0) &= u(x) && \text{en } \Omega. \end{aligned} \tag{1.3}$$

donde Ω es el dominio de estudio el cual va a ser fijado en cada experimento y ν es el coeficiente de viscosidad del fluido. La ecuación de Burgers sin viscosidad se obtiene al fijar $\nu = 0$ en la ecuación anterior.

En este trabajo nos concentraremos en la resolución del problema 4D-VAR con la ecuación de Burgers como ecuación de estado. Matemáticamente, el problema a resolver es el siguiente.

$$\begin{aligned} \min_{(y,u)\in\mathbb{R}^m\times\mathbb{R}^n} \quad & \frac{1}{2}\sum_{i=1}^{N}(z_i - \mathcal{H}_i(y_i))^T R_i^{-1}(z_i - \mathcal{H}_i(y_i)) + \frac{1}{2}(u-u^b)^T B^{-1}(u-u^b) \\ \text{sujeto a:} \quad & \\ & \text{la ecuación (1.3) discretizada.} \end{aligned} \tag{1.4}$$

donde $m = nN_t$ con N_t el número de puntos en la discretización temporal y n número de puntos en la discretización espacial. Cabe recalcar que la función objetivo está dada en espacios de dimensión finita, es por eso que el problema entero va a estar también en $(\mathbb{R}^{n \cdot N_t} \times \mathbb{R}^n)$ y por tanto la ecuación de estado necesita ser discretizada. Se tratará sobre este tema en los siguientes capítulos.

El problema anterior es un problema inverso bien condicionado teniendo en cuenta que la inclusión del término correspondiente a la información previa (*background*) hace las veces de la regularización de Tikhonov, lo que nos garantiza la existencia de soluciones. En el caso de no tener este término el problema sería un problema inverso mal condicionado. A continuación, se presentarán los conceptos más importantes sobre este tema.

1. Problemas Inversos

Los problemas inversos han sido de gran importancia no solo desde el punto de vista matemático sino también en las aplicaciones. En un inicio este tipo de problemas fueron concebidos en la geofísica, sin embargo, a lo largo del tiempo se han descubierto aplicaciones que van desde la estimación de parámetros en la vulcanología hasta el tratamiento de imágenes médicas.

J.B. Keller en [Keller, 1976] propuso que dos problemas pueden ser llamados inversos, si la formulación del primer problema involucra al otro. En problemas de la vida real, se puede diferenciar de mejor manera este hecho. Por ejemplo, el *problema directo* trata de predecir el comportamiento futuro de un sistema físico a partir del conocimiento del estado inicial y de las leyes físicas (incluyendo los parámetros) que lo describen. El *problema inverso*, por otro lado, trata de determinar el estado presente o las leyes que lo describen usando observaciones relacionadas indirectamente con el sistema.

Existen al menos dos posibles motivaciones para el uso de este tipo de problemas. La primera, es la necesidad de conocer estados pasados o los parámetros que describen las leyes físicas. La segunda, es la necesidad de encontrar la manera de influenciar un sistema a través de su estado inicial o de sus parámetros con la finalidad de obtener observaciones futuras lo más acercadas a un estado deseado. En nuestro caso, es justamente la segunda premisa la que nos motiva a trabajar en este problema y buscar maneras de mejorarlo.

2. Problemas inversos mal condicionados

Un problema inverso puede ser formulado matemáticamente a través de la siguiente expresión: Encontrar u tal que para cualquier v dado

$$K(u) = v. \tag{1.5}$$

con $K : \mathcal{H}_1 \to \mathcal{H}_2$ y $\mathcal{H}_i$ dos espacios de Hilbert. Entonces se tiene la siguiente definición.

DEFINICIÓN 1.1. *El operador K es bien condicionado si*

- *para cualquier $v \in \mathcal{H}_2$ existe $u \in \mathcal{H}_1$, llamado la solución, para el cual se cumple* (1.5)
- *la solución u es única; y*
- *la solución es estable con respecto a las perturbaciones de v. Es decir, si $K(\tilde{u}) = \tilde{v}$ y $K(u) = v$, entonces $\tilde{u} \to u$ siempre que $\tilde{v} \to v$.*

Obviamente, esta no es una definición matemática propiamente dicha; para precisar este hecho es necesario fijar un problema en concreto y trabajar sobre este.

Otra manera de entender cuando un problema es mal condicionado depende de la relación que existe entre los datos y las soluciones. Es decir, los problemas inversos mal condicionados son aquellos que con pequeños errores en los datos producen soluciones con errores muy grandes.

3. Regularización de problemas inversos mal condicionados

Los problemas inversos mal condicionados pueden ser reformulados de tal manera que se pueda garantizar existencia y unicidad en las soluciones y la dependencia continua en los datos. Este proceso es conocido como regularización y existen varios enfoques. Entre los principales tenemos las regularizaciones por filtrado, regularización aposteriori, regularización variacional y regularización iterativa. En este trabajo nos vamos a concentrar en la regularización variacional de problemas inversos, en particular en el problema de asimilación de datos. La idea básica de este enfoque es resolver un problema de optimización cuya función objetivo contiene dos términos: el primero mide el ajuste de las observaciones y la variable de estado. El segundo es un término que nos garantiza que las soluciones sean lo más acercadas a la realidad. Matemáticamente, este problema tiene la siguiente estructura

$$\min_{(y,u)} J(y,u) = \varphi(y, z_d) + \mathcal{R}(u),$$

donde $\varphi(y, z_d)$ es el término de fidelidad, aquel que mide el ajuste de las observaciones y la variable de estado. En el caso de los problemas 3D-VAR y 4D-VAR mencionados anteriormente, el término de fidelidad está dado por la expresión:

$$\varphi(z, u^b, y, u) = d\frac{1}{2}\sum_{i=1}^{N}(z_i - \mathcal{H}_i(y_i))^T R_i^{-1}(z_i - \mathcal{H}_i(y_i)) + \frac{1}{2}(u - u^b)^T B^{-1}(u - u^b).$$

El término $\mathcal{R}$ es el término de regularización el cuál consiste en utilizar la información que se conoce previamente de la solución. Es decir, si se conoce *a priori* que las soluciones a un cierto

problema son discontinuas y constantes a trozos, conviene utilizar un término regularizador que garantice que efectivamente las soluciones del problema mantengan dichas características.

Particularmente, en el campo de la restauración de imágenes, debido a que en este caso esperamos que las soluciones (imágenes) son constantes a trozos, un término de regularización ampliamente utilizado es la regularización de variación total. En la actualidad, se están proponiendo nuevos términos de regularización los cuales garanticen conservar las estructuras de funciones lineales a trozos, en particular estamos hablando del término de regularización de variación total generalizada de segundo orden.

A continuación se describen las funciones de regularización mencionadas anteriormente para espacios de dimensión finita.

Regularización de Variación Total (TV) La regularización de variación total está dada por la siguiente expresión:

$$\mathcal{R}(u) = \beta \sum_{i=1}^{n} |D_i u|,$$

con $\beta > 0$ y D es la matriz que corresponde a la discretización del gradiente y se nota por D_i su i–ésima fila.

Regularización de Variación Total Generalizada de Segundo Orden (TGV) La regularización de variación total generalizada tiene asociada dos parámetros $\alpha, \beta > 0$ y está dada por la siguiente expresión

$$\mathcal{R}(u) = \min_{w} \left\{ \alpha \sum_{i=1}^{n} |D_i u - w| + \beta \sum_{i=1}^{n} |E_i w| \right\},$$

donde D es la matriz que representa la discretización del gradiente y la matriz E es el gradiente simetrizado y se nota por E_i su i–ésima fila. De manera general, el gradiente simetrizado está dado por la siguiente expresión:

$$\mathcal{E}w = \frac{\nabla w + \nabla^T w}{2}.$$

Cabe recalcar que esta expresión está dada en términos funcionales. Por ejemplo, para una función w tal que $w(x, y) = (w_1(x, y), w_2(x, y))$ el operador $\mathcal{E}$ está dado por:

$$\mathcal{E}w = \begin{bmatrix} \partial_x w_1 & \dfrac{\partial_x w_2 + \partial_y w_1}{2} \\ \dfrac{\partial_x w_2 + \partial_y w_1}{2} & \partial_y w_2 \end{bmatrix}$$

Sin embargo, en nuestro caso, puesto que estamos trabajando con funciones con una sola dimensión espacial, el operador $\mathcal{E}$ coincide con ∇ y por tanto la matriz E corresponde también a la discretización del gradiente.

4. Problema de asimilación de datos con términos de regularización

El objetivo principal de la asimilación de datos es recuperar la condición inicial de la atmósfera a través de la combinación de información proveniente de observaciones meteorológicas y de pronósticos previos. De antemano conocemos que el estado de la atmósfera no tiene porque ser una función continua, es más, se prevé que está función tenga puntos de discontinuidad y de no diferenciabilidad. Con esto en mente, debemos garantizar que nuestras soluciones sean discontinuas, y para esto usamos los términos de regularización descritos en la sección precedente.

Como se mencionó anteriormente, el término correspondiente al background u^b hace las veces de regularización de Tikhonov lo cual permite garantizar que el problema dado en (1.6) es un problema bien condicionado. Sin embargo, la necesidad de utilizar regularizaciones más generales radica en que el término de Tikhonov no nos garantiza recuperar los puntos de discontinuidad (*sharp fronts*) de las funciones y por tanto la aproximación a la solución exacta será deficiente.

Todos los elementos mencionados anteriormente nos permiten formular los problemas en los que nos vamos a concentrar en este trabajo, los cuales tendrán la siguiente estructura:

$$\begin{aligned} &\min_{(y,u)} && \frac{1}{2}\sum_{i=1}^{N}(z_i - \mathcal{H}_i(y_i))^T R_i^{-1}(z_i - \mathcal{H}_i(y_i)) + \frac{1}{2}(u-u^b)B^{-1}(u-u^b) + \mathcal{R}(u) \\ &\text{sujeto a:} \\ &&& \text{Ecuación (1.3) discretizada.} \end{aligned} \tag{1.6}$$

El documento está organizado de la siguiente manera: en el capítulo 1 se presenta el problema de asimilación de datos propiamente dicho y la simplificación que vamos a considerar en este trabajo usando la ecuación de Burgers. El capítulo 2 está dedicado a la estimación estadística de la solución del problema de asimilación de datos y a mostrar la equivalencia que existe entre los métodos estadísticos y el enfoque variacional. El capítulo 3 está dedicado a la presentación de las condiciones de optimalidad de primer orden que nos permitirán presentar de manera formal el sistema de optimalidad de los problemas. Para dicho efecto, discutimos en primer lugar los esquemas de discretización para ecuaciones diferenciales parciales con términos de transporte, y presentamos un esquema completamente discretizado de la ecuación de estado. Por otro lado, teniendo en cuenta que las condiciones de optimalidad de los problemas de optimización no lineal dependen de las derivadas de primer orden es necesario incluir términos de regularización de las funciones objetivo para de este modo garantizar su diferenciabilidad. Al final del capítulo 3 se presenta la demostración de la convergencia de la soluciones de los problemas regularizados a los problemas originales. El capítulo 4 se concentra en proponer algoritmos iterativos de optimización no lineal que nos permita resolver estos

problemas. Específicamente vamos a utilizar el método del descenso más profundo, el método BFGS y un método de Newton globalizado. Además, para cada método mencionado en este trabajo, presentamos resultados teóricos para garantizar la convergencia a puntos estacionarios del problema. El capítulo 5 está dedicado a la presentación de los experimentos numéricos cuyo principal objetivo es mostrar que las soluciones del problema con la regularización de variación total generalizada son más acercadas a las soluciones exactas, en especial cuando se trata de funciones discontinuas y lineales a trozos. Presentamos varios experimentos cambiando el tamaño de las mallas, los cuales nos permiten mostrar la influencia que tienen los tamaños de la mallas en la recuperación de los puntos de discontinuidad de las funciones. Además, se diseñó un experimento en el cual se busca analizar la influencia que tiene la cantidad de observaciones utilizadas sobre la manera en la que se reconstruyen las soluciones. Finalmente, mostramos los resultados obtenidos con la ecuación de Burgers sin viscosidad, demostrando de esta manera que trabajar con un esquema de *discretizar–luego–optimizar* nos permite resolver problemas que en espacios funcionales no necesariamente tienen solución.

Capítulo 2

Enfoque estadístico del problema de asimilación de datos

El problema de asimilación de datos consiste en encontrar el estado más probable de la atmósfera utilizando información de observaciones y de pronósticos pasados. Recordamos el problema de asimilación de datos 4D-VAR mencionado en el capítulo anterior:

$$\begin{aligned}
&\min_{y,u} \quad J(y,u) = \frac{1}{2}\sum_{i=1}^{N}(z_i - \mathcal{H}_i(y_i))^T R_i^{-1}(z_i - \mathcal{H}_i(y_i)) + \frac{1}{2}(u-u^b)^T B^{-1}(u-u^b)\\
&\text{sujeto a:}\\
&\qquad y_i = u, \quad i = 1,\\
&\qquad y_{i+1} = \mathcal{M}_{i+1,i}(y_i), \quad \forall i = \{1,\dots,N\}.
\end{aligned}$$

Sabemos que efectivamente la función objetivo de este problema busca satisfacer las propiedades mencionadas anteriormente. Además, esta función objetivo puede ser interpretada desde un punto de vista estadístico usando el método de máxima verosimilitud o un enfoque bayesiano. Este capítulo está dedicado a mostrar las hipótesis bajo las cuales la función objetivo descrita anteriormente es equivalente a los métodos de estimación de parámetros estadísticos.

De manera general, el problema de asimilación de datos puede ser representado como: encontrar u tal que

$$\begin{aligned}
z_i &= \mathcal{M}_i(u) + v_i, \quad \forall i = 1,\dots,N\\
u^b &= u + v,
\end{aligned}$$

donde z_i son las observaciones en cada tiempo, u^b es la información previa, v_i y v son los errores sobre los cuales asumiremos las siguientes características:

- v_i, v siguen una distribución normal de media 0
- La matriz de covarianza de v_i es R_i, para $i = 1,\dots,N$.

- La matriz de covarianza de v es B.
- Los errores v_i y v no está correlacionados entre sí, ni con la variable u.

Dentro del campo de la estimación de parámetros estadísticos existen dos enfoques diferentes. El primero propuesto por Fisher en 1920 en el cual asumimos que la incognita u es una variable determinista; de esta suposición se deriva el método de máxima verosimilitud. El segundo enfoque, conocido como bayesiano, consiste en considerar que la incognita u es una variable aleatoria de la cual se conoce la función de probabilidad *a priori*. Además, usando la fórmula de Bayes, se puede obtener la distribución de probabilidad a posteriori mediante la cual podremos obtener estimadores como el de máxima probabilidad a posteriori (MAP) para la resolución del problema. En las siguientes secciones se desarrollará a más profundidad estos temas. Las ideas desarrolladas en este capítulo tomarón como referencia el libro [Lewis et al., 2006].

1. Método de máxima verosimilitud

El método de máxima verosimilitud fue desarrollado por Fisher, y se basa principalmente en la suposición de que la incógnita u es determinista y que conocemos de antemano las funciones de probabilidad condicional $p(z|u)$ y $p(u^b|u)$. Además, asumimos que las observaciones y la información previa son independientes entre sí. Así,

$$p(z, u^b|u) = p(z|u)p(u^b|u).$$

La teoría de probabilidad básica, nos indica que si consideramos la función de probabilidad condicional como una función de u obtenemos la conocida función de verosimilitud $L(u|z, u^b)$, la cual está dada por la siguiente expresión

$$L(u|z, u^b) = p(z|u)p(u^b|u),$$

y depende directamente de las funciones de distribución condicional que asumimos para z y para u^b. La idea básica del método es: dado un conjunto de observaciones z, u^b buscamos el valor de u que maximice la probabilidad o verosimilitud de obtener dicha muestra. En el caso de la asimilación de datos buscamos el estado de la atmósfera más probable dado que conocemos las observaciones y la información de pronósticos previos.

En las siguientes secciones se explicará como aplicar este método para los problemas 3D-VAR y 4D-VAR.

2. Método de estimación Bayesiana

A diferencia del método de máxima verosimilitud, en este caso consideramos que la incógnita u es una variable aleatoria de la cual se conoce *a priori* su distribución de probabilidad $p(u)$. Esta distribución debe resumir los conocimientos previos que tenemos sobre la variable u. Además, asumimos que las observaciones z contienen información sobre u la cual será mostrada en términos de la distribución condicional $p(z|u)$. La idea básica de este método es combinar esta información a través el uso de la fórmula de Bayes, para encontrar un estimador al cual denominaremos *estimador de Bayes*. Sin embargo, con esta técnica no podemos garantizar la unicidad del estimador de Bayes, por lo tanto se buscará el "mejor" estimador de Bayes, el cual será escogido de tal manera que el error en la estimación sea lo más pequeño posible; es decir, si definimos el error de la siguiente manera:

$$\tilde{u} = u - \hat{u},$$

donde $\hat{u}$ es el estimador dadas las observaciones z y una función de costo $c(.)$, buscamos minimizar el costo esperado dado por $E[c(u - \hat{u})]$.

La función de costo $c : \mathbb{R}^n \to \mathbb{R}$, debe satisfacer las siguientes condiciones:

1. $c(0) = 0$,
2. $c(\cdot)$ es una función no decreciente con respecto a la norma del argumento. Es decir, para cualquier par de vectores $a, b \in \mathbb{R}^n$ se tiene que

$$c(a) \leq c(b) \quad \text{si} \quad ||a|| \leq ||b||.$$

A continuación mostramos algunas de las funciones que se pueden utilizar.

Suma ponderada del error al cuadrado : Sea $W \in \mathbb{R}^{n\times n}$ una matriz simétrica y definida positiva. Entonces

$$c(u - \hat{u}) = (u - \hat{u})^T W (u - \hat{u}) = \| u - \hat{u} \|_W^2 .$$

Función de costo uniforme : Sea $\varepsilon > 0$ un número real fijo suficiente pequeño. Se define:

$$c(u - \hat{u}) = \begin{cases} 0, & \text{si } \| u \|_\infty \leq \varepsilon, \\ 1, & \text{caso contrario.} \end{cases}$$

Error absoluto : Esta función se utiliza en el caso específico en que la incógnita es un escalar y está dada por:

$$c(u - \hat{u}) = |u - \hat{u}|.$$

Función de costo simétrica y convexa : De manera general, cualquier función convexa y si-

métrica podría servir como función de costo.

El problema de estimación bayesiana se formula de la siguiente manera:

Dados $p(u)$, $p(z|u)$, las observaciones z y la función de costo $c(\cdot)$, debemos buscar un estimador $\hat{u}$ que minimice el costo esperado $B(\hat{u}) = E[c(\tilde{u})]$.

A continuación procedemos a desarrollar el valor del costo esperado para tener un expresión explícita la cual nos facilite los cálculos más adelante. Entonces,

$$B(\hat{u}) = E[c(\tilde{u})] = \int_{\mathbb{R}^m} \int_{\mathbb{R}^n} c(u - \hat{u}) p(u, z) du dz,$$

donde $p(u, z)$ es la función de distribución conjunta de u y z. Ahora utilizando la fórmula de Bayes sabemos que:

$$p(u|z) = \frac{p(z|u)p(u)}{p(z)},$$

donde la función

$$p(z) = \int_{\mathbb{R}^n} p(u, z) du = \int_{\mathbb{R}^n} p(z|u)p(u) du$$

es la distribución marginal de z. La función de distribución condicional $p(u|z)$ es conocida como la distribución a posteriori de u dadas las observaciones z. De la estructura que tiene la formula de Bayes, sabemos que esta función de distribución combina la información que se conoce a priori de la incógnita $p(u)$ y la información que proveen las observaciones en términos de la distribución condicional $p(z|u)$ de una manera natural. Ahora, combinando las fórmulas anteriores podemos concluir que:

$$B(\hat{u}) = \int_{\mathbb{R}^m} B(\hat{u}|z) p(z) dz,$$

donde

$$B(\hat{u}|z) = \int_{\mathbb{R}^n} c(u - \hat{u}) p(u|z) du.$$

Puesto que $p(z) \geq 0$, minimizar $B(\hat{u})$ es equivalente a minimizar $B(\hat{u}|z)$. Por tanto, para encontrar el estimador eficiente de Bayes nos vamos a concentrar en la minimización de $B(\hat{u}|z)$.

Dependiendo de la función de costo que se escoja obtenemos diferentes tipos de estimadores eficientes. Por ejemplo, tomando $c(\cdot)$ como la suma ponderada de los errores al cuadrado obtenemos el estimador de Bayes de mínimos cuadrados. Sin embargo, en este trabajo nos vamos a concentrar en el estimador máximo a posteriori (MAP), el cual se obtiene al elegir la función de costo $c(\cdot)$ como la función de costo uniforme para definir $B(u|z)$. Entonces

$$S_\varepsilon := \{u \in \mathbb{R}^n : \ \| u - \hat{u} \|_\infty > \varepsilon\},$$

y

$$S_\varepsilon^c := \mathbb{R}^n - S_\varepsilon = \{u \in \mathbb{R}^n : \ \| u - \hat{u} \|_\infty \leq \varepsilon\}. \tag{2.1}$$

A continuación presentamos una proposición que nos permitirá calcular de mejor manera

$B(\hat{u}|z)$.

Proposición 1. *Sea S_ε^c el conjunto definido en* (2.1)*. Se satisface que:*

$$V(S_\varepsilon^c) = (2\varepsilon)^n,$$

donde $V(S_\varepsilon^c)$ es el volumen del conjunto S_ε^c.

Demostración. Procedemos por inducción sobre n. Así, para $n = 1$ tenemos que

$$\begin{aligned} V(S_\varepsilon^c) &= \int_{S_\varepsilon^c} 1du = \int_{|u-\hat{u}|\leq\varepsilon} 1du \\ &= \int_{-\varepsilon}^{\varepsilon} 1dv = 2\varepsilon. \end{aligned}$$

Suponemos ahora que se cumple para n y vamos a mostrar que también es verdad para $n+1$. Así, tenemos

$$\begin{aligned} V(S_\varepsilon^c) &= \int_{|u^{n+1}-\hat{u}^{n+1}|\leq\varepsilon} \int_{|u^n-\hat{u}^n|\leq\varepsilon} \cdots \int_{|u^1-\hat{u}^1|\leq\varepsilon} 1du^1 \cdots du^n du^{n+1} \\ &= \int_{|u^{n+1}-\hat{u}^{n+1}|\leq\varepsilon} \left\{ \int_{|u^n-\hat{u}^n|\leq\varepsilon} \cdots \int_{|u^1-\hat{u}^1|\leq\varepsilon} 1du^1 \cdots du^n \right\} du^{n+1} \\ &= (2\varepsilon)^n \int_{|u^{n+1}-\hat{u}^{n+1}|\leq\varepsilon} 1du^{n+1} = (2\varepsilon)^{n+1}, \end{aligned}$$

de donde se tiene el resultado. □

Cabe recalcar que este resultado se tiene únicamente cuando tenemos el conjunto S_ε en función de la norma uniforme. Ya que para cualquier $u \in \mathbb{R}^n$, se tiene que:

$$\| u \|_\infty = \max_{1\leq i\leq n} |u_i|,$$

y por tanto

$$\| u \|_\infty \leq \varepsilon \Rightarrow |u_i| \leq \varepsilon, \quad \forall i = 1, \ldots, n.$$

Entonces tenemos

$$B(\hat{u}|z) = \int_{S_\varepsilon} p(u|z)du = 1 - \int_{S_\varepsilon^c} p(u|z)du.$$

Utilizando el teorema del valor medio en su versión integral podemos concluir que existe $\hat{u}$ tal que:

$$B(\hat{u}|z) = 1 - (2\varepsilon)^n p(\hat{u}|z).$$

Así por lo tanto, si queremos minimizar $B(\hat{u}|z)$ debemos maximizar $p(\hat{u}|z)$. Entonces el estimador Máximo A Posteriori de Bayes, al que denominaremos $\hat{u}_{MAP}$ es aquel que satisface

$$p(\hat{u}_{MAP}|z) \geq p(\hat{u}|z), \qquad \forall \hat{u}.$$

En las siguientes secciones mostraremos la forma de obtener este estimador para problemas específicos.

3. Problema 3D-VAR

Esta sección está dedicada a mostrar la equivalencia entre las estimaciones de máxima verosimilitud y estimación bayesiana para el problema 3D-VAR. En primer lugar, vamos a recordar la estructura que tiene este problema:

$$\begin{aligned} \min_{\mathbf{y},u} \quad & J(\mathbf{y},u) = \frac{1}{2}(z-\mathcal{H}(\mathbf{y}))^T R^{-1}(z-\mathcal{H}(\mathbf{y})) + \frac{1}{2}(u-u^b)^T B^{-1}(u-u^b) \\ \text{sujeto a:} \quad & \\ & y_i = u, \quad i = 1, \\ & y_{i+1} = \mathcal{M}_{i+1,i}(y_i), \quad \forall i = \{1,\dots,N\}. \end{aligned} \tag{2.2}$$

donde $\mathbf{y}$ es el arreglo que contiene todos los instantes de la variable de estado y_i. Por otro lado, ya que el operador $\mathcal{M}_i$ depende de los instantes anteriores de y y $y_1 = u$, se lo puede expresar como un solo operador $\mathcal{M}$ y la ecuación de estado puede ser representada como:

$$\mathbf{y} = \mathcal{M}(u).$$

En este caso, asumimos que conocemos una única observación z y el estado previo de la atmósfera u^b, independientes entre sí que satisfacen

$$\begin{aligned} z &= \mathcal{H}(\mathcal{M}(u)) + v, \\ u^b &= u + w, \end{aligned}$$

donde u es la incógnita, v, w son variables aleatorias que siguen una distribución normal con media 0 y matrices de covarianza R y B, respectivamente. Debido a la aditividad de la distribución normal sabemos que $z \sim \mathcal{N}_m(\mathcal{H}(\mathcal{M}(u)), R)$ y $u^b \sim \mathcal{N}_n(u, B)$, donde el subíndice de la distribución normal indica su orden.

3.1. Estimación de Máxima Verosimilitud

Teniendo en cuenta el método descrito en las secciones anteriores debemos calcular la función de probabilidad conjunta condicional de $p(z, u^b|u)$. Puesto que las observaciones son independientes y conocemos de antemano la distribución que siguen tenemos que:

$$\begin{aligned} p(z,u^b|u) = & \frac{1}{(2\pi)^{n+m/2}|B|^{1/2}|R|^{1/2}} \\ & \exp\left\{-\frac{1}{2}(u^b-u)^T B^{-1}(u^b-u) - \frac{1}{2}(z-\mathcal{H}(\mathcal{M}(u)))^T R^{-1}(z-\mathcal{H}(\mathcal{M}(u)))\right\}. \end{aligned}$$

Asumiendo que u es una variable determinista definimos la función de verosimilitud como:

$$L(u|z,u^b) = \frac{1}{(2\pi)^{n+m/2}|B|^{1/2}|R|^{1/2}} \exp\left\{-\frac{1}{2}(u^b-u)^T B^{-1}(u^b-u) - \frac{1}{2}(z-\mathcal{H}(\mathcal{M}(u)))^T R^{-1}(z-\mathcal{H}(\mathcal{M}(u)))\right\}.$$

Entonces, el estimador de máxima verosimilitud es aquel que maximiza la función de verosimilitud. Sin embargo, debido a la dificultad en la resolución de este problema se propone una modificación del mismo a través de la aplicación de la función logaritmo natural a la función de verosimilitud. La equivalencia entre ambos problemas se obtiene gracias a que la función logaritmo natural (ln) es monótona creciente. Así, debemos resolver el problema

$$\begin{aligned}\max_u \ln L(u|z,u^b) = \max_u \ln &\left(\frac{1}{(2\pi)^{n+m/2}|B|^{1/2}|R|^{1/2}}\right) \\ &- \left[\frac{1}{2}(u^b-u)^T B^{-1}(u^b-u) + \frac{1}{2}(z-\mathcal{H}(\mathcal{M}(u)))^T R^{-1}(z-\mathcal{H}(\mathcal{M}(u)))\right].\end{aligned}$$

Entonces, podemos concluir que este problema es equivalente al problema 3D–VAR dado en la ecuación (2.2).

3.2. Estimación Máxima A Posteriori

El método de estimación bayesiana, como se mencionó anteriormente, supone que la incógnita es una variable aleatoria de la cual conocemos a priori su función de distribución. En este caso, asumiremos que dado los pronósticos previos u^b la incógnita u sigue una distribución $u \sim \mathcal{N}_n(u^b, B)$. Además la función de distribución condicional $p(z|u)$ es

$$p(z|u) = \frac{1}{(2\pi)^{m/2}|R|^{1/2}} \exp\left\{-\frac{1}{2}(z-\mathcal{H}(\mathcal{M}(u)))^T R^{-1}(z-\mathcal{H}(\mathcal{M}(u)))\right\}.$$

Sabemos que para encontrar la estimación Máxima A Posteriori es suficiente maximizar $p(u|z)$ la cual usando la fórmula de Bayes está dada por:

$$p(u|z) = \frac{p(z|u)p(u)}{p(z)},$$

y nuevamente puesto que $p(z)$ es constante con respecto a u. El problema se reduce a resolver:

$$\max_u p(z|u)p(u),$$

lo cual, al igual que en el caso anterior es equivalente a resolver

$$\max_u \ln(p(z|u)) + \ln(p(u)),$$

debido a que la función ln es monótona creciente. Así, obtenemos el siguiente problema

$$\max_{u} -\left\{\frac{1}{2}(u^b-u)^T B^{-1}(u^b-u)+\frac{1}{2}(z-\mathcal{H}(\mathcal{M}(u)))^T R^{-1}(z-\mathcal{H}(\mathcal{M}(u)))\right\}.$$

Este problema es equivalente a resolver el problema 3D-VAR dada en la ecuación (2.2). Cabe recalcar que en este método no asumimos nada sobre los pronósticos previos u^b, el análisis se realiza en el sentido opuesto. Es decir, asumimos que la incógnita sigue una distribución normal que tiene como media el valor de u^b. Esta es la principal diferencia con el método de estimación de máxima verosimilitud a pesar de que al final ambos métodos se reducen a resolver el mismo problema.

4. Problema 4D-VAR

El problema 4D-VAR está dado por la siguiente expresión:

$$\begin{aligned}
&\min_{\mathbf{y},u} && J(\mathbf{y},u)=\frac{1}{2}\sum_{i=1}^{N}(z_i-\mathcal{H}_i(y_i))^T R_i^{-1}(z_i-\mathcal{H}_i(y_i))+\frac{1}{2}(u-u^b)^T B^{-1}(u-u^b)\\
&\text{sujeto a:} && \qquad (2.3)\\
&&& y_i=u, \quad i=1,\\
&&& y_{i+1}=\mathcal{M}_{i+1,i}(y_i), \quad \forall i=\{1,\dots,N\}.
\end{aligned}$$

Al igual que para el problema 3D-VAR asumimos que las N observaciones y la información previa son independientes entre sí y que satisfacen:

$$\begin{aligned}
z_i &= \mathbf{H}_i(\mathcal{M}(u))+v_i, \quad \forall i=1,\dots,n.\\
u^b &= u+w.
\end{aligned}$$

donde z_i son las observaciones, u^b es la información previa. Además, $v_i \sim \mathcal{N}_m(0,R_i)$ para cada $i=1,\dots,N$ y $w \sim \mathcal{N}_n(0,B)$.

4.1. Estimación de máxima verosimilitud

Siguiendo las ideas de la sección anterior sabemos que la función de distribución condicional está dada por la siguiente expresión

$$\begin{aligned}
p((z_i,u^b)|u) = &\frac{1}{(2\pi)^{Nm/2+n/2}\prod_{i=1}^{N}|R_i|^{1/2}|B|^{1/2}}\\
&\exp\left\{-\frac{1}{2}(u^b-u)^T B^{-1}(u^b-u)-\frac{1}{2}\sum_{i=1}^{N}(z_i-\mathcal{H}_i(\mathcal{M}(u)))^T R_i^{-1}(z_i-\mathcal{H}_i(\mathcal{M}(u)))\right\},
\end{aligned}$$

por tanto la función de verosimilitud está dada por

$$L(u|(z_i,u^b)) = \frac{1}{(2\pi)^{Nm/2+n/2}\prod_{i=1}^{N}|R_i|^{1/2}|B|^{1/2}} \exp\left\{-\frac{1}{2}(u^b-u)^TB^{-1}(u^b-u) - \frac{1}{2}\sum_{i=1}^{N}(z_i-\mathcal{H}_i(\mathcal{M}(u)))^TR_i^{-1}(z_i-\mathcal{H}_i(\mathcal{M}(u)))\right\}.$$

El estimador de máxima verosimilitud es aquel que resuelve el problema

$$\begin{aligned}&\max_u \ln(L(u|(z_1,\ldots,z_N,u^b)))\\ &\quad = \max_u -\left\{\frac{1}{2}(u^b-u)^TB^{-1}(u^b-u) + \frac{1}{2}\sum_{i=1}^{N}(z_i-\mathcal{H}_i(\mathcal{M}(u)))^TR_i^{-1}(z_i-\mathcal{H}_i(\mathcal{M}(u)))\right\},\end{aligned}$$

lo cual es equivalente a resolver el problema de asimilación de datos 4D-VAR dado en (2.3).

4.2. Estimación máxima a posteriori

Asumimos la independencia en las N observaciones y además suponemos que la incógnita es una variable aleatoria que sigue un distribución $u \sim \mathcal{N}(u^b, B)$. Usando un razonamiento similar al de la sección anterior tenemos que el estimador máximo a posteriori es aquel que resuelve el problema

$$\max_u \ln p(u|z) + \ln p(u),$$

es decir,

$$\max_u -\left\{\frac{1}{2}(u^b-u)^TB^{-1}(u^b-u) + \frac{1}{2}\sum_{i=1}^{N}(z_i-\mathcal{H}_i(\mathcal{M}(u)))^TR_i^{-1}(z_i-\mathcal{H}_i(\mathcal{M}(u)))\right\},$$

el cual es equivalente a resolver el problema 4D-VAR dado en (2.3). Los razonamientos realizados en las últimas dos secciones fueron desarrollados a partir de [Kalnay, 2003]

5. Problema 4D-VAR con regularización de variación total

Las soluciones del problema de asimilación de datos, como se mencionó anteriormente, no necesariamente son continuas. Esta información puede ser considerada como la información a priori de la incógnita. Es por eso de vital importancia el estudio de la interpretación estadística de los problemas inversos que involucran términos como el de variación total. Los resultdos presentados a continuación fueron desarrollados siguiendo las ideas de [Lee and Kitanidis, 2013]. Al igual que en los casos anteriores comenzamos recordando la

estructura que tiene el problema:

$$\begin{aligned} \min_{(y,u)} \quad & J(y,u) = \frac{1}{2}\sum_{i=1}^{N}(z_i - \mathcal{H}_i(y_i))^T R_i^{-1}(z_i - \mathcal{H}_i(y_i)) + \frac{1}{2}(u-u^b)^T B^{-1}(u-u^b) \\ & \quad +\beta \sum_{i=1}^{n} |D_i u| \\ \text{sujeto a:} \quad & \\ & y_i = u, \quad i = 1, \\ & y_{i+1} = \mathcal{M}_{i+1,i}(y_i), \quad \forall i = \{1,\ldots,N\}. \end{aligned} \tag{2.4}$$

Para este análisis definimos el vector $\bar{z} = (z, u^b)$ y el operador no lineal $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^n$, tal que

$$\mathcal{E}(u) = (\mathcal{H}(\mathcal{M}(u)), u)).$$

Así, el funcional objetivo del problema (2.4) puede ser reescrito como:

$$J(u) = \frac{1}{2}(\mathcal{E}(u) - \bar{z})^T G^{-1}(\mathcal{E}(u) - \bar{z}) + \beta \sum_{i=1}^{n} |D_i u|,$$

donde

$$G^{-1} = \begin{bmatrix} R^{-1} & 0 \\ 0 & B^{-1} \end{bmatrix}.$$

A continuación se presenta el análisis de estimación (MAP) para este problema. Comenzamos suponiendo que la función de distribución condicional de las observaciones $\bar{z}$ dado u, está dado por:

$$p(\bar{z}|u) = \frac{1}{(2\pi)^{m/2}|G|^{1/2}} \exp\left\{-\frac{1}{2}(\bar{z} - \mathcal{E}(u))^T G^{-1}(\bar{z} - \mathcal{E}(u))\right\}.$$

Ahora, como se mencionó en la introducción, la principal característica de la regularización de variación total es que nos permite conservar los puntos de discontinuidades o *sharp fronts* que tiene la solución. Con esta información apriori asumiremos que u sigue una distribución de Laplace de Du, cuya función de distribución es:

$$p(u) = \frac{1}{\theta^{2n}} \exp\left\{-\frac{1}{\theta}\sum_{i=1}^{n} |D_i u|\right\}.$$

Usando la fórmula de Bayes sabemos que la distribución a posteriori está dada por la expresión:

$$p(u|\bar{z}) = \frac{p(\bar{z}|u)p(u)}{p(\bar{z})},$$

puesto que $p(\bar{z})$ es constante con respecto a u. El estimador (MAP) del problema debe resolver el problema de

$$\max_u \ln(p(\bar{z}|u)) + \ln(p(u)),$$

es decir,

$$\max_{u} cte - \left\{ \frac{1}{2}(\bar{z} - \mathcal{E}(u))^T G^{-1}(\bar{z} - \mathcal{E}(u)) + \frac{1}{\theta}\sum_{i=1}^{n} |D_i u| \right\}.$$

El cual es equivalente al problema (2.4) con el parámetro $\beta = 1/\theta$. Dada la estructura del problema no se puede aplicar un método de máxima verosimilitud.

6. Problema 4D-VAR con regularización TGV

Este problema, al igual que el anterior, tienen la característica de conservar los *sharp fronts* en las soluciones y esta es la información a priori que usamos para la estimación de la probabilidad a posteriori. Recordando el problema que involucra la regularización TGV tenemos

$$\begin{aligned} \min_{y,u,w} \quad & J(y,u,w) = \frac{1}{2}\sum_{i=1}^{N}(z_i - \mathcal{H}_i(y_i))^T R_i^{-1}(z_i - \mathcal{H}_i(y_i)) + \frac{1}{2}(u-u^b)^T B^{-1}(u-u^b) \\ & \quad +\alpha\sum_{i=1}^{n}|D_i u - w_i| + \beta\sum_{i=1}^{n}|E_i w| \\ \text{sujeto a:} \quad & \\ & y_i = u, \quad i = 1, \\ & y_{i+1} = \mathcal{M}_{i+1,i}(y_i), \quad \forall i = \{1,\dots,N\}. \end{aligned} \tag{2.5}$$

Realizando los siguientes cambios de variable $\bar{z} = (z, u^b)$ y $x = (u, w)$ y definiendo los operadores $\mathcal{E}(x) = \mathcal{E}(u,w) = (\mathcal{H}(\mathcal{M}(u)), u)$ y la matriz

$$\mathbb{D} = \begin{bmatrix} \alpha D & -\alpha\mathbb{I} \\ 0 & \beta E \end{bmatrix}.$$

El funcional objetivo del problema (2.5) se puede reescribir como:

$$J(x) = \frac{1}{2}(\mathcal{E}(x) - \bar{z})^T G^{-1}(\mathcal{E}(x) - \bar{z}) + \sum_{i=1}^{2n-1}|\mathbb{D}_i x|.$$

Asumimos que la distribución de probabilidad condicional dado el vector $\bar{z}$ está dada por:

$$p(\bar{z}|x) = \frac{1}{(2\pi)^{m/2}|G|^{1/2}}\exp\left\{-\frac{1}{2}(\bar{z} - \mathcal{E}(x))^T G^{-1}(\bar{z} - \mathcal{E}(x))\right\},$$

y la información a priori sobre x se asume que sigue una distribución de Laplace de $\mathbb{D}x$ dada por la expresión:

$$p(x) = \exp\left\{-\sum_{i=1}^{2n-1}|\mathbb{D}_i x|\right\}.$$

Usando la fórmula de Bayes sabemos que la distribución a posteriori está dada por

$$p(x|\bar{z}) = \frac{p(\bar{z}|x)p(x)}{p(\bar{z})},$$

y por un razonamiento similar al anterior, podemos concluir que el estimador MAP debe resolver el problema

$$\max_x \ln(p(\bar{z}|x)) + \ln(p(x)),$$

es decir,

$$\max_u -\left\{ \frac{1}{2}(\bar{z} - \mathcal{E}(x))^T G^{-1}(\bar{z} - \mathcal{E}(x)) + \sum_{i=1}^{2n-1} |\mathbb{D}_i x| \right\},$$

el cual es equivalente a resolver el problema (2.5).

Capítulo 3

Condiciones de Optimalidad

El principal objetivo de este capítulo es derivar las condiciones de optimalidad del problema de asimilación de datos con la ecuación de Burgers como ecuación de estado. Este resultado será obtenido utilizando el enfoque de *discretizar–luego–optimizar*; el cual como su nombre lo indica, se basa en la discretización completa del problema para luego obtener las condiciones de optimalidad del mismo. Así, se abordarán temas como la discretización de la ecuación de estado. Además, debido a la no diferenciabilidad de las funciones objetivo se presentan técnicas de regularización, las cuales nos permiten calcular las primeras derivadas de las mismas y así poder presentar formalmente los sistema de optimalidad para los problemas de interés. Una vez obtenido el sistema de optimalidad, sabremos la estructura de la ecuación adjunta y usando resultados de ecuaciones en diferencias podremos garantizar la existencia y unicidad de soluciones. Finalmente, puesto que se resolverán los problemas regularizados se mostrará que las soluciones de los problemas regularizados convergen hacia las soluciones de los problemas originales.

1. Ecuación de Burgers

Esta sección está dedicada al estudio de la ecuación de Burgers. En lo que sigue analizaremos aspectos teóricos sobre la ecuación de estado como el análisis de existencia y unicidad de las soluciones. Además, presentamos las técnicas de discretización utilizados para la resolución numérica de la misma.

1.1. Existencia y unicidad de soluciones

El análisis de existencia y unicidad de las soluciones se la realizará tomando como referencia la Tesis doctoral [Volkwein, 1997]. El mismo que será realizado en espacios de dimensión infinita.

Comenzamos la sección presentando la notación que se usará. Se definen los espacios $V = H_0^1(\Omega)$, $H = L^2(\Omega)$ y $W(V) = W(0,T;V)$ cuya definición está dada por

$$W(0,T;V) = \{\varphi\colon \varphi \in L^2(V), \varphi_t \in L^2(V^*)\}.$$

donde φ_t es la primera derivada temporal de la función $\varphi(x,t)$. Además, notamos al producto interno usual del espacio de Hilbert H por $\langle\cdot,\cdot\rangle_H$, mientras que el producto $\langle\cdot,\cdot\rangle_{V^*,V}$ es el producto en dualidad de V un espacio real normado y V^* su correspondiente espacio dual.

Consideramos la ecuación de Burgers con condiciones de borde tipo Dirichlet homogéneas dada en la ecuación (1.3) con la siguiente estructura:

$$\begin{aligned} \frac{\partial y(x,t)}{\partial t} - \nu\frac{\partial^2 y(x,t)}{\partial x^2} + \frac{\partial y(x,t)}{\partial x}\cdot y(x,t) &= f(x,t) && \text{en } Q = \Omega\times[0,T],\\ y(x,t) &= 0 && \text{sobre } \Gamma\times[0,T],\\ y(x,0) &= u(x) && \text{en } \Omega. \end{aligned}$$

donde ν es el parámetro de viscosidad del fluido, $f \in L^2(V^*)$ y $u \in L^2(\Omega)$. Notaremos por $y(t) = y(\cdot,t)$ a las funciones que dependen solo de x cuando t es fijo.

Definición 1. *Una función $y \in W(V)$ se llama solución débil de* (1.3) *si*

$$\langle y_t(t),\varphi\rangle_{V^*,V} + \nu\langle y(t),\varphi\rangle_V + b(y(t),y(t),\varphi) = \langle f(t),\varphi\rangle_{V^*,V} \tag{3.1}$$

para todo $\varphi \in V$ y para casi todo $t \in [0,T]$ y

$$y(0) = u \tag{3.2}$$

se satisface. Además, notamos por $b(\varphi,\phi,\psi)$ a la forma tri–lineal dada por

$$b(\varphi,\phi,\psi) = \frac{1}{3}\int_\Omega (\varphi\phi)'\psi + \varphi\phi'\psi dx$$

para todo $\varphi,\phi,\psi \in H^1(\Omega)$.

El siguiente Teorema nos garantiza la existencia de una única solución $y \in W(V)$ que satisface (3.1)– (3.2)

Teorema 1. *Existe una única solución débil para* (1.3) *si $f \in L^2(V^*)$ y $u \in L^2(\Omega)$.*

Demostración. La demostración de este Teorema fue tomada de [Volkwein, 1997], Teorema 2.2, Capítulo 5. □

1.2. Discretización

La ecuación de Burgers es una ecuación de Advección–Difusión ya que contiene un término de difusión dado por

$$-\nu \frac{\partial^2 y}{\partial x^2},$$

y el término advectivo o de transporte dado por

$$y \frac{\partial y}{\partial x}.$$

Además, ya que vamos a tomar $\nu \ll 1$, se puede considerar a esta como una ecuación cuyo término de transporte es dominante. La discretización de la ecuación será realizada utilizando el esquema de diferencias finitas para las variables espaciales y un esquema de Euler semi-implícito para la variable temporal.

Discretización espacial

La idea principal del esquema de discretización con diferencias finitas se basa en la partición del dominio espacial. Esta partición puede ser realizada uniformemente o con mallas adaptivas.

Los elementos de la partición $\{x_j\}_{j=0}^{n}$ son de la forma:

$$x_j = jh, \qquad \forall j = 1, \ldots, n,$$

donde $h = 1/n+1$ y n es la cantidad de puntos de la partición. A este tipo de particiones se les conoce como *uniformes* y corresponden al tipo de particiones con las que vamos a trabajar.

Los puntos de la partición servirán como referencia para la creación de vectores en los cuales cada entrada corresponde al valor de la función en un punto de la malla. Una vez obtenidos estos vectores se propone el uso de un operador diferencial discreto para la segunda derivada espacial, el cual tiene la siguiente estructura:

$$y'' \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}.$$

Reemplazando en la ecuación de Burgers tenemos que, para cada i perteneciente a la partición del dominio espacial,

$$\frac{\partial y_i}{\partial t} - \nu \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + y_i (y_i)' = f_i,$$

donde $h = 1/n+1$ es el tamaño de la discretización y n es la cantidad de puntos de la partición.

Ahora nos corresponde discretizar el término $y_i(y_i)'$. Siguiendo la lógica anterior se podría pensar que usar un esquema con diferencias centradas podría funcionar sin inconvenientes. Sin embargo, de trabajos pasados se conoce que este tipo de discretización aplicada a ecuaciones cuyo término de transporte es dominante, genera oscilaciones en las soluciones. Este

tipo de oscilaciones no corresponden a la dinámica del sistema en estudio y por tanto generan soluciones incorrectas. De [Griebel et al., 1998] se sabe que utilizando mallas suficientemente finas se puede evitar este fenómeno. Sin embargo, utilizar mallas muy finas resulta en el incremento considerable del tamaño de los sistemas a resolverse lo cual para problemas de dos o tres dimensiones es poco conveniente.

Otra solución al problema de estabilidad en este tipo de ecuaciones consiste en reemplazar el esquema de diferencias centradas por un esquema de diferencias hacia *atrás* si el coeficiente del término de transporte es positivo y un esquema de diferencias hacia *adelante* en caso contrario. Matemáticamente, si se desea discretizar el término $\beta u'$, esta condición se puede expresar de la siguiente manera:

$$\begin{aligned} \text{si } \beta > 0 \Rightarrow u_i' &= \frac{u_i - u_{i-1}}{h} \\ \text{si } \beta < 0 \Rightarrow u_i' &= \frac{u_{i+1} - y_i}{h}. \end{aligned}$$

La principal ventaja de usar este esquema es que los valores propios de la matriz resultante al discetizar completamente la ecuación tendrán su parte real positiva lo que resulta en una estabilización de la discretización, ver [Griebel et al., 1998].

Aplicando estos resultados a nuestro problema en el cual el coeficiente del término de transporte es y_i se tiene la siguiente regla:

$$\begin{aligned} \text{si } y_i > 0 \Rightarrow y_i' &= \frac{y_i - y_{i-1}}{h} \\ \text{si } y_i < 0 \Rightarrow y_i' &= \frac{y_{i+1} - y_i}{h}. \end{aligned}$$

Así, tenemos que escoger un esquema de discretización para cada componente del vector y. Por lo tanto, se puede definir la siguiente matriz, la cual notaremos por U_p y la estructura de cada una de sus filas está dada por (3.3).

$$(U_p)_{i\cdot} = \frac{1}{h} \begin{cases} U_p(i,i) = 1 \text{ y } U(i,i-1) = -1 & \text{si } y_i > 0 \text{ y } i \neq 1, \\ U_p(i,i) = 1 & \text{si } y_i > 0 \text{ y } i = 1, \\ U_p(i,i) = -1 \text{ y } U(i,i+1) = 1 & \text{si } y_i < 0 \text{ y } i \neq n, \\ U_p(i,i) = -1 & \text{si } y_i < 0 \text{ y } i = n, \\ U_p(i,j) = 0 & \text{si } j \neq i. \end{cases} \tag{3.3}$$

Es importante mencionar que el esquema *Upwinding* produce un error local de discretización de $O(h)$ al contrario del $O(h^2)$ que se obtiene en el caso de las diferencias centradas. Sin embargo, el efecto de estabilización que ofrece este esquema es de vital importancia para obtener la solución correcta del problema.

Por otro lado, en el caso del término de segundo orden se usará la conocida matriz corres-

pondiente al Laplaciano discreto, dada por la siguiente expresión:

$$A = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

Discretización temporal

Para la discretización temporal usaremos un esquema semi–implícito de Euler dada por la siguiente expresión:

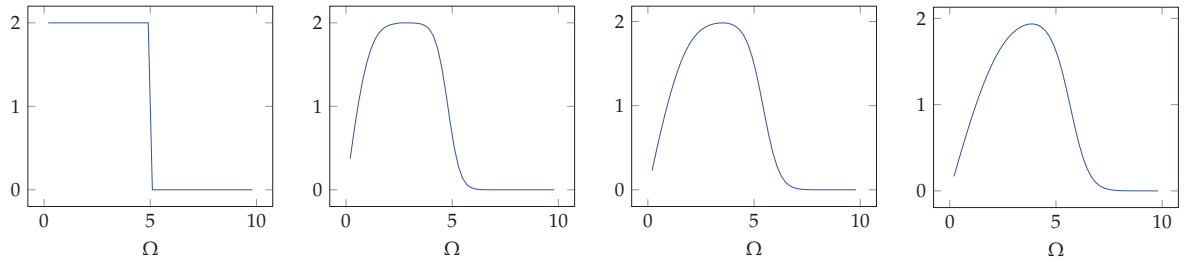
$$\frac{y^{j+1} - y^j}{\Delta t} + \nu A y^{j+1} + y^j U_p^j(y^{j+1}) = f^{j+1},$$

donde $\Delta t = 1/(N_t + 1)$ y N_t es el número de puntos en la discretización temporal. La principal ventaja de escoger este tipo de esquemas es que el sistema a resolver es lineal y tiene la siguiente estructura:

$$y^{j+1} + \Delta t \nu A y^{j+1} + \Delta t \text{diag}(y^j) U_p^j y^{j+1} = \Delta t f^{j+1} + y^j, \tag{3.4}$$

para $j = 2, \dots, N_t$. Mientras que la condición inicial dice que $y^1 = u$.

En la Figura 1 se presentan 4 instantes diferentes de la solución de la ecuación de Burgers obtenida con el esquema presentado anteriormente. Para su solución utilizamos los siguientes parámetros, $f \equiv 0, \nu = 0{,}6$. Tomamos 50 puntos de discretización espacial y 100 en la temporal. Además, la condición inicial para este problema es $u = 2\mathbb{I}_{\{x:\, 0 \le x \le 0{,}5\}}$ donde $\mathbb{I}_A$ es la función indicatriz del conjunto A.



(a) t=0 **(b)** t=0.25 **(c)** t=0.5 **(d)** t=0.75

Figura 1. Instantáneas de la ecuación de Burgers con $\nu = 0{,}6$

Otra opción válida en el caso de la discretización de la ecuación de estado consiste en realizar una discretización completa en el tiempo y en el espacio, la cual nos permita escribir

la ecuación de estado como una sola ecuación no lineal. Para esto, comenzamos notando que la solución de la ecuación de Burgers $y \in \mathbb{R}^{n \times N_t}$ puede ser reescrita como un vector con la siguiente estructura:

$$\mathbf{y} = \begin{bmatrix} y^1 \\ \vdots \\ y^{N_t} \end{bmatrix},$$

donde $y^i \in \mathbb{R}^n$; de este modo $\mathbf{y} \in \mathbb{R}^m$ con $m = n \cdot N_t$. En este contexto definimos además las siguientes matrices:

$$\mathbb{A} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & A & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A \end{bmatrix}, \qquad \mathbb{U} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & U_p^1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & U_p^{N_t-1} \end{bmatrix}$$

$$\mathbb{Z} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & \operatorname{diag}(y^1) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \operatorname{diag}\,(y^{N_t-1}) \end{bmatrix}$$

y

$$\mathbb{E} = \frac{1}{\Delta t} \begin{bmatrix} \mathbb{I} & 0 & 0 & \dots & 0 & 0 \\ -\mathbb{I} & \mathbb{I} & 0 & \dots & 0 & 0 \\ 0 & -\mathbb{I} & \mathbb{I} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\mathbb{I} & \mathbb{I} \end{bmatrix}, \tag{3.5}$$

donde 0 es la matriz de ceros de tamaño $n \times n$. Además, definimos el vector

$$\mathbf{f}(u) = \begin{bmatrix} u/\Delta t \\ f^2 \\ \vdots \\ f^{N_t} \end{bmatrix},$$

así la ecuación de estado puede ser reescrita como

$$e(\mathbf{y}, u) = \mathbb{E}\mathbf{y} + \nu\mathbb{A}\mathbf{y} + \mathbb{Z}(\mathbf{y})\mathbb{U}\mathbf{y} - \mathbf{f}(u) = 0. \tag{3.6}$$

Como mencionamos anteriormente, la solución de la ecuación de estado se redujo a la solución de un sistema no lineal. Debido al tamaño de las matrices definidas anteriormente, este esquema no es aplicable desde el punto de vista computacional; por esta razón, para su solución procedemos a resolver los N_t sistemas lineales resultantes dados en la ecuación (3.4).

Nota: La matriz U_p dada en (3.3) depende directamente de la variable $\mathbf{y}$; sin embargo, de ahora en adelante asumiremos que esta matriz es constante y corresponde a la discretización del gradiente. El esquema *upwinding* será usado únicamente para la resolución numérica del problema.

2. Asimilación de datos con la ecuación de Burgers

De lo mencionado en la introducción sabemos que el problema de asimilación de datos con la regularización de variación total con el cual vamos a trabajar está dado por la expresión:

$$\begin{aligned}\min_{(y,u)\in\mathbb{R}^m\times\mathbb{R}^n} J(y,u) = &\frac{1}{2}\sum_{i=1}^{N}(z_i-\mathcal{H}_i(y_i))^T R_i^{-1}(z_i-\mathcal{H}_i(y_i)) + \frac{1}{2}(u-u^b)^T B^{-1}(u-u^b)\\ &+\beta\sum_{i=1}^{n}|D_i u|\end{aligned} \tag{3.7}$$

sujeto a:

$$e(\mathbf{y},u)=0,$$

donde $e(\mathbf{y},u)=0$ está dada en (3.6) y corresponde a la ecuación completamente discretizada. Además, la matriz del gradiente discreto es una matriz de $n-1$ filas y n columnas la cual nos garantiza la estabilidad de la discretización y está dada por la siguiente expresión:

$$D=\frac{1}{h}\begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0\\ -1 & 1 & 0 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \dots & -1 & 1\end{bmatrix}. \tag{3.8}$$

Por otro lado, el problema con regularización de variación total generalizada está dado

por:

$$\begin{aligned}\min_{(y,u,w)\in\mathbb{R}^m\times\mathbb{R}^n\times\mathbb{R}^n} J(y,u,w) &= \frac{1}{2}\sum_{i=1}^{N}(z_i-\mathcal{H}_i(y_i))^T R_i^{-1}(z_i-\mathcal{H}_i(y_i)) \\ &+\frac{1}{2}(u-u^b)^T B^{-1}(u-u^b) \\ &+\alpha\sum_{i=1}^{n}|D_i u-w_i|+\beta\sum_{i=1}^{n}|E_i w| \\ \text{sujeto a:}\qquad & e(\mathbf{y},u)=0,\end{aligned} \tag{3.9}$$

donde $e(\mathbf{y},u)=0$ está dada en (3.6), D esta dada en (3.8). La matriz E que corresponde al gradiente simetrizado, en este caso debido a que trabajamos únicamente con una variable espacial coincide con el gradiente usual discretizado dado por la siguiente expresión:

$$E=\frac{1}{h}\begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ 0 & 0 & -1 & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & 1 \\ 0 & 0 & 0 & \dots & -1 \end{bmatrix}.$$

3. Discretización del operador $\mathcal{H}(\cdot)$

Esta sección está dedicada al estudio del operador $\mathcal{H}(\cdot)$, cuya principal función es extraer la información de la variable de estado ubicada en los puntos de la malla que coinciden con información que poseemos de las observaciones. Para esto, asumimos en primer lugar que conocemos de antemano la ubicación de las observaciones tanto en el tiempo como en el espacio. Así, denominaremos n_o la cantidad de puntos en el espacio en el que se toman las observaciones y por N_o a la cantidad de instantes en el tiempo en los que se toman dichas observaciones. Además, recordamos la definición del índice $m=nN_t$ y definimos $m_o=n_oN_o$. De este modo las observaciones están ordenadas en un vector $\mathbf{z}\in\mathbb{R}^{m_o\times 1}$, mientras que $\mathbf{y}\in\mathbb{R}^{m\times 1}$. Así, el operador $\mathcal{H}$ está dado por:

$$\begin{array}{rccc}\mathcal{H}: & \mathbb{R}^{m\times 1} & \to & \mathbb{R}^{m_o\times 1} \\ & \mathbf{y} & \mapsto & \mathcal{H}(\mathbf{y}).\end{array}$$

Este operador no necesariamente es lineal; sin embargo, en este trabajo lo asumiremos. Así, vamos a deducir la existencia de matrices que describen este operador. En primer lugar, definimos el espacio

$$\mathcal{O}:=\{(i,j)\colon 1\le i\le n, 1\le j\le N_t\colon \text{ existe una observación en el punto } i \text{ en el instante } j\}.$$

Entonces definimos la matriz H_i la cual corresponde a extraer de una matriz identidad $n \times n$ las filas que corresponden a los indices en $\mathcal{O}$. Luego definimos la matriz H la cual es una matriz diagonal por bloques tal que el bloque H_{jj} está dado por

$$H_{jj} = \begin{cases} H_i & \text{si } (i,j) \in \mathcal{O}, \\ 0 & \text{si no.} \end{cases}$$

Finalmente, definimos la matriz $S = F \otimes \mathbb{I}_{n_o \times n_o}$, con F la matriz que está formada al extraer las filas de la matriz identidad $N_t \times N_t$ que corresponden a los instantes en los que se toman las observaciones. La operación $(\otimes)$ es el producto tensorial de Kronecker. Antes de presentar la forma que tendrán las funciones objetivo vamos a mostrar un ejemplo simple de la forma que tienen las matrices descritas en esta sección.

EJEMPLO 1. Sea $n = 2$, $N_t = 3$, $N = 1$ y $M = 2$. Entonces $n_o = 2$ y $N_o = 2$ y tenemos el conjunto

$$\mathcal{O} = \{(1,1), (2,1), (1,3), (2,3)\}.$$

Esto significa que tenemos observaciones en todos los puntos del espacio pero solo en el primer y último instante en el tiempo. Entonces, siguiendo los lineamientos descritos anteriormente, la matriz H está dada por la siguiente expresión:

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

con H_i la matriz

$$H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

La matriz S está dada por

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Podemos verificar entonces que si $\mathbf{y} = [1,2,3,4,5,6]^T$, se tiene que

$$SH\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 5 \\ 6 \end{bmatrix},$$

Las entradas del vector resultante corresponden a las ubicaciones definidas por el espacio $\mathcal{O}$, por tanto podrán ser comparadas con las observaciones.

Utilizando estas matrices la función objetivo del problema con la regularización TV está dada por:

$$J(\mathbf{y}, u) = \frac{1}{2}(\mathbf{z} - SH\mathbf{y})^T R^{-1}(\mathbf{z} - SH\mathbf{y}) + \frac{1}{2}(u - u^b)^T B^{-1}(u - u^b) + \beta \sum_{i=1}^{n} |D_i u|.$$

De la misma manera, se puede reemplazar el operador $\mathcal{H}(\cdot)$ en la función objetivo con la regularización TGV obteniendo la siguiente función objetivo:

$$J(\mathbf{y}, u, w) = \frac{1}{2}(\mathbf{z} - SH\mathbf{y})^T R^{-1}(\mathbf{z} - SH\mathbf{y}) + \frac{1}{2}(u - u^b)^T B^{-1}(u - u^b) + \alpha \sum_{i=1}^{n} |D_i u - w_i| + \beta \sum_{i=1}^{n} |E_i w|$$

4. Función objetivo no diferenciable

La resolución numérica de los problemas de optimización no lineal asume la diferenciabilidad de la función objetivo. En el caso de los problemas (3.7) y (3.9) sus funciones objetivo contienen términos no diferenciables, por lo tanto es necesario realizar cambios a las mismas de tal manera que se pueda garantizar la diferenciabilidad y así poder aplicar los algoritmos de optimización usuales. Los términos no diferenciables en estas funciones objetivo corresponden a la función valor absoluto, la cual contiene un solo punto de no diferenciabilidad.

En este trabajo, se propone el uso de la regularización de Huber la cual tiene un buen desempeño debido a que su modificación es de carácter local. Proponemos el uso de la regularización de Huber usual para los algoritmos que asuman la diferenciabilidad de la función objetivo y a la cual de ahora en adelante nos referiremos como la regularización de Huber $\mathcal{C}^1$, haciendo referencia a que es solo una vez continuamente diferenciable. Además, se propone el uso de la regularización de Huber dos veces diferenciable o regularización de Huber $\mathcal{C}^2$ para aquellos algoritmos que lo requieran. Cada una de estas funciones tienen asociada un parámetro γ el cual a medida que va tomando valores más grandes se va acercando cada vez más a la función del valor absoluto. De ahora en adelante nos referiremos a esta regularización como la γ–regularización del problema. La regularización de Huber $\mathcal{C}^1$ está dada por la siguiente

expresión

$$H_\gamma(t) = \begin{cases} \dfrac{\gamma}{2}t^2 & \text{if } |t| \leq \dfrac{1}{\gamma}, \\ |t| - \dfrac{1}{2\gamma} \text{ if } |t| > \dfrac{1}{\gamma}. \end{cases} \tag{3.10}$$

Mientras que la regularización de Huber $\mathcal{C}^2$ está dada por la expresión:

$$H_\gamma(t) = \begin{cases} |t| + C_1 + K & \text{si } t \in \mathcal{A}, \\ \dfrac{\gamma}{2}t^2 & \text{si } t \in \mathcal{B}, \\ G|t| + \dfrac{H}{2}|t|^2 + \dfrac{C}{3}|t|^3 + D & \text{si } t \in \mathcal{I}, \end{cases} \tag{3.11}$$

donde:

$$\mathcal{A} := \left\{ t \in \mathbb{R} \colon \gamma|t| \geq 1 + \frac{1}{2\gamma} \right\},$$
$$\mathcal{B} := \left\{ t \in \mathbb{R} \colon \gamma|t| \leq 1 - \frac{1}{2\gamma} \right\},$$
$$\mathcal{I} := \left\{ t \in \mathbb{R} \colon |\gamma|t| - 1| \leq \frac{1}{2\gamma} \right\},$$

las constantes $u_1 = \dfrac{1}{\gamma}\left(1 - \dfrac{1}{2\gamma}\right)$, $u_2 = \dfrac{1}{\gamma}\left(1 + \dfrac{1}{2\gamma}\right)$ y:

$$G = 1 - \frac{(2\gamma+1)^2}{8\gamma} \quad H = \frac{\gamma}{2}(2\gamma+1)$$
$$C = -\frac{\gamma^3}{2} \quad D = \left(\frac{\gamma}{2} - \frac{H}{2}\right)u_1^2 - G|u_1| - \frac{C}{3}|u_1|^3$$
$$C_1 = \frac{\gamma}{2}u_1^2 - u_2 \quad K = G(u_2 - u_1) + \frac{H}{2}(u_2^2 - u_1^2) + \frac{C}{3}(u_2^3 - u_1^3)$$

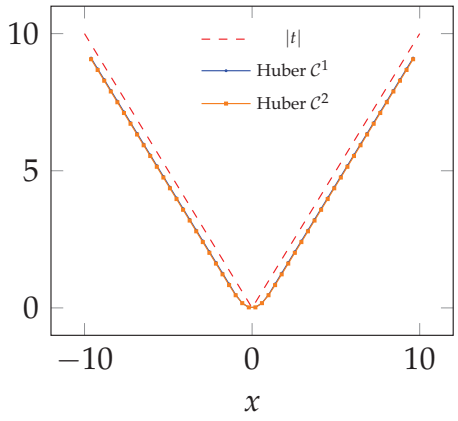
En la Figura 2, se muestra la comparación entre la función valor absoluto y las regularizaciones de Huber $\mathcal{C}^1$ y $\mathcal{C}^2$ con $\gamma = 1$.

La elección de la función de regularización se la realizará dependiendo del algoritmo que se use para la solución del problema. Los algoritmos de optimización utilizados en este trabajo utilizan la información de las primeras y segundas derivadas de las regularizaciones, por tanto presentamos a continuación sus expresiones.

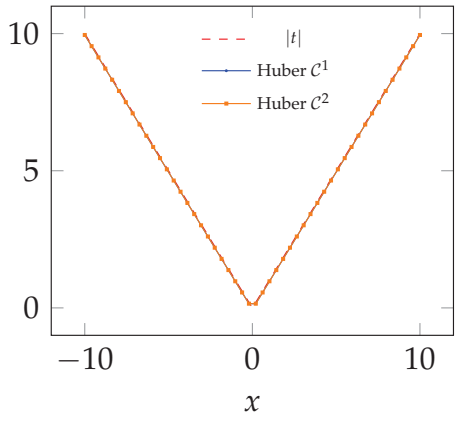
La derivada de primer orden de la regularización de Huber $\mathcal{C}^1$ está dada por la expresión (3.12):

$$h_\gamma(t) = \frac{\gamma t}{\text{máx}\{\gamma|t|, 1\}} \tag{3.12}$$

Mientras que la primera derivada de la regularización de Huber $\mathcal{C}^2$ está dada en la expre-



(a) Regularizaciones de Huber $\mathcal{C}^1$ y $\mathcal{C}^2$ y la función $|\cdot|$ con $\gamma = 1$



(b) Regularizaciones de Huber $\mathcal{C}^1$ y $\mathcal{C}^2$ y la función $|\cdot|$ con $\gamma = 10$

Figura 2. Comparación de Regularizaciones y la función $|\cdot|$

sión (3.13)

$$h_\gamma(t) = \begin{cases} \dfrac{t}{|t|} & \text{si } t \in \mathcal{A}, \\ \gamma t & \text{si } t \in \mathcal{B}, \\ \dfrac{t}{|t|}\left(1 - \dfrac{\gamma}{2}\left(1 - \gamma|t| + \dfrac{1}{2\gamma}\right)^2\right) & \text{si } t \in \mathcal{I}. \end{cases} \tag{3.13}$$

En la Figura 3, se muestran las derivadas de primer orden de las dos regularizaciones.

Finalmente, puesto que para el método de Newton globalizado necesitamos conocer la segunda derivada de la regularización de Huber $\mathcal{C}^2$, mostramos su expresión en la ecuación (3.14)

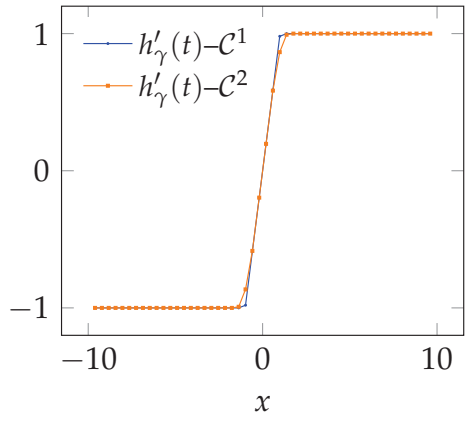
$$h'_\gamma(t) = \begin{cases} \left[\dfrac{1}{|t|} - \dfrac{(t)\odot(t)}{|t|^3}\right] & \text{si } t \in \mathcal{A}, \\ \gamma 1 & \text{si } t \in \mathcal{B}, \\ \left\{\left(1 - \dfrac{\gamma}{2}\theta_\gamma^2\right)\left[\dfrac{1}{|t|} - \dfrac{(t)\odot(t)}{|t|^3}\right] + \gamma^2\theta_\gamma\dfrac{t\odot t}{|t|^2}\right\} & \text{si } t \in \mathcal{I}, \end{cases} \tag{3.14}$$

donde $\theta_\gamma(t) = 1 - \gamma|t| + \dfrac{1}{2\gamma}$. En el caso de que $t \in \mathbb{R}$ el producto $(\odot)$ puede ser reemplazado por el producto usual. Sin embargo, en el caso de que la variable sea un vector entonces el producto $(\odot)$ representa el producto componente a componente.

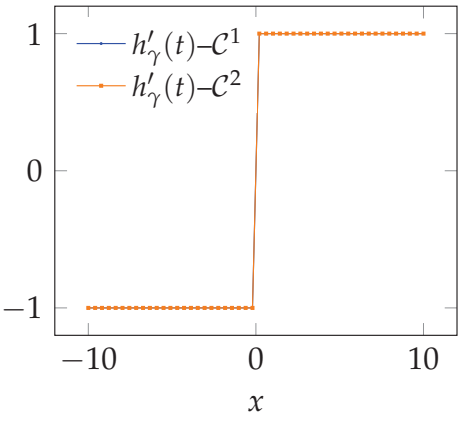
Una vez que hemos introducido las diferentes funciones de regularización procedemos a mostrar la estructura que tendrán las funciones objetivo cuando incluimos la γ–regularización. En este caso no vamos a especificar el tipo de regularización de Huber que se va a utilizar.

La función objetivo γ–regularizada del problema con la regularización de variación total está dada en la ecuación (3.15)

$$J_\gamma(\mathbf{y}, u) = \frac{1}{2}(\mathbf{z} - SH\mathbf{y})^T R^{-1}(\mathbf{z} - SH\mathbf{y}) + \frac{1}{2}(u - u^b)^T B^{-1}(u - u^b) + \beta\sum_{i=1}^{n} H_\gamma(D_i u), \tag{3.15}$$



(a) Primera derivada de las regularizaciones de Huber $\mathcal{C}^1$ y $\mathcal{C}^2$ con $\gamma = 1$



(b) Primera derivada de las regularizaciones de Huber $\mathcal{C}^1$ y $\mathcal{C}^2$ con $\gamma = 10$

Figura 3. Comparación de las primeras derivadas de las regularizaciones

y la función objetivo γ–regularizada para el problema con la regularización de variación total generalizada está dada en la ecuación (3.16)

$$J_\gamma(\mathbf{y}, u, w) = \frac{1}{2}(\mathbf{z} - SH(\mathbf{y}))^T R^{-1}(\mathbf{z} - SH(\mathbf{y})) + \frac{1}{2}(u - u^b)^T B^{-1}(u - u^b) + \alpha \sum_{i=1}^{n} H_\gamma(D_i u - w_i) + \beta \sum_{i=1}^{n-1} \frac{1}{2} H_\gamma(E_i w). \tag{3.16}$$

De este modo los problemas γ–regularizados se escriben como sigue:

$$\begin{aligned} &\min_{(\mathbf{y},u)} && J_\gamma(\mathbf{y}, \mathbf{u}) \text{ definida en (3.15)} \\ &\text{sujeto a:} \\ & && e(\mathbf{y}, u) = 0, \end{aligned} \tag{3.17}$$

y

$$\begin{aligned} &\min_{(\mathbf{y},u,w)} && J_\gamma(\mathbf{y}, u, w) \text{ definida en (3.16)} \\ &\text{sujeto a:} \\ & && e(\mathbf{y}, u) = 0. \end{aligned} \tag{3.18}$$

5. Convergencia de las soluciones de los problemas γ–regularizados a las soluciones del problema original

En esta sección vamos a analizar la convergencia de las soluciones del problema γ– regularizado a las soluciones del problema original. Para esto vamos a usar las ideas desarrolladas en [De los Reyes, 2015]. El análisis se lo realiza para la regularización de Huber $\mathcal{C}^2$, sin embargo,

estos resultados pueden ser extendidos a la regularización de Huber $\mathcal{C}^1$ sin ningún problema. Comenzamos esta sección con un Lema que garantiza la convergencia de la función de regularización de Huber $\mathcal{C}^2$ a la función valor absoluto.

Lema 1. *Sea $a \in \mathbb{R}$, entonces*

$$\lim_{\gamma\to\infty} H_\gamma(a) = |a|.$$

Demostración. Notemos primeramente que $H_\gamma(a) \leq |a|$ para todo $a \in \mathbb{R}$. Para poder usar la definición de la regularización de Huber dada en (3.11) analizamos tres casos por separado

Si $\gamma|a| \geq 1 + \dfrac{1}{2\gamma}$:

De la definición dada en (3.11) tenemos

$$\begin{aligned} 0 \leq &\big||a| - H_\gamma(a)\big| = \big||a| - |a| - C_1 - K\big| = |C_1 + K| \\ &= \left|\frac{1}{2\gamma} - \frac{1}{2\gamma^2} + \frac{1}{8\gamma^3} - \frac{1}{\gamma} - \frac{1}{2\gamma^2} + \frac{1}{\gamma^2} - \frac{1}{2\gamma} - \frac{1}{2\gamma^2} - \frac{1}{8\gamma^3}\right. \\ &\left. + \frac{1}{\gamma} + \frac{1}{2\gamma^2} - \frac{1}{2\gamma} - \frac{1}{24\gamma^3}\right| \\ &= \big| - \frac{1}{2\gamma} - \frac{1}{24\gamma^3}\big| = \frac{12\gamma^2+1}{24\gamma^3} \xrightarrow[\gamma\to+\infty]{} 0. \end{aligned}$$

Del teorema del sándwich podemos concluir que

$$H_\gamma(a) \xrightarrow[\gamma\to+\infty]{} |a|.$$

Si $\gamma|a| \leq -\dfrac{1}{2\gamma}$:

Usando nuevamente la definición en (3.11) tenemos que

$$0 \leq \big||a| - H_\gamma(a)\big| =\leq |a| + \frac{\gamma|a|^2}{2}.$$

Puesto que $|a| \leq 1 - \dfrac{1}{2\gamma}$ entonces tenemos

$$\begin{aligned} 0 \leq &\big||a| - H_\gamma(a)\big| \leq \frac{1}{\gamma} - \frac{1}{2\gamma^2} + \frac{1}{2\gamma}\left(1 - \frac{1}{2\gamma}\right)^2, \\ =&\frac{3}{2\gamma} - \frac{1}{\gamma^2} + \frac{1}{8\gamma^3}. \end{aligned}$$

Tomando el límite cuando $\gamma \to +\infty$ y del teorema del sándwich podemos concluir que

$$H_\gamma(a) \xrightarrow[\gamma\to+\infty]{} |a|.$$

Si $|\gamma|a| - 1| \leq {}^1/_{2\gamma}$:

De argumentos similares a los de los casos anteriores se tiene que

$$\begin{aligned}
0 \leq & \left| |a| - G|a| - \frac{H}{2}|a|^2 - \frac{C}{3}|a|^3 - D \right| \leq |a| + G|a| + \frac{H}{2}|a|^2 + \frac{C}{3}|a|^3 + D, \\
\leq & (1+G)\left(1+\frac{1}{2\gamma}\right)\frac{1}{\gamma} + \frac{H}{2\gamma^2}\left(1+\frac{1}{2\gamma}\right)^2 + \frac{C}{3\gamma^3}\left(1+\frac{1}{2\gamma}\right)^3 \\
& + \left(\frac{\gamma}{2} - \frac{H}{2}\right)\frac{1}{\gamma^2}\left(1-\frac{1}{2\gamma}\right)^2 - \frac{A}{\gamma}\left(1-\frac{1}{2\gamma}\right) - \frac{C}{3\gamma^3}\left(1-\frac{1}{2\gamma}\right)^3, \\
= & \frac{3}{2\gamma} + \frac{3}{4\gamma^2} + \frac{1}{12\gamma^3}.
\end{aligned}$$

lo que nos permite concluir que

$$H_\gamma(a) \xrightarrow[\gamma \to +\infty]{} |a|. \qquad \square$$

Para poder probar la convergencia de las soluciones, en primer lugar, vamos a probar la acotación de las variables de estado de los problemas γ–regularizados. Puesto que la función objetivo depende únicamente en unas pocas componentes de $\mathbf{y}$, vamos a usar la definición de la ecuación de estado para garantizar su acotación. Para la demostración del siguiente Lema vamos a usar el resultado conocido como suma por partes el cual enunciamos a continuación.

Lema 2 (Suma por partes)**.** *Sean* $\{a_k\}$ *y* $\{b_k\}$ *dos sucesiones. Entonces*

$$\sum_{k=m}^{n} a_k(b_{k+1} - b_k) = [a_{n+1}b_{n+1} - a_m b_m] - \sum_{k=m}^{n} b_{k+1}(a_{k+1} - a_k)$$

El siguiente Lema muestra el resultado de acotación de la variable de estado.

Lema 3. *Sea* $\mathbf{y}$ *la solución de la ecuación de estado* $e(\mathbf{y}, u) = 0$ *dada en* (3.6)*, entonces se tiene que*

$$\| y^j \| < \| u \| + \Delta t \sum_{i=1}^{j} \| f^i \|, \qquad \forall j = 1, \ldots, N_t.$$

Demostración. Vamos a comenzar por recordar la ecuación de estado $e(\mathbf{y}, u) = 0$ dada por:

$$\mathbb{E}\mathbf{y} + \nu\mathbb{A}\mathbf{y} + \mathbb{Z}(\mathbf{y})\mathbb{U}\mathbf{y} - \mathbf{f}(u) = 0.$$

donde $\mathbf{y} = \begin{bmatrix} y^1 \\ \vdots \\ y^{N_t} \end{bmatrix}$ y $\mathbf{f}(u) = \begin{bmatrix} u/\Delta t \\ \vdots \\ f^{N_t} \end{bmatrix}$.

Entonces, analizando componente a componente en el tiempo

Para $j = 1$**:**

$$\frac{y^1}{\Delta t} - \frac{u}{\Delta t} = 0$$

Para $j = 2, \ldots, N_t$

$$\frac{y^j - y^{j-1}}{\Delta t} + \nu A y^j + \text{diag}(y^{j-1}) U_p y^j - f^j = 0.$$

Vamos a analizar únicamente los términos para $j = 2, \ldots, N_t$. Así, multiplicando por $(y^j)^T$ se tiene que

$$(y^j)^T \left(\frac{y^j - y^{j-1}}{\Delta t} \right) + \nu (y^j)^T A y^j + (y^j)^T \text{diag}(y^{j-1}) U_p y^j - (y^j)^T f^j = 0. \tag{3.19}$$

Analizamos por separado el tercer término de la ecuación anterior. Así, recordando que asumimos que la matriz U_p es la discretización del gradiente, se tiene que

$$(y^j)^T \text{diag}(y^{j-1}) U_p y^j = \sum_{i=1}^{n} y_i^j y_i^{j-1} \left(\frac{y_i^j - y_{i-1}^j}{h} \right).$$

Usando, el Lema 2 se tiene que

$$(y^j)^T \text{diag}(y^{j-1}) U_p y^j = \frac{1}{h} \left[y_n^j y_n^{j-1} y_n^j - y_1^j y_1^{j-1} y_1^j \right] - \sum_{i=1}^{n} y_i^j y_i^{j-1} \left(\frac{y_i^j - y_{i-1}^j}{h} \right).$$

Podemos notar, que el último término del lado derecho es igual a $(y^j)^T \text{diag}(y^{j-1}) U_p y^j$, por tanto

$$(y^j)^T \text{diag}(y^{j-1}) U_p y^j = \frac{1}{2h} \left[y_n^j y_n^{j-1} y_n^j - y_1^j y_1^{j-1} y_1^j \right].$$

Usando las condiciones de borde del problema, las cuales indican que $y_n^j = y_1^j = 0$, se tiene que

$$(y^j)^T \text{diag}(y^{j-1}) U_p y^j = 0.$$

Asi, de (3.19) tenemos que

$$\| y^j \|^2 = (y^j)^T y^{j-1} + \Delta t (y^j)^T f^j - \nu \Delta t (y^j)^T A y^j$$

puesto que A, la matriz de la discretización del Laplaciano, es definida positiva se tiene que $\nu \Delta t (y^j)^T A y^j > 0$. Entonces,

$$\| y^j \|^2 < (y^j)^T y^{j-1} + \Delta t (y^j)^T f^j.$$

Aplicando la desigualdad de Cauchy–Schwarz se tiene que

$$\| y^j \|^2 < \| y^j \| \| y^{j-1} \| + \Delta t \| y^j \| \| f^j \|,$$

y por tanto

$$\| y^j \| < \| y^{j-1} \| + \Delta t \| f^j \|,$$

para todo $j = 2, \ldots, N_t$. Recordemos además, que para $j = 1$ se tiene que

$$\| y^1 \| = \| u \| .$$

Entonces, usando recursivamente este resultado se tiene que

$$\begin{aligned}
\| y^j \| &< \| y^{j-1} \| + \Delta t \| f^j \| \\
&< \| y^{j-2} \| + \Delta t \left(\| f^j \| + \| f^{j-1} \| \right) \\
&\vdots \\
&< \| y^1 \| + \Delta t \sum_{i=1}^{j} \| f^i \| \\
&= \| u \| + \Delta t \sum_{i=1}^{j} \| f^i \| .
\end{aligned}$$

de donde se tiene el resultado. □

El siguiente Teorema nos muestra la convergencia de las soluciones del problema γ– regularizado a las soluciones del problema original para la regularización de variación total.

Teorema 2. *Sea* $(\mathbf{y}_\gamma, u_\gamma)$ *solución del problema* (3.17)*, donde* $\mathbf{y}_\gamma$ *satisface* $e(\mathbf{y}_\gamma, u_\gamma) = 0$*. Sea además* $(\mathbf{y}, u)$ *solución del problema* (3.7)*. Entonces,*

$$(\mathbf{y}_\gamma, u_\gamma) \to (\mathbf{y}, u),$$

para $\gamma \to \infty$*.*

Demostración. Recordemos que la matriz B por definición es una matriz de covarianza simétrica y definida positiva, lo que nos permite utilizar su descomposición espectral y concluir que

$$(u_\gamma - u^b)^T B^{-1} (u_\gamma - u^b) \geq \lambda_{min}(B^{-1}) \| u_\gamma - u^b \|^2, \tag{3.20}$$

donde $\lambda_{min}(B^{-1}) > 0$ es el valor propio más pequeño de B^{-1}. Entonces, de la definición de la función objetivo $J_\gamma(\mathbf{y}, u)$ se tiene que:

$$(u_\gamma - u^b)^T B^{-1} (u_\gamma - u^b) \leq J_\gamma(\mathbf{y}_\gamma, u_\gamma). \tag{3.21}$$

Usando la optimalidad de $(\mathbf{y}_\gamma, u_\gamma)$ se sigue que

$$J_\gamma(\mathbf{y}_\gamma, u_\gamma) \leq J_\gamma(\mathbf{y}, u) \tag{3.22}$$

donde $(\mathbf{y}, u)$ es la solución del problema (3.7). Además, puesto que $H_\gamma(a) \leq |a|$ para todo $a \in \mathbb{R}$ se tiene que

$$J_\gamma(\mathbf{y}, u) \leq J(\mathbf{y}, u). \tag{3.23}$$

Usando, los resultados de (3.20)–(3.23) podemos concluir que:

$$\| u_\gamma - u^b \| \leq \sqrt{\frac{J(\mathbf{y}, u)}{\lambda_{min}(B^{-1})}}.$$

Finalmente de la desigualdad triangular inversa se tiene que

$$\| u_\gamma \| \leq \sqrt{\frac{J(\mathbf{y}, u)}{\lambda_{min}(B^{-1})}} + \| u^b \| =: \eta. \tag{3.24}$$

la cual es una constante independiente de γ y la notaremos por η. Este resultado nos permite concluir que la sucesión $\{u_\gamma\}$ es uniformemente acotada con respecto a γ. Usando el Teorema de Bolzano–Weierstrass sabemos que existe una subsucesión a la que notaremos por $\{u_\gamma\}$ convergente en $\mathbb{R}^n$ y a su límite lo notaremos por $\bar{u}$.

Por otro lado, usando el Lema 3 tenemos que

$$\begin{aligned}
\| \mathbf{y}_\gamma \|^2 &= \sum_{j=1}^{N_t} \| y_\gamma^j \|^2 \\
&< \sum_{j=1}^{N_t} (\| u_\gamma \| + \Delta t \sum_{i=1}^{j} \| f^i \|)^2 \\
&= \sum_{j=1}^{N_t} \left[\| u_\gamma \|^2 + 2\Delta t \| u_\gamma \| \left(\sum_{i=1}^{j} \| f^i \| \right) + \Delta t^2 \left(\sum_{i=1}^{j} \| f^i \| \right)^2 \right].
\end{aligned}$$

Usando (3.24) podemos concluir que

$$\| \mathbf{y}_\gamma \|^2 \leq N_t \eta^2 + 2\eta \Delta t \left(\sum_{j=1}^{N_t} \sum_{i=1}^{j} \| f^i \| \right) + \Delta t^2 \sum_{j=1}^{N_t} \left(\sum_{i=1}^{j} \| f^i \| \right)^2.$$

la cual es independiente de γ y por tanto la sucesión $\{\mathbf{y}_\gamma\}$ es uniformemente acotada con respecto a γ. Así, usando nuevamente el Teorema de Bolzano–Weierstrass sabemos que existe una subsucesión a la cual notaremos por $\{\mathbf{y}_\gamma\}$ convergente en $\mathbb{R}^m$ y a su límite lo notaremos por $\bar{\mathbf{y}}$.

Del análisis anterior hemos construido un candidato a solución del problema (3.7). Así, en primer lugar debemos probar que este candidato $(\bar{\mathbf{y}}, \bar{u})$ es factible, es decir, que $e(\bar{\mathbf{y}}, \bar{u}) = 0$. Usando la definición de la ecuación de estado dada en (3.6) tenemos que:

$$e(\mathbf{y}(u_\gamma), u_\gamma) = \mathbb{E}\mathbf{y}(u_\gamma) + \nu \mathbb{A}\mathbf{y}(u_\gamma) + \mathbb{Z}_\gamma \mathbb{U}\mathbf{y}(u_\gamma) - \mathbf{f}(u_\gamma) = 0,$$

donde $\mathbb{Z}_\gamma = \mathbb{Z}(\mathbf{y}(u_\gamma))$. Entonces, tomando el límite cuando $\gamma \to \infty$ se tiene que

$$\mathbb{E}\bar{\mathbf{y}} + \nu \mathbb{A}\bar{\mathbf{y}} + \lim_{\gamma \to \infty} (\mathbb{Z}_\gamma \mathbb{U}\mathbf{y}(u_\gamma)) - \mathbf{f}(\bar{u}) = 0,$$

de este modo, para poder garantizar que $\bar{\mathbf{y}}$ satisface la ecuación de estado es suficiente probar que

$$\lim_{\gamma\to\infty} \mathbb{Z}_\gamma \mathbb{U}\mathbf{y}(u_\gamma) = \overline{\mathbb{Z}}\mathbb{U}\bar{\mathbf{y}},$$

donde $\overline{\mathbb{Z}} = \mathbb{Z}(\bar{\mathbf{y}})$.

Usando la definición de las matrices $\mathbb{Z}$, $\mathbb{U}$ y analizando componente a componente en el tiempo el sistema tenemos que calcular

$$\lim_{\gamma\to\infty} \operatorname{diag}(y^{j-1}(u_\gamma))U_p y^j(u_\gamma),$$

para $j = 2, \ldots, N_t$. Realizando un análisis componente a componente en el espacio y recordando que la matriz U_p representa el gradiente discreto tenemos que calcular el siguiente límite:

$$\lim_{\gamma\to\infty} (y_i^{j-1})(u_\gamma) \left(\frac{(y_i^j)(u_\gamma) - (y_{i-1}^j)(u_\gamma)}{h} \right),$$

para $i = 1 \ldots, n$. Entonces, puesto que $\mathbf{y}(u_\gamma) = \mathbf{y}_\gamma \to \bar{\mathbf{y}}$ para $\gamma \to \infty$, tenemos que

$$(y_i^{j-1})_\gamma \to \bar{y}_i^{j-1}$$

y

$$\left(\frac{(y_i^j)(u_\gamma) - (y_{i-1}^j)(u_\gamma)}{h} \right) \to \left(\frac{\bar{y}_i^j - \bar{y}_{i-1}^j}{h} \right).$$

Por tanto, usando la propiedad del producto de límites se sigue que

$$(y_i^{j-1}) \left(\frac{(y_i^j)(u_\gamma) - (y_{i-1}^j)(u_\gamma)}{h} \right) \to \bar{y}_i^{j-1} \left(\frac{\bar{y}_i^j - \bar{y}_{i-1}^j}{h} \right).$$

El resultado anterior se lo tiene para todo $i = 1 \ldots, n$. Así, nuevamente puesto que la matriz U_p es el gradiente discreto se tiene que

$$\lim_{\gamma\to\infty} \operatorname{diag}(y^{j-1}(u_\gamma))U_p y^j(u_\gamma) = \operatorname{diag}(\bar{y}^{j-1})U_p \bar{y}^j,$$

para todo $j = 2, \ldots, N_t$. Finalmente para $j = 1$ sabemos que las matrices son iguales a cero por tanto se tiene el resultado automáticamente y esto nos permite concluir que

$$\lim_{\gamma\to\infty} \mathbb{Z}_\gamma \mathbb{U}\mathbf{y}(u_\gamma) = \overline{\mathbb{Z}}\mathbb{U}\bar{\mathbf{y}}.$$

Consecuentemente,

$$\mathbb{E}\bar{\mathbf{y}} + \nu\mathbb{A}\bar{\mathbf{y}} + \overline{\mathbb{Z}}\mathbb{U}\bar{\mathbf{y}} - \mathbf{f}(\bar{u}) = 0,$$

es decir, $e(\bar{\mathbf{y}}, \bar{u}) = 0$.

Finalmente, del Lema 1 sabemos que $H_\gamma(a) \to |a|$ para $\gamma \to \infty$, por lo tanto

$$J(\bar{\mathbf{y}}, \bar{u}) = \lim_{\gamma \to \infty} J_\gamma(\mathbf{y}_\gamma, u_\gamma) \leq J(\mathbf{y}, u).$$

Por otro lado, puesto que $(\mathbf{y}, u)$ es una solución global del problema (3.7) podemos concluir que $(\bar{\mathbf{y}}, \bar{u})$ también es una solución global del mismo y por tanto se obtiene el resultado. □.

El resultado de convergencia de las soluciones del problema γ–regularizado con la regularización de variación total generalizada puede ser demostrado usando las ideas de la demostración anterior.

6. Sistema de optimalidad

En esta sección vamos a derivar el sistema de optimalidad de los problemas γ– regularizados. Además, como se mencionó anteriormente asumiremos que la matriz de discretización $\mathbb{U}$ no depende de $\mathbf{y}$, es decir, es una matriz constante cuyas matrices asociadas U_p representan el gradiente discreto.

6.1. Derivadas de $e(\mathbf{y}, u)$

En los siguientes párrafos vamos a presentar el cálculo de las derivadas de la ecuación de estado para poder caracterizar el sistema de optimalidad del problema de asimilación de datos.

Lema 4. *Sea $e(\mathbf{y}, u)$ dada en* (3.6)*, entonces las siguientes igualdades se satisfacen*

$$e_y(\mathbf{y}, u)^T\mathbf{p} = \mathbb{E}^T\mathbf{p} + \nu\mathbb{A}^T\mathbf{p} + \mathbb{Z}\mathbb{U}^T\mathbf{p} + diag(\mathbb{U}\mathbf{y})\mathbf{p},$$
$$e_u(\mathbf{y}, u)^T\mathbf{p} = (-p^1/\Delta t, 0, 0, \cdots, 0)^T.$$

Demostración. Para la primera igualdad vamos a calcular la derivada direccional en la dirección $\mathbf{p}$. Así,

$$e_y(\mathbf{y}, u)^T\mathbf{p} = \mathbb{E}^T\mathbf{p} + \mathbb{A}^T\mathbf{p} + (\mathbb{Z}\mathbb{U}\mathbf{y})'(\mathbf{p}).$$

Usando la regla de la derivada del producto, tenemos:

$$e_y(\mathbf{y}, u)^T\mathbf{p} = \mathbb{E}^T\mathbf{p} + \mathbb{A}^T\mathbf{p} + \mathbb{Z}\mathbb{U}^T\mathbf{p} + \text{diag}\,(\mathbb{U}\mathbf{y})\mathbf{p},$$

de donde se tiene el resultado.

Para la segunda igualdad, puesto que $e_u(\mathbf{y}, u) : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^{m\times n}$ y solo el primer término

depende de u se tiene:

$$e_u(\mathbf{y}, u) = \begin{bmatrix} -\frac{1}{\Delta t}\mathbb{I} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Por esta razón evaluando esta derivada en la dirección

$$\mathbf{p} = \begin{bmatrix} p^1 \\ p^2 \\ \vdots \\ p^{N_t} \end{bmatrix}$$

donde $p^i \in \mathbb{R}^n$. Entonces se tiene el resultado. □

6.2. Condiciones de optimalidad

Esta sección está dedicada a mostrar los teoremas que nos permiten enunciar los sistemas de optimalidad para los problemas regularizados TV y TGV usando la regularización de Huber $\mathcal{C}^1$; sin embargo, los sistemas de optimalidad para los problemas TV y TGV con la regularización de Huber $\mathcal{C}^2$ pueden ser obtenidos usando las mismas herramientas.

Teorema 3. *El sistema de optimalidad del problema* (3.17) *está dado por la siguiente expresión:*

$$\begin{cases} \mathbb{E}\mathbf{y} + \nu\mathbb{A}\mathbf{y} + \mathbb{Z}\mathbb{U}\mathbf{y} - \mathbf{f}(u) = 0 & \textit{(Ecuación de estado)} \\ \mathbb{E}^T\mathbf{p} + \nu\mathbb{A}^T\mathbf{p} + \mathbb{Z}\mathbb{U}^T\mathbf{p} + diag(\mathbb{U}\mathbf{y})\mathbf{p} - H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z}) = 0 & \textit{(Ecuación adjunta)} \\ B^{-1}(u - u^b) + \beta D^T\left(\frac{\gamma Du}{\max\{\gamma|Du|, 1\}}\right) + \frac{1}{\Delta t}p^1 = 0 & \textit{(Ecuación del gradiente)} \end{cases}$$

donde $\mathbf{p} = [p^1, p^2, \ldots, p^{N_t}]^T \in \mathbb{R}^m$, *con* $p^i \in \mathbb{R}^n$ *para cada* $i = 1, \ldots, N_t$.

Demostración. Para la caracterización del sistema de optimalidad de este problema usamos el Teorema 3.3 de [De los Reyes, 2015] el cual nos dice que:

Sea (y^*, u^*) una solución óptima de (3.17). Si

$$e_y(y^*, u^*)$$

es biyectiva, entonces existe un estado adjunto p tal que:

$$\begin{cases} e(y^*, u^*) = 0, \\ e_y(y^*, u^*)^T p = \nabla_y J(y^*, u^*), \\ e_u(y^*, u^*)^T p = \nabla_u J(y^*, u^*). \end{cases}$$

Por el momento se asume la biyectividad de $e_y(y^*, u^*)$. Este tema será abordado al final del capítulo. De este modo, la demostración se reduce al cálculo de las derivadas de $J_\gamma(\mathbf{y}, u)$ y $e(\mathbf{y}, u)$.Usando las reglas de la derivada del producto y la derivada de la función de Huber se satisfacen las siguientes igualdades:

$$\begin{aligned} \nabla_y J_\gamma(\mathbf{y}, u) &= H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z}), \\ \nabla_u J_\gamma(\mathbf{y}, u) &= B^{-1}(u - u^b) + \beta D^T \left(\frac{\gamma Du}{\text{máx}\{\gamma |Du|, 1\}} \right), \end{aligned}$$

donde las operaciones de división y máx son componente a componente. Finalmente, las derivadas de la ecuación de estado fueron calculadas en el Lema 4. Combinando estos resultados se concluye la demostración. □

A continuación enunciamos el Teorema que caracteriza el sistema de optimalidad para el problema (3.18) con la regularización de variación total generalizada.

Teorema 4. *El sistema de optimalidad asociado al problema* (3.18) *está dado por la siguiente expresión:*

$$\begin{cases} \mathbb{E}\mathbf{y} + \nu \mathbb{A}\mathbf{y} + \mathbb{Z}\mathbb{U}\mathbf{y} - \mathbf{f} = 0 \\ \\ \mathbb{E}^T\mathbf{p} + \nu \mathbb{A}^T\mathbf{p} + \mathbb{Z}\mathbb{U}^T\mathbf{p} + diag(\mathbb{U}\mathbf{y})\mathbf{p} - H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z}) = 0 \\ \\ B^{-1}(u - u^b) + \alpha D^T \left(\dfrac{\gamma\,(Du - w)}{\text{máx}\{\gamma |Du - w|, 1\}} \right) + \dfrac{1}{\Delta t} p^1 = 0 \\ \\ -\alpha \left(\dfrac{\gamma\,(Du - w)}{\text{máx}\{\gamma |Du - w|, 1\}} \right) + \beta E^T \left(\dfrac{\gamma Ew}{\text{máx}\{\gamma |Ew|, 1\}} \right) = 0 \end{cases}$$

Demostración. De las ideas desarrolladas en la demostración del Teorema 3 basta calcular las derivadas de la función objetivo. Usando las reglas de derivación del producto y la derivada de la regularización de Huber $\mathcal{C}^1$ dada en (3.12) tenemos que:

$$\begin{aligned} \nabla_y J_\gamma(\mathbf{y}, u, w) &= H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z}), \\ \nabla_u J_\gamma(\mathbf{y}, u, w) &= \alpha D^T \left(\frac{\gamma(Du - hw)}{\text{máx}\{\gamma |Du - hw|, 1\}} \right), \\ \nabla_w J_\gamma(\mathbf{y}, u, w) &= -\alpha \left(\frac{\gamma(Du - hw)}{\text{máx}\{\gamma |Du - hw|, 1\}} \right) + \beta E^T \left(\frac{\gamma Ew}{\text{máx}\{\gamma |Ew|, 1\}} \right). \end{aligned}$$

Usando nuevamente el Teorema 3.3 de [De los Reyes, 2015] y combinando este resultado con el Lema 4 se obtiene el resultado. □

6.3. Estado adjunto

En esta sección vamos a discutir sobre las principales características de la ecuación adjunta dada por la expresión:

$$\mathbb{E}^T\mathbf{p} + \nu\mathbb{A}^T\mathbf{p} + \mathbb{Z}\mathbb{U}^T\mathbf{p} + \text{diag}(\mathbb{U}\mathbf{y})\mathbf{p} - H^TS^TR^{-1}(SH\mathbf{y} - \mathbf{z}) = 0 \tag{3.25}$$

El Teorema 3.3 de [De los Reyes, 2015] asume la invertibilidad de la derivada de la ecuación de estado con respecto a y, es decir, de $e_y(\mathbf{y}, u)$. En este caso el análisis de existencia y unicidad de las soluciones de la ecuación adjunta que se va a realizar a continuación es equivalente a mostrar la invertibilidad de $e_y(\mathbf{y}, u)$ es por esta razón de vital importancia su análisis.

En primer lugar, notamos que a diferencia de los resultados en dimensión infinita en los que se puede garantizar que la condición final del estado adjunto es igual a cero; en este caso, la condición final es distinta a cero y podemos además, dar una expresión explícita de dicha condición. Analizando componente a componente en el tiempo el sistema correspondiente a la ecuación adjunta dado en la ecuación (3.25), tenemos

Para $j = N_t$:

$$p^{N_t} = \left({}^1\!/_{\Delta t}\mathbb{I} + \nu A^T + \text{diag}(y^{N_t-1})U_p^T + \text{diag}(U_p y^{N_t})\right)^{-1}\left[H^TS^TR^{-1}(SH\mathbf{y} - \mathbf{z})\right]_{N_t},$$

donde $\left[H^TS^TR^{-1}(SH\mathbf{y} - \mathbf{z})\right]_j$ son las j–ésimas n entradas del vector $H^TS^TR^{-1}(SH\mathbf{y} - \mathbf{z})$.

Para $j = 2, \ldots, N_t - 1$:

$$p^j = \left({}^1\!/_{\Delta t}\mathbb{I} + \nu A^T + \text{diag}(y^{j-1})U_p^T + \text{diag}(U_p y^j)\right)^{-1}\left({}^1\!/_{\Delta t}p^{j+1} + \left[H^TS^TR^{-1}(SH\mathbf{y} - \mathbf{z})\right]_j\right) \tag{3.26}$$

Para $j = 1$:

$$(\mathbb{I} + \Delta t\text{diag}\,(U_p y^1))p^1 = p^2 + \Delta t\left[H^TS^TR^{-1}(SH\mathbf{y} - \mathbf{z})\right]_1.$$

Para garantizar la existencia de soluciones de la ecuación adjunta vamos a usar los resultados de ecuaciones en diferencias. En particular usamos el siguiente Teorema tomado de [Elaydi, 2005].

Teorema 5. *El problema a valores iniciales*

$$\begin{cases} y(k+n) + p_1(n)y(k+n-1) + \ldots + p_k(n)y(n) = g(n) \\ y(n_0) = a_0 \\ y(n_0+1) = a_1 \\ \vdots \\ y(n_0+k+1) = a_{k-1} \end{cases}$$

tiene una única solución $y(n)$.

Ya que en nuestro caso la ecuación adjunta está definida hacia atrás vamos a realizar en primer lugar el siguiente cambio de variable

$$q^j = p^{N_t - j + 1}.$$

Usando las expresiones mencionadas anteriormente tenemos que

$$q^1 = \left(\frac{1}{\Delta t}\mathbb{I} + \nu A^T + \text{diag}\,(y^{N_t-1})U_p^T y^{N_t} + \text{diag}\,(U_p y^{N_t}) \right)^{-1} [H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z})]_{N_t} =: a_1$$

y

$$q^{n+1} + \alpha(n)q^n = \varphi(n), \qquad \forall n$$

con

$$\alpha(n) = \frac{1}{\Delta t}\left(\frac{1}{\Delta t} + \nu A^T \text{diag}\,(y^n)U_p^T + \text{diag}\,(U_p y^{n+1}) \right)^{-1}$$

y

$$\varphi(n) = \left({}^1\!/_{\Delta t}\mathbb{I} + \nu A^T + \text{diag}(y^n)U_p^T + \text{diag}(U_p y^{n+1}) \right)^{-1} \left[H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z}) \right]_{n+1}.$$

De las definiciones anteriores tenemos un polinomio característico que es no homogéneo y no lineal. Usando Teorema 5 podemos concluir que el sistema anterior tiene solución única q^n, para cualquier $n = 1, \ldots$ y por tanto existe solución única $p^{N_t - n + 1}$ para todo $n = 1, \ldots$.

De esta manera hemos mostrado la existencia y unicidad de las soluciones de la ecuación adjunta para cualquier n. De este resultado podemos además concluir que la matriz asociada a los sistemas lineales obtenidos al analizar componente a componente en el tiempo de la ecuación adjunta es invertible y la presentamos a través del siguiente resultado.

Corolario 1. *La matriz*

$$\frac{1}{\Delta t}\mathbb{I} + \nu A^T + diag\,(y^{j-1})U_P^T + diag\,(U_p y^j)$$

es invertible para todo $j = 2, \ldots, N_t$.

Capítulo 4

Métodos numéricos para la solución del problema

La solución numérica de los problemas de asimilación de datos regularizados puede ser realizada utilizando el método de Newton para resolver el sistema de optimalidad, o usando algoritmos iterativos para la solución del problema de optimización. En este trabajo nos concentramos en el estudio de los algoritmos iterativos para la resolución de problemas de optimización no lineal aplicados al problema de asimilación de datos variacional sujeto a la ecuación de Burgers. A lo largo de este capítulo vamos a discutir sobre los algoritmos iterativos de optimización como el método del descenso más profundo, el método BFGS y un método de Newton globalizado para problemas de optimización. El objetivo principal de proponer y estudiar estos tres algoritmos es encontrar numéricamente el mejor algoritmo para la resolución de problema en términos del número de iteraciones y de la reconstrucción de las soluciones. Para los tres métodos mencionados anteriormente estudiamos resultados teóricos que nos garanticen la convergencia del método a puntos estacionarios del problema.

La naturaleza del problema de asimilación de datos con las regularizaciones de variación total y variación total generalizada genera inconvenientes al momento de escoger el paso de descenso con el algoritmo usual del *backtracking*. Es por esta razón que en este trabajo se estudió también el uso de nuevos algoritmos que permitan la búsqueda del paso de descenso. En particular, estudiamos el algoritmo de búsqueda lineal polinomial.

1. Resultados preliminares

Esta sección está dedicada a mostrar resultados preliminares que nos permitirán demostrar la convergencia de los algoritmos presentados en las siguientes secciones.

Comenzamos esta sección presentado un resultado de acotación del estado adjunto. Este resultado es análogo al mostrado en [Volkwein, 1997] Lema 3.4, página 83. La demostración de

este resultado será realizada usando el Teorema de Gronwall discreto [Quarteroni et al., 2010] el cual se presenta a continuación

Lema 5 (Lemma de Gronwall discreto)**.** *Sea k_n una sucesión no negativa y φ_n una sucesión tal que*

$$\begin{cases} \varphi_0 \leq g_0, \\ \varphi_n \leq g_0 + \sum_{s_0}^{n-1} p_s + \sum_{s=0}^{n-1} k_s \varphi_s, \qquad n \geq 1. \end{cases}$$

Si $g_0 \geq 0$ y $p_n \geq 0$ para cualquier $n \geq 0$, entonces

$$\varphi_n \leq \left(g_0 + \sum_{s_0}^{n-1} p_s \right) \exp\left(\sum_{s=0}^{n-1} k_s \right), \qquad n \geq 1.$$

A continuación presentamos el Lema de acotación de la variable adjunta.

Lema 6. *El estado adjunto $\mathbf{p}$ asociado a las variables $(\mathbf{y}, u)$ del problema* (3.17) *satisface*

$$\| \mathbf{p} \|_\infty \leq \rho \, \| H^T S^T R^{-1} (SH\mathbf{y} - \mathbf{z}) \|_\infty,$$

con $\rho = \sum_{i=1}^{N_t} \| G_j^{-1} \|_\infty \exp\left(\sum_{i=1}^{N_t} \frac{\| G_j^{-1} \|_\infty}{\Delta t} \right)$, donde la matriz G_j está definida por

$$\begin{aligned} G_j &= \frac{1}{\Delta t}\mathbb{I} + \nu A^T + diag(y^{N_t - j}) U_p^T + diag\,(U_p y^{N_t - j + 1}), \qquad j = 1, \ldots, N_t - 1 \\ G_{N_t} &= \frac{1}{\Delta t}\mathbb{I} + diag(U_p y^1). \end{aligned}$$

Demostración. Comenzamos por recordar que la ecuación adjunta está dada por la siguiente expresión

$$\mathbb{E}^T \mathbf{p} + \nu \mathbb{A}^T \mathbf{p} + \mathbb{Z}\mathbb{U}^T \mathbf{p} + \text{diag}\,(\mathbb{U}\mathbf{y})\mathbf{p} - H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z}) = 0.$$

Si definimos,

$$p^j = q^{N_t - j + 1},$$

entonces, $\mathbf{p} = \mathbb{P}\mathbf{q}$ donde $\mathbb{P}$ es una matriz de permutación de la forma

$$\mathbb{P} = \begin{bmatrix} 0 & \ldots & 0 & 0 & \mathbb{I} \\ 0 & \ldots & 0 & \mathbb{I} & 0 \\ 0 & \ldots & \mathbb{I} & 0 & 0 \\ \vdots & \iddots & \vdots & \vdots & \vdots \\ \mathbb{I} & \ldots & 0 & 0 & 0 \end{bmatrix}. \tag{4.1}$$

Así, obtenemos el sistema

$$\mathbb{E}^T \mathbb{P}\mathbf{q} + \nu \mathbb{A}^T \mathbb{P}\mathbf{q} + \mathbb{Z}\mathbb{U}^T \mathbb{P}\mathbf{q} + \text{diag}\,(\mathbb{U}\mathbf{y})\mathbb{P}\mathbf{q} - H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z}) = 0.$$

Analizando el sistema anterior componente a componente en el tiempo se tienen tres casos diferentes

Para $j = 1$:

$$\left(\frac{1}{\Delta t}\mathbb{I} + \text{diag}\,(U_p y^1)\right) q^{N_t} = \left[H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z})\right]_1 + \frac{1}{\Delta t} q^{N_t - 1},$$

donde $\left[H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z})\right]_j$ son las j–ésimas n entradas del vector $H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z})$.

Para $j = 2, \ldots, N_t - 1$:

$$\left(\frac{1}{\Delta t}\mathbb{I} + \nu A^T + \text{diag}(y^{j-1}) U_p^T + \text{diag}\,(U_p y^j)\right) q^{N_t - j + 1} = \left[H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z})\right]_j + \frac{1}{\Delta t} q^{N_t - j}.$$

Para $j = N_t$:

$$\left(\frac{1}{\Delta t}\mathbb{I} + \nu A^T + \text{diag}(y^{N_t - 1}) U_p^T + \text{diag}\,(U_p y^{N_t})\right) q^1 = \left[H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z})\right]_{N_t}.$$

Definimos entonces

$$G_j = \frac{1}{\Delta t}\mathbb{I} + \nu A^T + \text{diag}(y^{N_t - j}) U_p^T + \text{diag}\,(U_p y^{N_t - j + 1}), \qquad j = 1, \ldots, N_t - 1,$$
$$G_{N_t} = \frac{1}{\Delta t}\mathbb{I} + \text{diag}(U_p y^1).$$

Así,

$$G_1 q^1 = \left[H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z})\right]_{N_t},$$
$$G_j q^j = \left[H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z})\right]_j + \frac{1}{\Delta t} q^{j-1}, \quad j = 2, \ldots, N_t$$

Podemos notar que G_j corresponde a la matriz descrita en el Corolario 1 cuyo enunciado garantiza que esta matriz es invertible para todo $j = 1, \ldots, N_t$. Por tanto, despejando la variable q^j, tomando normas a ambos lados de la desigualdad y aplicando la definición de norma matricial y la desigualdad triangular obtenemos

$$\| q^1 \|_\infty \leq \| G_1^{-1} \|_\infty \left\| \left[H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z})\right]_{N_t} \right\|_\infty,$$
$$\| q^j \|_\infty \leq \| G_j^{-1} \|_\infty \left\| \left[H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z})\right]_j \right\|_\infty + \frac{1}{\Delta t} \| G_j^{-1} \|_\infty \| q^{j-1} \|_\infty, \qquad j = 2, \ldots, N_t.$$

Aplicando el Lema 5 tenemos que

$$\| q^j \|_\infty \leq \| G_j^{-1} \|_\infty \exp\left(\frac{\| G_j^{-1} \|_\infty}{\Delta t}\right) \left\| \left[H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z})\right]_j \right\|_\infty \quad \forall j = 1, \ldots, N_t.$$

Puesto que todos los términos son positivos tenemos que

$$\begin{aligned}\| q^j \|_\infty &\leq \sum_{j=1}^{N_t} \| G_j^{-1} \| \exp\left(\frac{\| G_j^{-1} \|_\infty}{\Delta t}\right) \left\| \left[H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z})\right]_j \right\|_\infty, \\ &\leq \| H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z}) \|_\infty \sum_{j=1}^{N_t} \| G_j^{-1} \|_\infty \exp\left(\frac{\| G_j^{-1} \|_\infty}{\Delta t}\right).\end{aligned}$$

Así de la definición de la norma $\| \cdot \|_\infty$ podemos concluir que

$$\| \mathbf{q} \|_\infty \leq \rho \| H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z}) \|_\infty,$$

donde $\rho = \sum_{j=1}^{N_t} \| G_j^{-1} \|_\infty \exp\left(\frac{\| G_j^{-1} \|_\infty}{\Delta t}\right)$. Finalizamos la demostración recordando que $\mathbf{q} = \mathbb{P}^{-1}\mathbf{p}$ y puesto que $\mathbb{P}$ es una matriz de permutación se tiene que

$$\| \mathbf{p} \|_\infty \leq \rho \| H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z}) \|_\infty,$$

□

Los resultados de convergencia de los algoritmos que se muestran a continuación asumen el cumplimiento de las condiciones de factibilidad de la búsqueda lineal las cuales están dadas por las siguientes expresiones

$$f(u^k + s_k d^k) < f(u^k), \quad \forall k = 1, 2, \ldots, \tag{4.2}$$

$$f(u^k + s_k d^k) - f(u^k) \xrightarrow{k\to\infty} 0 \Rightarrow \frac{(\nabla f(u^k)^T d^k)}{\| d^k \|} \xrightarrow{k\to\infty} 0. \tag{4.3}$$

También se asume como verdadera la condición del ángulo dada por:

$$-(\nabla f(u^k)^T) d^k \geq \eta \| \nabla f(u^k) \| \| d^k \| \tag{4.4}$$

para algún $\eta \in (0, 1)$.

La demostración de convergencia de los algoritmos será realizada aplicando el siguiente resultado tomado de [De los Reyes, 2015].

Teorema 6. *Sea f continuamente diferenciable y acotada inferiormente. Sea $\{u^k\}$, $\{d^k\}$ y $\{s_k\}$ sucesiones generadas por un algoritmo de descenso que satisfacen las condiciones de factibilidad de la búsqueda*

lineal y la condición del ángulo. Entonces

$$\lim_{k\to\infty} \nabla f(u^k) = 0,$$

y todo punto de acumulación de $\{u^k\}$ *es un punto estacionario de* f.

La demostración de la condición (4.4) dependerá directamente de la definición de la dirección de descenso para cada algoritmo. Por otro lado, la demostración de que los pasos de descenso $\{s_k\}$ satisfacen las condiciones (4.2) y (4.3) se realizará usando el siguiente resultado que consiste en una versión modificada al presentado en [De los Reyes, 2015] y la proposición 9.1 página 36 de [Ulbrich and Ulbrich, 2012]:

Teorema 7. *Sea* ∇f *uniformemente continuo sobre el conjunto de nivel*

$$N_0^o := \left\{u + d \colon f(u) \leq f(u^0), \parallel d \parallel \leq \rho\right\}, \tag{4.5}$$

para algún $\rho > 0$. *Si las iteraciones generadas por el método de descenso, con* $\{s_k\}$ *que satisface la regla de Armijo. Además, si existe una función* $\varphi : [0, +\infty) \to [0, +\infty)$ *monótona creciente tal que las direcciones generadas* $\{d^k\}$ *satisfacen*

$$\parallel d^k \parallel \geq \varphi\left(\frac{-\nabla f(u^k)^T d^k}{\parallel d^k \parallel}\right). \tag{4.6}$$

entonces $\{s_k\}$ *satisface las condiciones de factibilidad de la búsqueda lineal.*

Para la demostración de este resultado basta mostrar que las direcciones $\{d_k\}$ satisfacen (4.6), lo cual se hace a partir de la definición de cada algoritmo. Además, es necesario mostrar que ∇f es uniformemente continuo donde f corresponde al funcional reducido dado por $f(u) = J_\gamma(\mathbf{y}(u), u)$. Puesto que ∇f no depende del algoritmo que se escoja vamos a demostrar a continuación este resultado, el cual será aplicando posteriormente para la demostración de convergencia en cada algoritmo. Este resultado nos permite concluir además que la función objetivo es continuamente diferenciable, supuesto necesario para demostración de la convergencia de los algoritmos.

En primer lugar, comenzamos por mostrar un resultado que nos garantiza que la función objetivo reducida $f(u) = J_\gamma(\mathbf{y}(u), u)$ es radialmente no acotada. Este resultado será aplicado más adelante para mostrar que el conjunto de nivel dado en (4.5) es compacto.

Lema 7. *Sea* $f(u) = J_\gamma(\mathbf{y}(u), u)$ *definida en* (3.15) *es radialmente no acotada. Es decir,*

$$f(u) \to +\infty, \qquad \textit{cuando } \parallel u \parallel \to +\infty.$$

Demostración. Usando la definición de la función reducida $f(u)$ tenemos que:

$$\begin{aligned} f(u) &= \frac{1}{2}(\mathbf{z} - SH\mathbf{y}(u))^T R^{-1}(\mathbf{z} - SH\mathbf{y}(u)) + \frac{1}{2}(u - u^b)^T B^{-1}(u - u^b) + \beta \sum_{i=1}^{n} H_\gamma(D_i u) \\ &\geq \frac{1}{2}(u - u^b)^T B^{-1}(u - u^b) \\ &\geq \frac{\lambda_{min}(B^{-1})}{2} \parallel u - u^b \parallel^2 \end{aligned}$$

donde $\lambda_{min}(B^{-1}) > 0$ puesto que B es una matriz de covarianza simétrica y definida positiva. Entonces, tomando u tal que $\parallel u \parallel \to +\infty$ se tiene que:

$$\parallel u - u^b \parallel^2 \to +\infty.$$

Por tanto, podemos concluir que

$$f(u) \to +\infty.$$ □

De la misma manera podemos mostrar que la función objetivo asociada al problema con la regularización de variación total generalizada es radialmente no acotada.

A continuación mostramos un resultado que nos garantiza que el conjunto de nivel es compacto. Este resultado nos será de gran ayuda al momento de demostrar que el gradiente del funcional reducido es uniformemente continuo.

Lema 8. *El conjunto de nivel N_0^ρ dado en* (4.5) *es compacto.*

Demostración. Puesto que estamos trabajando en espacios de dimensión finita para mostrar que N_0^ρ es compacto es suficiente mostrar que es cerrado y acotado.

N_0^ρ **es acotado** :

Procedemos por reducción al absurdo, entonces suponemos que N_0^ρ no es acotado, es decir que existe $u_k + d \in N_0^\rho$ tal que

$$\parallel u_k + d \parallel \to +\infty.$$

Ademas, ya que $\parallel d \parallel \leq \rho$ se tiene que

$$\parallel u_k \parallel \to +\infty.$$

Sin embargo, puesto que $u_k + d \in N_0^\rho$ para todo k se tiene que

$$f(u_k) \leq f(u^0),$$

es decir, la sucesión $\{f(u_k)\}$ es acotada. Por otro lado, $f(u)$ del Lema 7 es radialmente no acotada, entonces

$$f(u_k) \to +\infty.$$

Lo que produce una contradicción, por tanto, N_0^ρ es acotado.

N_0^ρ **es cerrado** :

En este caso procedemos por la caracterización de sucesiones de un conjunto cerrado, es decir, tomamos una sucesión $\{u_k + d\} \subset N_0^\rho$ convergente tal que $u_k + d \to u + d$. Entonces, vamos a probar que $u + d \in N_0^\rho$. Así, usando la definición del conjunto N_0^ρ tenemos que

$$f(u_k) \leq f(u^0).$$

Por otro lado, de la continuidad de la función $f(\cdot)$ sabemos que:

$$f(u_k) \to f(u).$$

Entonces, se puede concluir que

$$f(u) \leq f(u^0),$$

y por tanto $u + d \in N_0^\rho$, ya que $\| d \| < \rho$ para algún $\rho > 0$. □

Finalmente, con este resultado procedemos a la demostración de la continuidad uniforme del gradiente del funcional reducido. Debido a que la ecuación de estado es no lineal no existe una expresión explícita para el funcional reducido, por esta razón vamos a utilizar la definición del gradiente del funcional reducido en términos del estado adjunto la cual está dada por la siguiente expresión:

$$\begin{aligned} \nabla f(u) &= -e_u(\mathbf{y}, u)^T p + \nabla_u J(\mathbf{y}, u) \\ &= \left(\frac{1}{\Delta t}\right) p^1 + B^{-1}(u - u^b) + \beta D^T h_\gamma(Du), \end{aligned} \tag{4.7}$$

donde $h_\gamma(\cdot)$ es la derivada de la regularización de Huber dependiendo del algoritmo que se utilice. Los algoritmos, del descenso más profundo y BFGS utilizan la regularización de Huber $\mathcal{C}^1$ por tanto $h_\gamma(\cdot)$ está dada por (3.12). Mientras para el método de Newton globalizado se utiliza la derivada dada en (3.13).

Para a demostración de la continuidad uniforme del gradiente vamos a usar el Lema de Banach de operadores inversos dado en [Ulbrich and Ulbrich, 2012] cuyo enunciado se muestra a continuación

Lema 9 (Lemma de Banach de operadores inversos)**.** *El conjunto $\mathcal{M} \subset \mathbb{R}^{n\times n}$ de las matrices invertibles es abierto y la aplicación $M \in \mathcal{M} \mapsto M^{-1}$ es continua. Específicamente, para toda matriz $G \in \mathcal{M}$ y toda matriz $H \in \mathbb{R}^{n\times n}$ tal que $\| G^{-1}H \| < 1$ (o equivalentemente $\| G^{-1} \| \| H \| < 1$). Entonces $G + H$ es invertible y se tiene*

$$\| (G+H)^{-1} \| \leq \frac{\| G^{-1} \|}{1 - \| G^{-1}H \|},$$
$$\| (G+H)^{-1} - G^{-1} \| \leq \frac{\| G^{-1} \| \| G^{-1}H \|}{1 - \| G^{-1}H \|}.$$

Una vez enunciado el Lemma que vamos a usar, procedemos a la presentación del resultado que garantiza la continuidad uniforme del gradiente del funcional reducido.

Lema 10. *Sea* ∇f *dado por* (4.7) *es uniformemente continuo sobre el conjunto de nivel* N_0^ρ *para* Δt *suficientemente pequeño.*

Demostración. De la definición de continuidad uniforme debemos probar que: para todo $\varepsilon > 0$, existe $\delta > 0$ independiente de u,v tal que

$$\| u - v \| \leq \delta \quad \Rightarrow \quad \| \nabla f(u) - \nabla f(v) \| \leq \varepsilon.$$

Sean $u, v \in N_0^\rho$, restando las ecuaciones (4.7) en las variables u y v, se tiene que

$$\| \nabla f(u) - \nabla f(v) \| \leq \frac{1}{\Delta t} \| p^1(u) - p^1(v) \| + \| B^{-1}(u - v) \| + \beta \| D^T(h_\gamma(Du) - h_\gamma(Dv)) \|, \tag{4.8}$$

donde $\| \cdot \|$ es la norma euclídea y $h_\gamma(Du)$ está dada en (3.13). Así, comenzamos por analizar el último termino, el cual corresponde a la primera derivada de la regularización de Huber $\mathcal{C}^2$. Puesto que la regularización de Huber $\mathcal{C}^2$ es una función 2 veces continuamente diferenciable, entonces $h_\gamma(Du)$ es localmente Lipschitz y del Teorema del valor medio tenemos que

$$\| h_\gamma(Du) - h_\gamma(Dv) \| \leq \gamma \| D \| \| u - v \| .$$

Este resultado se obtiene de la definición de $h'_\gamma(\cdot)$, puesto que $\| h'_\gamma(x) \| \leq \gamma$ para cualquier $x \in \mathbb{R}^n$. En el caso del segundo término de (4.8), usamos la definición de norma matricial inducida por la norma euclídea y se tiene que

$$\| B^{-1}(u - v) \| \leq \| B^{-1} \| \| u - v \| .$$

Finalmente, nos queda por acotar el primer término. Para esto usaremos un razonamiento similar al utilizado en el Lema 6. Así, realizando el cambio de variable

$$\mathbf{p} = \mathbb{P}\mathbf{q},$$

con $\mathbb{P}$ la matriz de permutación dada en (4.1). Entonces, la ecuación adjunta está dada por la expresión

$$\mathbb{E}^T\mathbb{P}\mathbf{q} + \nu\mathbb{A}^T\mathbb{P}\mathbf{q} + \mathbb{Z}\mathbb{U}^T\mathbb{P}\mathbf{q} + \text{diag}\,(\mathbb{U}\mathbf{y})\mathbb{P}\mathbf{q} - H^TS^TR^{-1}(SH\mathbf{y} - \mathbf{z}) = 0.$$

Restando dos ecuaciones adjuntas que dependan de u y v tenemos

$$\begin{aligned}\mathbb{E}^T\mathbb{P}(\mathbf{q}(u) - \mathbf{q}(v)) + \nu\mathbb{A}^T\mathbb{P}(\mathbf{q}(u) - \mathbf{q}(v)) + \mathbb{Z}(u)\mathbb{U}^T\mathbb{P}\mathbf{q}(u) - \mathbb{Z}(v)\mathbb{U}^T\mathbb{P}\mathbf{q}(v) \\ + \text{diag}(\mathbb{U}\mathbf{y}(u))\mathbb{P}\mathbf{q}(u) - \text{diag}(\mathbb{U}\mathbf{y}(v))\mathbb{P}\mathbf{q}(v) - H^TS^TR^{-1}SH(\mathbf{y}(u) - \mathbf{y}(v)) = 0\end{aligned}$$

Analizando el sistema anterior componente a componente en el tiempo se distinguen tres casos

Para $j = 1$:

$$\frac{1}{\Delta t}(q^{N_t}(u) - q^{N_t}(v)) + \text{diag}\,(U_p y^1(u))q^{N_t}(u) - \text{diag}\,(U_p y^1(v))q^{N_t}(v) = \left[H^T S^T R^{-1} S H(\mathbf{y}(u) - \mathbf{y}(v))\right]_1 + \frac{1}{\Delta t}(q^{N_t-1}(u) - q^{N_t-1}(v)).$$

Entonces, sumando y restando el término $\text{diag}(U_p y^1(u))q^{N_t}(v)$

$$\begin{aligned}\frac{1}{\Delta t}(q^{N_t}(u) - q^{N_t}(v)) &+ \text{diag}\,(U_p y^1(u))(q^{N_t}(u) - q^{N_t}(v)) \\ &+ [\text{diag}\,(U_p y^1(u)) - \text{diag}\,(U_p y^1(v))]q^{N_t}(v) \\ &= [H^T S^T R^{-1} S H(\mathbf{y}(u) - \mathbf{y}(v))]_1 + \frac{1}{\Delta t}(q^{N_t-1}(u) - q^{N_t-1}(v)).\end{aligned}$$

Reordenando términos tenemos

$$\begin{aligned}&\left[1/\Delta t\mathbb{I} + \text{diag}\,(U_p y^1(u))\right](q^{N_t}(u) - q^{N_t}(v)) \\ &= \left[H^T S^T R^{-1} S H(\mathbf{y}(u) - \mathbf{y}(v))\right]_1 + 1/\Delta t(q^{N_t-1}(u) - q^{N_t-1}(v)) \\ &- \left[\text{diag}\,(U_p y^1(u)) - \text{diag}\,(U_p y^1(v))\right] q^{N_t}(v).\end{aligned}$$

Despejando el termino $(q^{N_t}(u) - q^{N_t}(v))$, tomando normas a ambos lados, aplicando la definición de la norma matricial inducida por la norma euclídea y utilizando desigualdad triangular se obtiene

$$\begin{aligned}\|q^{N_t}(u) - q^{N_t}(v)\| \leq &\left\| \left[1/\Delta t\mathbb{I} + \text{diag}\,(U_p y^1(u))\right]^{-1} \right\| \| H^T S^T R^{-1} S H \| \| \mathbf{y}(u) - \mathbf{y}(v) \| \\ &+ 1/\Delta t \left\| \left[1/\Delta t\mathbb{I} + \text{diag}\,(U_p y^1(u))\right]^{-1} \right\| \| q^{N_t-1}(u) - q^{N_t-1}(v) \| \\ &+ \left\| \left[1/\Delta t\mathbb{I} + \text{diag}\,(U_p y^1(u))\right]^{-1} \right\| \| \text{diag}\,(U_p y^1(u)) - \text{diag}\,(U_p y^1(v)) \| \| q^{N_t}(v) \|\end{aligned}$$

Para $j = 2, \ldots, N_t - 1$:

$$\begin{aligned}\left(\frac{1}{\Delta t}\mathbb{I} + \nu A^T\right)&(q^{N_t-j+1}(u) - q^{N_t-j+1}(v)) + \text{diag}\,(y^{j-1}(u))U_p^T q^{N_t-j+1}(u) \\ &- \text{diag}\,(y^{j-1}(v))U_p^T q^{N_t-j+1}(v) + \text{diag}\,(U_p y^j(u))q^{N_t-j+1}(u) \\ &- \text{diag}\,(U_p y^j(v))q^{N_t-j+1}(v) = [H^T S^T R^{-1} S H(\mathbf{y}(u) - \mathbf{y}(v))]_j \\ &+ \frac{1}{\Delta t}(q^{N_t-j}(u) - q^{N_t-j}(v)).\end{aligned}$$

Sumando y restando los términos

$$\text{diag}\,(y^{j-1}(u))U_p^T q^{N_t-j+1}(v)$$

y diag $(U_p y^j(u))q^{N_t-j+1}(v)$ tenemos

$$\left(\frac{1}{\Delta t}\mathbb{I}+\nu A^T\right)(q^{N_t-j+1}(u)-q^{N_t-j+1}(v))+\text{diag}\,(y^{j-1}(u))U_p^T\left(q^{N_t-j+1}(u)-q^{N_t-j+1}(v)\right)$$
$$+\left[\text{diag}(y^{j-1}(u))-\text{diag}\,(y^{j-1}(v))\right]U_p^T q^{N_t-j+1}(v)$$
$$+\text{diag}\,(U_p y^j(u))\left(q^{N_t-j+1}(u)-q^{N_t-j+1}(v)\right)$$
$$+\left[\text{diag}\,(U_p y^j(u))-\text{diag}\,(U_p y^j(v))\right]q^{N_t-j+1}(v)$$
$$=\left[H^T S^T R^{-1} SH(\mathbf{y}(u)-\mathbf{y}(v))\right]_j+\frac{1}{\Delta t}\left(q^{N_t-j}(u)-q^{N_t-j}(v)\right).$$

Reordenando términos, despejando $q^j(u)-q^j(v)$, tomando normas a ambos lados, aplicando la definición de la norma matricial inducida y la desigualdad triangular se sigue que

$$\|q^{N_t-j+1}(u)-q^{N_t-j+1}(v)\|\leq\left\|\left[1/\Delta t\mathbb{I}+\nu A^T+\text{diag}\,(y^{j-1}(u))U_p^T+\text{diag}\,(U_p y^j(u))\right]^{-1}\right\|$$
$$\{\|\,\text{diag}\,(y^{j-1}(u))-\text{diag}\,y^{j-1}(v)\,\|\|\,U_p\,\|\|\,q^{N_t-j+1}(v)\,\|$$
$$+\|\,\text{diag}\,(U_p y^j(u))-\text{diag}\,(U_p y^j(v))\,\|\|\,q^{N_t-j+1}(v)\,\|$$
$$+\|\,H^T S^T R^{-1} SH\,\|\|\,\mathbf{y}(u)-\mathbf{y}(v)\,\|+1/\Delta t\,\|\,q^{N_t-j}(u)-q^{N_t-j}(v)\,\|\}.$$

La matriz $1/\Delta t\mathbb{I}+\nu A^T+\text{diag}\,(y^{j-1}(u))U_p^T+\text{diag}\,(U_p y^j(u))$ corresponde a la matriz del Corolario 1 y por tanto es invertible.

Para $j=N_t$:

$$\left(\frac{1}{\Delta t}\mathbb{I}+\nu A^T\right)(q^1(u)-q^1(v))+\text{diag}\,(y^{N_t-1}(u))U_p^T q^1(u)-\text{diag}\,(y^{N_t-1}(v))U_p^T q^1(v)$$
$$+\text{diag}\,(U_p y^{N_t}(u))q^1(u)+\text{diag}\,(U_p y^{N_t}(v))q^1(v)=\left[H^T S^T R^{-1} SH(\mathbf{y}(u)-\mathbf{y}(v))\right]_{N_t}.$$

Sumando y restando los términos diag $(y^{N_t-1}(u))U_p^T q^1(v)$ y diag $(U_p y^{N_t}(u))q^1(v)$ tenemos

$$\left(\frac{1}{\Delta t}\mathbb{I}+\nu A^T\right)(q^1(u)-q^1(v))+\text{diag}\,(y^{N_t-1}(u))U_p^T\left(q^1(u)-q^1(v)\right)$$
$$+\left[\text{diag}\,(y^{N_t-1}(u))-\text{diag}\,(y^{N_t-1}(v))\right]U_p^T q^1(v)$$
$$+\text{diag}\,(U_p y^{N_t}(u))\left(q^1(u)-q^1(v)\right)+\left[\text{diag}\,(U_p y^{N_t}(u))-\text{diag}\,(U_p y^{N_t}(v))\right]q^1(v)$$
$$=\left[H^T S^T R^{-1} SH(\mathbf{y}(u)-\mathbf{y}(v))\right]_{N_t}.$$

Entonces, despejando el termino $[q^1(u)-q^1(v)]$, tomando normas a ambos lados, utilizando la definición de la norma matricial inducida por la norma euclídea y la desigual-

dad triangular tenemos

$$\begin{aligned}\| q^1(u) - q^1(v) \| \leq & \left\| \left[{}^1/_{\Delta t}\mathbb{I} + \nu A^T + \text{diag}\,(y^{N_t-1}(u))U_p^T + \text{diag}\,(U_p y^{N_T-1}(u)) \right]^{-1} \right\| \\ & \{\| \text{diag}\,(y^{N_t-1}(u)) - \text{diag}\,(y^{N_t-1}(v)) \|\| U_p \|\| q^1(v) \| \\ & + \| \text{diag}\,(U_p y^{N_t-1}(u)) - \text{diag}\,(U_p y^{N_t-1}(v)) \|\| q^1(v) \| \\ & + \| H^T S^T R^{-1} S H \|\| \mathbf{y}(u) - \mathbf{y}(v) \|\}.\end{aligned}$$

Al igual que en el caso anterior aplicando el Corolario 1 la matriz

$${}^1/_{\Delta t}\mathbb{I} + \nu A^T + \text{diag}\,(y^{j-1}(u))U_p^T + \text{diag}\,(U_p y^j(u)),$$

es invertible.

Definimos las matrices $G_j(u,v)$ y $F_j(u,v)$ por

$$\begin{aligned}G_j(u,v) &= \frac{1}{\Delta t}\mathbb{I} + \nu A + \text{diag}\,(y^{N_t-j}(u))U_p^T + \text{diag}\,(U_p y^{N_t-j+1}(u)), \\ F_j(u,v) &= \left[\text{diag}\,(y^{N_t-j}(u)) - \text{diag}\,(y^{N_t-j}(v))\right] U_p^T \\ &\quad + \left[\text{diag}\,(U_p y^{N_t-j+1}(u)) - \text{diag}\,(U_p y^{N_t-j+1}(v))\right],\end{aligned}$$

para $j = 1, \ldots, N_t - 1$ y

$$G_{N_t}(u,v) = \frac{1}{\Delta t}\mathbb{I} + \text{diag}\,(U_p y^1(u)) \tag{4.9}$$

$$F_{N_t}(u,v) = \text{diag}\,(U_p y^1(u)) - \text{diag}\,(U_p y^1(v)). \tag{4.10}$$

Entonces se tienen los siguientes resultados

$$\begin{aligned}\| q^1(u) - q^1(v) \| \leq & \| G_1(u,v)^{-1} \| \\ & \left\{ \| F_1(u,v) \|\| q^1(v) \| + \| H^T S^T R^{-1} S H \|\| \mathbf{y}(u) - \mathbf{y}(v) \| \right\} \\ \| q^j(u) - q^j(v) \| \leq & \| G_j(u,v)^{-1} \|\| F_j(u,v) \|\| q^j(v) \| \\ & + \| G_j(u,v)^{-1} \|\| H^T S^T R^{-1} S H \|\| \mathbf{y}(u) - \mathbf{y}(v) \| \\ & + {}^1/_{\Delta t} \| G_j(u,v)^{-1} \|\| q^{j-1}(u) - q^{j-1}(v) \| .\end{aligned}$$

Usando el Lema 6 tenemos que

$$\begin{aligned}
\| q^1(u) - q^1(v) \| \leq \rho c_\infty \| G_1(u,v)^{-1} \| \| F_1(u,v) \| \| H^T S^T R^{-1}(SH\mathbf{y}(v) - \mathbf{z}) \|_\infty \\
+ \| G_1(u,v)^{-1} \| \| H^T S^T R^{-1} SH \| \| \mathbf{y}(u) - \mathbf{y}(v) \|, \\
\| q^j(u) - q^j(v) \| \leq \rho c_\infty \| G_j(u,v)^{-1} \| \| F_j(u,v) \| \| H^T S^T R^{-1}(SH\mathbf{y}(v) - \mathbf{z}) \|_\infty \\
+ \| G_j(u,v)^{-1} \| \| H^T S^T R^{-1} SH \| \| \mathbf{y}(u) - \mathbf{y}(v) \| \\
+ 1/\Delta t \| G_j(u,v)^{-1} \| \| q^{j-1}(u) - q^{j-1}(v) \|,
\end{aligned}$$

donde c_∞ es la constante de equivalencia entre la norma euclideana y la norma infinito. Entonces aplicando el Lema 5 tenemos que

$$\begin{aligned}
\| q^j(u) - q^j(v) \| \leq \\
\rho c_\infty \exp\left(\| G_j(u,v)^{-1} \| \right) \| G_j(u,v)^{-1} \| \| F_j(u,v) \| \| H^T S^T R^{-1}(SH\mathbf{y}(v) - \mathbf{z}) \|_\infty \\
+ \exp\left(\| G_j(u,v)^{-1} \| \right) \| G_j(u,v)^{-1} \| \| H^T S^T R^{-1} SH \| \| \mathbf{y}(u) - \mathbf{y}(v) \| .
\end{aligned}$$

para todo $j = 1, \ldots, N_t$. En particular, para $j = N_t$ se tiene

$$\begin{aligned}
\|q^{N_t}(u) - q^{N_t}(v) \| \leq \\
\rho c_\infty \| G_{N_t}(u,v)^{-1} \| \exp(\| G_{N_t}(u,v)^{-1} \|) \| F_{N_t}(u,v) \| \| H^T S^T R^{-1}(SH\mathbf{y}(v) - \mathbf{z}) \|_\infty \\
+ \exp\left(\| G_{N_t}(u,v)^{-1} \| \right) \| G_{N_t}(u,v)^{-1} \| \| H^T S^T R^{-1} SH \| \| \mathbf{y}(u) - \mathbf{y}(v) \| .
\end{aligned} \tag{4.11}$$

Usando el cambio de variable sabemos que $q^{N_t} = p^1$ entonces tenemos que

$$\begin{aligned}
\| p^1(u) - p^1(v) \| \leq \\
\rho c_\infty \| G_{N_t}(u,v)^{-1} \| \exp(\| G_{N_t}(u,v)^{-1} \|) \| F_{N_t}(u,v) \| \| H^T S^T R^{-1}(SH\mathbf{y}(v) - \mathbf{z}) \|_\infty \\
+ \exp\left(\| G_{N_t}(u,v)^{-1} \| \right) \| G_{N_t}(u,v)^{-1} \| \| H^T S^T R^{-1} SH \| \| \mathbf{y}(u) - \mathbf{y}(v) \| .
\end{aligned}$$

Ahora, vamos a acotar cada término de la expresión anterior. En primer lugar, vamos a acotar la matriz $G_{N_t}^{-1}$ utilizando el Lema 9. De la definición de la matriz G_{N_t} dada en (4.9) tenemos

$$G_{N_t} = 1/\Delta t \mathbb{I} + \text{diag}\,(U_p y^1(u))),$$

donde claramente la primera matriz es invertible y su inversa está dada por $\Delta t \mathbb{I}$. Por tanto, para garantizar que

$$\| \Delta t[\text{diag}\,(U_p y^1(u))] \| < 1$$

basta tomar

$$\Delta t < \frac{1}{\| \text{diag}\,(U_p y^1(u)) \|}. \tag{4.12}$$

Una vez verificada esta hipótesis podemos garantizar que G_{N_t} es invertible y además se tiene

la siguiente cota

$$\| G_{N_t}(u,v)^{-1} \| \leq \frac{\Delta t}{1 - \Delta t \| U_p u \|},$$

y esto se obtiene ya que de la ecuación de estado sabemos que $y^1(u) = u$. Entonces, aplicando las propiedades de las normas matriciales inducidas por la norma euclídea tenemos que

$$\| G_{N_t}(u,v)^{-1} \| \leq \frac{\Delta t}{1 - \Delta t \| U_p \| \| u \|}.$$

Finalmente, puesto que $u \in N_0^\rho$ y del Lema 8 este conjunto es compacto podemos garantizar que existe $K_0 > 0$ tal que $\| u \| \leq K_0$. De aquí podemos concluir que

$$\| G_{N_t}(u,v)^{-1} \| \leq \frac{\Delta t}{1 - \Delta t K_0 \| U_p \|}. \tag{4.13}$$

Además, puesto que la función $\exp(\cdot)$ es creciente se tiene que

$$\exp(\| G_{N_t}(u,v)^{-1} \|) \leq \exp\left(\frac{\Delta t}{1 - \Delta t K_0 \| U_p \|}\right). \tag{4.14}$$

Ahora, de la definición de la matriz $F_{N_t}(u,v)$ dada en (4.10) tenemos que

$$\begin{aligned} \| F_{N_t}(u,v) \| &= \| \operatorname{diag}(U_p y^1(u)) - \operatorname{diag}(U_p y^1(v)) \| \\ &= \| U_p(u - v) \| \leq \| U_p \| \| u - v \|. \end{aligned} \tag{4.15}$$

Acotando el término $\| H^T S^T R^{-1}(SH\mathbf{y}(v) - \mathbf{z}) \|_\infty$ tenemos que

$$\begin{aligned} \| H^T S^T R^{-1}(SH\mathbf{y}(v) - \mathbf{z}) \|_\infty &\leq \| H^T S^T R^{-1} \|_\infty \| SH\mathbf{y}(v) - \mathbf{z} \|_\infty \\ &\leq \| H^T S^T R^{-1} \|_\infty \{ \| SH \|_\infty \| \mathbf{y}(v) \|_\infty + \| \mathbf{z} \|_\infty \}. \end{aligned}$$

Aplicando el resultado obtenido en el Lema 3 se tiene que

$$\begin{aligned} \| \mathbf{y}(v) \|_\infty^2 &\leq c_\infty \sum_{j=1}^{N_t} \| y^j(v) \|^2 \\ &< c_\infty \sum_{j=1}^{N_t} (\| v \| + \Delta t \sum_{i=1}^{j} \| f^i \|)^2 \\ &= c_\infty N_t \| v \|^2 + 2c_\infty \Delta t \| v \| \left(\sum_{j=1}^{N_t} \sum_{i=1}^{j} \| f^i \| \right) + c_\infty \Delta t^2 \sum_{j=1}^{N_t} \left(\sum_{i=1}^{j} \| f^i \| \right)^2. \end{aligned}$$

donde c_∞ es la constante de equivalencia de la norma euclídea y la norma infinito. Usando nuevamente el hecho de que $v \in N_0^\rho$ y este conjunto gracias al Lema 8 es compacto; es decir,

existe $K_0 > 0$ tal que $\| v \| \leq K_0$. Entonces:

$$\| \mathbf{y}(v) \|_\infty \leq c_\infty N_t K_0^2 + 2c_\infty \Delta t K_0 \left(\sum_{j=1}^{N_t} \sum_{i=1}^{j} \| f^i \| \right) + c_\infty \Delta t^2 \sum_{j=1}^{N_t} \left(\sum_{i=1}^{j} \| f^i \| \right)^2 =: C(f, K_0) \quad (4.16)$$

Notamos a la constante por $C(f, K_0)$ para resaltar su independencia de v. Así,

$$\| H^T S^T R^{-1} (SH\mathbf{y}(v) - \mathbf{z}) \|_\infty \leq \| H^T S^T R^{-1} \|_\infty \{ \| SH \|_\infty C(f, K_0) + \| \mathbf{z} \|_\infty \} \quad (4.17)$$

Finalmente, vamos a acotar el término $\| \mathbf{y}(u) - \mathbf{y}(v) \|$. Para este propósito vamos a utilizar argumentos similares a los usados para acotar el término del estado adjunto. Comenzamos recordando la expresión de la ecuación de estado dada en la ecuación (3.6)

$$e(\mathbf{y}, u) = \mathbb{E}\mathbf{y} + \nu \mathbb{A}\mathbf{y} + \mathbb{Z}(\mathbf{y})\mathbb{U}\mathbf{y} - \mathbf{f}(u) = 0.$$

Analizando componente a componente en el tiempo podemos diferenciar dos casos

Para $j = 1$:

$$\frac{y^1}{\Delta t} - \frac{u}{\Delta t} = 0$$

Para $j = 2, \ldots, N_t$:

$$\frac{y^j - y^{j-1}}{\Delta t} + \nu A y^j + \text{diag}\,(y^{j-1}) U_p y^j - f^j = 0$$

Restamos dos ecuaciones en las variables u y v, y obtenemos

$$\frac{y^j(u) - y^j(v)}{\Delta t} + \nu A(y^j(u) - y^j(v)) + \text{diag}\,(y^{j-1}(u)) U_p y^j(u) - \text{diag}\,(y^{j-1}(v)) U_p y^j(v) = \frac{y^{j-1}(u) - y^{j-1}(v)}{\Delta t}.$$

Sumando y restando el término $\text{diag}\,(y^{j-1}(u)) U_p y^j(v)$ tenemos

$$\frac{y^j(u) - y^j(v)}{\Delta t} + \nu A(y^j(u) - y^j(v)) + \text{diag}\,(y^{j-1}(u)) U_p (y^j(u) - y^j(v)) + [\text{diag}\,(y^{j-1}(u)) - \text{diag}\,(y^{j-1}(v))] U_p y^j(v) = \frac{y^{j-1}(u) - y^{j-1}(v)}{\Delta t},$$

Reordenando términos se sigue que

$$[y^j(u) - y^j(v)] = [y^{j-1}(u) - y^{j-1}(v)] - \nu \Delta t A [y^j(u) - y^j(v)] - \Delta t \text{diag}\,(y^{j-1}(u)) U_p [y^j(u) - y^j(v)] - \Delta t [\text{diag}\,(y^{j-1}(u)) - \text{diag}\, y^{j-1}(v)] U_p y^j(u)$$

Tomando normas a ambos lados de la igualdad anterior y aplicando la desigualdad triangular

obtenemos

$$\begin{aligned}\| y^j(u) - y^j(v) \| \leq & \| y^{j-1}(u) - y^{j-1}(v) \| + \nu\Delta t \| A \| \| y^j(u) - y^j(v) \| \\ & + \Delta t \| U_p \| y^{j-1}(u) \| \| y^j(u) - y^j(v) \| \\ & + \Delta t \| U_p \| y^j(v) \| \| y^{j-1}(u) - y^{j-1}(v) \|\end{aligned}$$

Utilizando la cota que obtuvimos para y^j dada en el Lema 3 tenemos que

$$\begin{aligned}\| y^j(u) - y^j(v) \| \leq & \| y^{j-1}(u) - y^{j-1}(v) \| + \nu\Delta t \| A \| \| y^j(u) - y^j(v) \| \\ & + \Delta t \| U_p \| \left(\| u \| + \sum_{i=1}^{j-1} \| f^i \| \right) \| y^j(u) - y^j(v) \| \\ & + \Delta t \| U_p \| \left(\| v \| + \sum_{i=1}^{j} \| f^i \| \right) \| y^{j-1}(u) - y^{j-1}(v) \|\end{aligned}$$

Finalmente recordando que $u, v \in N_0^\rho$ y este conjunto es compacto gracias al Lema 8 sabemos que existe $K_0 > 0$ tal que $\| u \| \leq K_0$ y $\| v \| \leq K_0$. Por tanto, se tiene que

$$\begin{aligned}\| y^j(u) - y^j(v) \| \leq & \| y^{j-1}(u) - y^{j-1}(v) \| + \nu\Delta t \| A \| \| y^j(u) - y^j(v) \| \\ & + \Delta t \| U_p \| \left(K_0 + \sum_{i=1}^{j-1} \| f^i \| \right) \| y^j(u) - y^j(v) \| \\ & + \Delta t \| U_p \| \left(K_0 + \sum_{i=1}^{j} \| f^i \| \right) \| y^{j-1}(u) - y^{j-1}(v) \|\end{aligned}$$

Definiendo la constante

$$C(f, K_0) = K_0 + \sum_{i=1}^{j} \| f^i \|,$$

tenemos que

$$\begin{aligned}\| y^j(u) - y^j(v) \| \leq & \| y^{j-1}(u) - y^{j-1}(v) \| + \nu\Delta t \| A \| \| y^j(u) - y^j(v) \| \\ & + \Delta t C(f, K_0) \| U_p \| \| y^j(u) - y^j(v) \| + \Delta t C(f, K_0) \| U_p \| \| y^{j-1}(u) - y^{j-1}(v) \|\end{aligned}$$

Reordenando términos se sigue que

$$\begin{aligned}(1 - \Delta t(\nu \| A \| + C(f, K_0) \| U_p \|)) & \| y^j(u) - y^j(v) \| \\ & \leq (1 + \Delta t C(f, K_0) \| U_p \|) \| y^{j-1}(u) - y^{j-1}(v) \|\end{aligned}$$

Tomando

$$\Delta t < \frac{1}{\nu \| A \| + C(f, K_0) \| U_p \|}, \tag{4.18}$$

podemos garantizar que

$$\| y^j(u) - y^j(v) \| \leq \frac{1 + \Delta t C(f, K_0) \| U_p \|}{1 - \Delta t(\nu \| A \| + C(f, K_0) \| U_p \|)} \| y^{j-1}(u) - y^{j-1}(v) \|,$$

usando este resultado recursivamente tenemos

$$\begin{aligned}
\| y^j(u) - y^j(v) \| &\leq \frac{1 + \Delta t C(f, K_0) \| U_p \|}{1 - \Delta t(\nu \| A \| + C(f, K_0) \| U_p \|)} \| y^{j-1}(u) - y^{j-1}(v) \| \\
&\leq \left(\frac{1 + \Delta t C(f, K_0) \| U_p \|}{1 - \Delta t(\nu \| A \| + C(f, K_0) \| U_p \|)} \right)^2 \| y^{j-2}(u) - y^{j-2}(v) \| \\
&\vdots \\
&\leq \left(\frac{1 + \Delta t C(f, K_0) \| U_p \|}{1 - \Delta t(\nu \| A \| + C(f, K_0) \| U_p \|)} \right)^{j-1} \| y^1(u) - y^1(v) \| \\
&= \left(\frac{1 + \Delta t C(f, K_0) \| U_p \|}{1 - \Delta t(\nu \| A \| + C(f, K_0) \| U_p \|)} \right)^{j-1} \| u - v \|.
\end{aligned}$$

Finalmente, usando la definición de la norma infinito y su equivalencia con la norma euclídea, podemos concluir que existe una constante a la que denominaremos $\tilde{C} > 0$ tal que

$$\| \mathbf{y}(u) - \mathbf{y}(v) \| \leq \tilde{C} \| u - v \|. \tag{4.19}$$

Reemplazando los resultados obtenidos en (4.13)– (4.19) en (4.11) se tiene que si definimos

$$\omega := \left(\frac{\Delta t}{1 - \Delta t K_0 \| U_p \|} \right) \exp \left(\frac{\Delta t}{1 - \Delta t K_0 \| U_p \|} \right)$$

entonces

$$\begin{aligned}
&\| p^1(u) - p^1(v) \| \leq \tilde{C}\omega \| H^T S^T R^{-1} S H \| \| u - v \| \\
&+ \rho c_\infty \omega C(f, K_0) \| H^T S^T R^{-1} \|_\infty \| S H \|_\infty \| U_p \| \| u - v \| \\
&+ \rho c_\infty \omega \| \mathbf{z} \|_\infty \| H^T S^T R^{-1} \|_\infty \| U_p \| \| u - v \| \\
&\leq C(K_0, f, U_p, H, S, R, z, \tilde{C}, \Delta t,) \| u - v \|,
\end{aligned}$$

donde resaltamos la dependencia de la constante $C > 0$ en las matrices y datos que definen el problema pero su independencia de u y v. Finalmente, reemplazando este resultado en la ecuación (4.8) tenemos que

$$\| \nabla f(u) - \nabla f(v) \| \leq \left\{ \frac{C}{\Delta t} + \| B^{-1} \| + \gamma \| D \|^2 \right\} \| u - v \|.$$

Por tanto, basta tomar

$$\delta = \frac{\varepsilon}{C/\Delta t + \| B^{-1} \| + \gamma \| D \|^2},$$

el cual es independiente de u y v, esto nos permite concluir no solo la continuidad uniforme de ∇f sino la Lipschitz continuidad del mismo. Finalmente, cabe recalcar que para que el resultado obtenido anteriormente sea correcto el parámetro Δt tiene que satisfacer al mismo tiempo las condiciones (4.12) y (4.18) lo que nos indica que tan pequeño tiene que ser este parámetro. □

En el caso de los algoritmos que usan la regularización de Huber $\mathcal{C}^1$ la única diferencia con la demostración anterior radica en la forma de acotar el término $\| h_\gamma(Du) - h_\gamma(Dv) \|$.

A continuación presentamos la forma en acotar este término cuando $h_\gamma(\cdot)$ está dada por (3.12).

Sean $u, v \in N_0^o$ entonces, se tiene

$$\begin{aligned}
\| h_\gamma(Du) - h_\gamma(Dv) \| &= \left|\left| \frac{\gamma Du}{\max\{\gamma|Du|,1\}} - \frac{\gamma Dv}{\max\{\gamma|Dv|,1\}} \right|\right|, \\
&= \left|\left| \frac{\gamma Du \max\{\gamma|Dv|,1\} - \gamma Du \max\{\gamma|Du|,1\}}{\max\{\gamma|Du|,1\}\max\{\gamma|Dv|,1\}} \right.\right. \\
&\quad \left.\left. + \frac{\gamma Du \max\{\gamma|Du|,1\} - \gamma Dv \max\{\gamma|Du|,1\}}{\max\{\gamma|Du|,1\}\max\{\gamma|Dv|,1\}} \right|\right| \\
&\leq \frac{\gamma^2 \| Du \|\| D(u-v) \|}{\max\{\gamma|Du|,1\}\max\{\gamma|Du|,1\}} + \frac{\gamma \| D \|\| u-v \|}{\max\{\gamma|Dv|,1\}}, \\
&\leq \frac{\gamma^2 \| Du \|\| D \|\| u-v \|}{\max\{\gamma|Du|,1\}\max\{\gamma|Dv|,1\}} + \frac{\gamma \| D \|\| u-v \|}{\max\{\gamma|Dw|,1\}}, \\
&\leq \frac{2\gamma \| D \|\| u-v \|}{\max\{\gamma|Dv|,1\}}, \\
&\leq 2\gamma \| D \|\| u-v \| .
\end{aligned}$$

Las desigualdades se obtuvieron a partir de la Lipschitz continuidad de la función $\max\{x, 0\}$ y puesto que $\max\{\gamma|Du|, 1\} < 1$. De este modo, podemos concluir que:

$$\| h_\gamma(Du) - h_\gamma(Dv) \| \leq 2\gamma \| D \|\| u-v \| .$$

El resto de la demostración coincide con la del Lema 10. Puesto que el resultado se tiene para ambas definiciones del gradiente, de ahora en adelante no se hará distinción al respecto.

2. Método del descenso más profundo

Uno de los métodos más simples y más utilizados en la teoría de optimización no lineal clásica es el método del descenso más profundo. En nuestro caso, adaptamos dicho algoritmo al caso de la resolución de los problemas de asimilación de datos usando las regularizaciones TV y TGV. Ya que para este algoritmo se necesita únicamente la primera derivada de la función de Huber, vamos a usar la regularización de Huber $\mathcal{C}^1$ dada en la ecuación (3.10) y su primera

derivada dada en la ecuación (3.12).

2.1. Presentación del algoritmo

Los pasos que describen el método para la resolución del problema de asimilación de datos con regularización TV se describe en el Algoritmo 1.

Algorithm 1 Método del descenso más profundo para el problema TV

1: Inicializar $k = 0$, u_0
2: **repeat**
3: Calcular y^k la solución de la ecuación de estado dada en (3.6).
4: Calcular p^k la solución de la ecuación adjunta dada en (3.25).
5: Actualizar la ecuación del gradiente:

$$d^k = -\left[\frac{1}{\Delta t}(p^1)^k + B^{-1}(u^k - u^b) + \beta D^T\left(\frac{\gamma D u^k}{\text{máx}\{\gamma|Du^k|, 1\}}\right)\right]$$

6: Calcular el paso de descenso s usando algún procedimiento de búsqueda lineal que cumpla las condiciones (4.2) y (4.3).
7: Actualizar $u^{k+1} = u^k + sd^k$.
8: $k \leftarrow k + 1$.
9: **until** Un criterio de parada sea satisfecho

De la misma manera, se puede proponer un método del descenso más profundo que nos permita resolver el problema de asimilación de datos con la regularización TGV, el cual queda descrito a través de los pasos indicados en el Algoritmo 2

2.2. Análisis de convergencia

Como se mencionó al inicio de este capítulo vamos a mostrar la convergencia de este método en particular cuando usamos la regla de Armijo para el cálculo del paso de descenso. Recordando lo mencionado en secciones anteriores nuestro principal objetivo en este contexto es usar el resultado dado por el Teorema 6. Así para aplicar este teorema basta verificar que la dirección de descenso satisface la condición del ángulo dada en (4.4). Puesto que en el algoritmo del descenso más profundo se tiene que

$$d^k = -\nabla f(u^k),$$

entonces

$$\frac{-(\nabla f(u^k))^T(-\nabla f(u^k))}{\| \nabla f(u^k) \|^2} = \frac{\| \nabla f(u^k) \|^2}{\| \nabla f(u^k) \|^2} = 1 \geq \eta.$$

Algorithm 2 Método del descenso más profundo para el problema TGV

1: Inicializar $k = 0$, u_0
2: **repeat**
3: Calcular y^k la solución de la ecuación de estado dada en (3.6).
4: Calcular p^k la solución de la ecuación adjunta dada en (3.25).
5: Actualizar la ecuación del gradiente:

$$\begin{bmatrix} (d_u)^k \\ (d_w)^k \end{bmatrix} = - \begin{bmatrix} \frac{1}{\Delta t}(p^1)^k + B^{-1}(u^k - u^b) + \alpha D^T \left(\frac{\gamma D u^k - w^k}{\text{máx}\{\gamma |D u^k - w^k|, 1\}} \right) \\ -\alpha \left(\frac{\gamma D u^k - w^k}{\text{máx}\{\gamma |D u^k - w^k|, 1\}} \right) + \beta \left(\frac{\gamma E w^k}{\text{máx}\{\gamma |E w^k|, 1\}} \right) \end{bmatrix}$$

6: Calcular el paso de descenso s usando algún procedimiento de búsqueda lineal que cumpla las condiciones (4.2) y (4.3).
7: Actualizar $u^{k+1} = u^k + s(d_u)^k$ y $w^{k+1} = w^k + s(d_w)^k$.
8: $k \leftarrow k + 1$.
9: **until** Un criterio de parada sea satisfecho

para algún $\eta \in (0, 1)$ fijo. Por tanto, la dirección d^k satisface la condición del ángulo. Finalmente, para mostrar que los pasos de descenso $\{s_k\}$ satisfacen las condiciones de factibilidad (4.2) y (4.3) vamos a usar el Teorema 7. En primer lugar, debemos notar que la dirección de descenso $d^k = -\nabla f(u^k)$ satisface la condición (4.6). En efecto, tomando la función $\varphi : [0, +\infty) \to [0, +\infty)$ como la identidad se tiene que

$$\varphi \left(\frac{-(\nabla f(u^k))^T (d^k)}{\| d^k \|} \right) = \frac{-(\nabla f(u^k))^T (d^k)}{\| d^k \|} = \frac{-(\nabla f(u^k))^T (-\nabla f(u^k))}{\| \nabla f(u^k) \|} = \| \nabla f(u^k) = \| d^k \| .$$

Finalmente, usando el Lema 10 sabemos que $\nabla f(u)$ es uniformemente continuo sobre el conjunto de nivel N_0^ρ. Así, podemos concluir que el método de descenso más profundo con la regla de Armijo satisface

$$\lim_{k \to \infty} \nabla f(u^k) = 0.$$

Es decir, se tiene la convergencia del algoritmo a un punto estacionario del problema.

3. El método BFGS

El método BFGS es un método ampliamente utilizado, el cual usa información de la curvatura de la función objetivo. Esta información nos permite incrementar la tasa del convergencia del método con respecto al método del descenso más profundo. De los resultados clásicos de optimización no lineal sabemos que este método es globalmente convergente si usamos la regla de Wolfe para la búsqueda lineal. Sin embargo, en este trabajo proponemos un método BFGS globalizado el cual nos garantice que en todas las iteraciones se satisfaga la condición de la

curvatura. Con esta consideración podemos garantizar que la regla de Armijo es suficiente para la convergencia global del método. Al igual que en el caso anterior este método asume que la función objetivo es una vez continuamente diferenciable, es por esto que para este algoritmo usaremos la regularización de Huber $\mathcal{C}^1$. El método para el problema con la regularización de variación total se describe en el Algoritmo 3

Algorithm 3 BFGS Globalizado para el problema con regularización TV

1: Inicializar $k = 0$, u_0, B_0
2: **repeat**
3: Calcular y^k la solución de la ecuación de estado dada en (3.6).
4: Calcular p^k la solución de la ecuación adjunta dada en (3.25).
5: Actualizar la ecuación del gradiente

$$G^k = G(u^k) = \left[\frac{1}{\Delta t}(p^1)^k + B^{-1}(u^k - u^b) + \beta D^T\left(\frac{\gamma D u^k}{\text{máx}\{\gamma|Du^k|, 1\}}\right)\right]$$

6: Actualizar $r_k = u^{k+1} - u^k$
7: Actualizar $t_k = G^{k+1} - G^k$
8: **if** $r_k^T t_k > 0$ **then**
9: Actualizar

$$B_{k+1} = B_k + \frac{(r_k - B_k t_k)s_k^T + r_k(r_k - B_k t_k)^T}{r_k^T r_k} - \frac{(r_k - B_k t_k)_k^t}{(r_k^T t_k)^2} r_k r_k^T$$

10: Calcular la dirección de descenso como solución del sistema

$$B_k d = -G^k$$

11: **Else**
12:
13: Fijar la dirección de descenso como $d = -G^k$
14: **EndIf**
15:
16: Calcular el paso de descenso s usando alguna regla de búsqueda lineal que cumpla las condiciones (4.2) y (4.3)..
17: Calcular $u^{k+1} = u^k + sd^k$.
18: $k \leftarrow k + 1$.
19: **until** Un criterio de parada sea satisfecho

Para el problema regularizado TGV (3.18) vamos a usar de igual manera el método BFGS el cuál está dado en el Algoritmo 4

Algorithm 4 BFGS globalizado para el problema con regularización TGV

Inicializar $k = 0$, u_0, B_0
repeat
  Calcular y^k solución de la ecuación de estado dada en (3.6).
  Calcular p^k solución de la ecuación adjunta dada en (3.25).
  Actualizar la ecuación del gradiente

$$G(u^k, w^k) = \begin{bmatrix} \frac{1}{\Delta t}(p^1)^k + B^{-1}(u^k - u^b) + \alpha D^T \left(\frac{\gamma D u^k - w^k}{\text{máx}\{\gamma |D u^k - w^k|, 1\}} \right) \\ -\alpha \left(\frac{\gamma \left(D u^k - w^k \right)}{\text{máx}\{\gamma |D u^k - w^k|, 1\}} \right) + \beta E^T \left(\frac{\gamma E w^k}{\text{máx}\{\gamma |E w^k|, 1\}} \right) \end{bmatrix}$$

  Actualizar $r_k = u^{k+1} - u^k$
  Actualizar $t_k = G^{k+1} - G^k$

  if $r_k^T t_k > 0$ **then**
    Actualizar

$$B_{k+1} = B_k + \frac{(r_k - B_k t_k) s_k^T + r_k (r_k - B_k t_k)^T}{r_k^T r_k} - \frac{(r_k - B_k t_k)_k^t}{(r_k^T t_k)^2} r_k r_k^T$$

    Calcular la dirección de descenso como solución del sistema

$$B_k [(d_u)^k, (d_w)^k]^T = -G^k$$

  Else

    Tomar la dirección como la del descenso más profundo

$$[(d_u)^k, (d_w)^k]^T = -G^k$$

  EndIf

  Calcular el paso de descenso s usando alguna regla de búsqueda lineal que cumpla las condiciones (4.2) y (4.3)..
  Actualizar $u^{k+1} = u^k + s(d_u)^k$ y $w^{k+1} = w^k + s(d_w)^k$
  $k \leftarrow k + 1$.
until Un criterio de parada sea satisfecho

3.1. Análisis de convergencia

En primer lugar presentamos un resultado que nos garantiza la convergencia del método

Teorema 8. *Sea $f : \mathbb{R}^n \to \mathbb{R}$, $\mathcal{C}^1$ y $x_0 \in \mathbb{R}^n$ tal que el conjunto de nivel N_0^ρ es compacto. Entonces, el algoritmo BFGS está bien definido y si además el número de condición de la sucesión de matrices $\{B_k\}$ es uniformemente acotado entonces todo punto de acumulación de $\{x_k\}$ es un punto estacionario de f.*

Demostración. Siguiendo las ideas de la demostración el método BFGS clásico dadas en la Proposición 13.11 de [Ulbrich and Ulbrich, 2012], solo nos resta por garantizar que en todos los pasos la condición de la curvatura sea satisfecha. Supongamos entonces que las matrices B_k tienen un número de condición uniformemente acotado por la constante c_B. Debido a que las iteraciones en las que no se cumple la condición de la curvatura la matriz es la identidad cuyo número de condición es 1 entonces junto con la hipótesis de la existencia de la constante c_B, todas las matrices están uniformemente acotadas por $C = \text{máx}\{1, c_B\}$. Es claro además, que la función objetivo de nuestros problemas γ–regularizados es $\mathcal{C}^1$. Finalmente, el conjunto de nivel es compacto del Lema 8. Una vez verificadas todas la hipótesis se procede de manera similar al caso del método BFGS clásico y se obtiene el resultado. □

El Teorema 8 nos muestra que cada punto de acumulación de la sucesión generada por el algoritmo converge a un punto estacionario. Queremos probar además que este algoritmo converge a un punto estacionario aplicando el Teorema 6. En primer lugar, vamos a mostrar que la dirección de descenso $d^k = -B_k^{-1}\nabla f(u^k)$ satisface la condición del ángulo. Entonces, de los resultados obtenidos en el capítulo 8 de [Wright and Nocedal, 1999] sabemos que existen constantes $0 < m < M$ tal que

$$m \parallel q \parallel^2 \leq q^T B_k^q \leq M \parallel q \parallel^2 \tag{4.20}$$

para todo $q \in \mathbb{R}^n$. Entonces,

$$\begin{aligned}\frac{-\nabla f(x^k)^T d^k}{\parallel d^k \parallel \parallel \nabla f(x^k) \parallel} &= \frac{\nabla f(x^k) B_k^{-1} \nabla f(x^k)}{\parallel B_k^{-1}\nabla f(x^k) \parallel \parallel d^k \parallel}, \\ &= \frac{\nabla f(x^k) B_k^{-1} B_k B_k^{-1} \nabla f(x^k)}{\parallel B_k \nabla f(x^k) \parallel \parallel B_k B_k^{-1} \nabla f(x^k) \parallel}.\end{aligned}$$

Tomando $q = B_k^{-1}\nabla f(x^k)$ tenemos que

$$\frac{-\nabla f(x^k)^T d^k}{\parallel d^k \parallel \parallel \nabla f(x^k) \parallel} = \frac{q^T B_k q}{\parallel q \parallel \parallel B_k q \parallel}.$$

Usando (4.20) y teniendo en cuenta que esta condición implica que $\| B_k \| \leq M$ entonces:

$$\frac{-\nabla f(x^k)^T d^k}{\| d^k \| \| \nabla f(x^k) \|} \geq \frac{m \| q \|^2}{M \| q \|^2} = \frac{m}{M} =: \eta < 1.$$

Finalmente, para mostrar que los pasos de descenso $\{s_k\}$ satisfacen las condiciones de factibilidad (4.2) y (4.3) usamos el Teorema 7. Comenzamos por mostrar que la dirección $d^k = -B_k^{-1}\nabla f(u^k)$ satisface la condición (4.6), entonces puesto que las matrices $\{B_k\}$ satisfacen (4.20), definimos la función $\varphi : [0, +\infty) \to [0, +\infty)$ tal que

$$\varphi(x) = \frac{x}{M}$$

con M la constante dada en (4.20). Además, la función $\varphi(\cdot)$ es monótona creciente. Entonces, se tiene que

$$\begin{aligned} \frac{-\nabla f(x^k)^T d^k}{\| d^k \|} &= \frac{\nabla f(x^k)^T B_k^{-1} B_k B_k^{-1} \nabla f(x_k)}{\| d^k \|}, \\ &= \frac{(d^k)^T B_k d^k}{\| d^k \|} \leq \frac{M \| d^k \|^2}{\| d^k \|} = M \| d^k \|, \end{aligned}$$

de donde

$$\frac{1}{M}\left(\frac{-\nabla f(x^k)^T d^k}{\| d^k \|}\right) = \varphi\left(\frac{-\nabla f(x^k)^T d^k}{\| d^k \|}\right) \leq \| d^k \| .$$

Así usando el resultado dado en el Lema 10 nos permite concluir que todo punto de acumulación de la sucesión $\{u^k\}$ es un punto estacionario de f.

4. Método de Newton globalizado

Esta sección está dedicada a introducir las técnicas utilizadas para proponer un método Newton globalizado con la finalidad de resolver los problemas 4D-VAR con regularización TV y TGV. Teniendo en cuenta que el método asume que la función objetivo es dos veces diferenciable utilizaremos la regularización de Huber $\mathcal{C}^2$ dada en la ecuación (3.11), su primera derivada dada en la ecuación (3.13) y su segunda derivada dada en (3.14).

4.1. Presentación del método

La idea del método propuesto es la siguiente:

1. Utilizar el método de Newton para problemas de optimización.
2. Formular el problema en términos de las variables primales y duales.
3. Proyectar la matriz de segundo orden.

El método de Newton para problemas de optimización con restricciones de igualdad es reformulado utilizando variables auxiliares de tal manera que nos permitan calcular la información de segundo orden en función del Lagrangeano del problema, el método fue tomado de [De los Reyes, 2015].

Comenzaremos este análisis con el cálculo de las derivadas del Lagrageano del problema 4D-VAR con la regularización de variación total. Así, se define

$$\begin{aligned}\mathcal{L}(y,u,p) = &\frac{1}{2}(SH\mathbf{y}-\mathbf{z})^T R^{-1}(SH\mathbf{y}-\mathbf{z}) + \frac{1}{2}(u-u^b)^T B^{-1}(u-u^b) + \beta\sum_{i=1}^{n} H_\gamma(D_i u)\\ &- p^T(\mathbb{E}\mathbf{y} + \nu\mathbb{A}\mathbf{y} + \mathbb{Z}(\mathbf{y})\mathbb{U}\mathbf{y} - \mathbf{f}(u))\end{aligned}$$

Entonces, el gradiente del Lagrangeano está dado por la expresión:

$$\nabla_{(y,u)}\mathcal{L}(y,u,p) = \begin{bmatrix} H^T S^T R^{-1}(HS\mathbf{y}-\mathbf{z}) - \mathbb{E}^T p - \nu\mathbb{A}^T p - \mathbb{Z}\mathbb{U}^T p - \operatorname{diag}(\mathbb{U}\mathbf{y})p \\ (1/\Delta t)p^1 + B^{-1}(u-u^b) + \beta D^T h_\gamma(Du) \end{bmatrix}$$

Utilizando las ideas desarrolladas en [Calatroni et al., 2015], vamos a añadir la variable dual q, la cual nos permite estabilizar el método. Así, tenemos el nuevo gradiente

$$\nabla_{(y,u)}\mathcal{L}(y,u,p) = \begin{bmatrix} H^T S^T R^{-1}(HS\mathbf{y}-\mathbf{z}) - \mathbb{E}^T p - \nu\mathbb{A}^T p - \mathbb{Z}\mathbb{U}^T p - \operatorname{diag}(\mathbb{U}\mathbf{y})p \\ (1/\Delta t)p^1 + B^{-1}(u-u^b) + \beta D^T q \\ q - h_\gamma(Du) \end{bmatrix}.$$

Su segunda derivada es:

$$\nabla^2(y,u,q)\mathcal{L}(y,u,q,p) = \begin{bmatrix} H^T S^T R^{-1} HS - \mathbb{K} & 0 & 0 \\ 0 & B^{-1} & 0 \\ 0 & -QD & \mathbb{I} \end{bmatrix}$$

donde:

$$\mathbb{K} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & \operatorname{diag}(U_p^T p^2) + \operatorname{diag}(U_p p^2) & 0 & \cdots & 0 \\ 0 & 0 & \operatorname{diag}(U_p^T p^3) + \operatorname{diag}(U_p p^3) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \operatorname{diag}(U_p^T p^{N_t}) + \operatorname{diag}(U_p p^{N_t}) \end{bmatrix}. \tag{4.21}$$

Debemos añadir también la información de segundo orden de la regularización de Huber $\mathcal{C}^2$ dada en (3.14), por la expresión:

$$h'_\gamma(Du) = \begin{cases} \left[\dfrac{1}{|Du|} - \dfrac{(Du)\odot(Du)}{|Du|^3}\right] & \text{si } i \in \mathcal{A}, \\ \gamma 1 & \text{si } i \in \mathcal{B}, \\ \left\{(1-\dfrac{\gamma}{2}\theta_\gamma^2)\left[\dfrac{1}{|Du|} - \dfrac{(Du)\odot(Du)}{|Du|^3}\right] + \gamma^2\theta_\gamma\dfrac{(Du)\odot(Du)}{|Du|^2}\right\} & \text{si } i \in \mathcal{I}. \end{cases}$$

Sin embargo, añadir directamente esta información no es suficiente para garantizar la convergencia del método, es por esto que siguiendo las ideas desarrolladas en [Calatroni et al., 2015] realizamos una proyección la cual consiste en reemplazar el término

$$\frac{(Du)\odot(Du)}{|Du|^3}$$

por

$$\frac{q}{\text{máx}\{1,|q|\}}\odot\frac{Du}{|Du|^2},$$

donde q es la variable dual que acabamos de introducir. Además, todas las operaciones mostradas en la expresión anterior se las realiza componente a componente. Para efectuar esta última operación se construyen matrices diagonales cuyos elementos de la diagonal coincidan con los términos mencionados anteriormente y luego se los multiplica normalmente. Así la matriz Q, aquella que mencionamos en la derivada del Lagrangeano y que corresponde a la información de segundo orden que vamos a incluir en el método, está dada por la expresión:

$$Q=\begin{cases}\left[\dfrac{1}{|Du|}-\dfrac{q}{\text{máx}\{1,|q|\}}\odot\dfrac{Du}{|Du|^2}\right] & \text{si } i\in\mathcal{A},\\ \gamma\mathbb{I} & \text{si } i\in\mathcal{B},\\ \left\{(1-\dfrac{\gamma}{2}\theta_\gamma^2)\left[\dfrac{1}{|Du|}-\dfrac{q}{\text{máx}\{1,|q|\}}\odot\dfrac{Du}{|Du|^2}\right]+\gamma^2\theta_\gamma\dfrac{Du}{|Du|}\odot\dfrac{Du}{|Du|}\right\} & \text{si } i\in\mathcal{I}.\end{cases} \tag{4.22}$$

Además, de los capítulos anteriores sabemos que

$$e'(y,u,q)=\begin{bmatrix}\mathbb{E}+\nu\mathbb{A}+\mathbb{Z}\mathbb{U}+\text{diag}(\mathbb{U}\mathbf{y})\\ -(1/\Delta t)\begin{bmatrix}\mathbb{I}\\0\\\vdots\\0\end{bmatrix}\\ 0\end{bmatrix} \tag{4.23}$$

Entonces, la dirección debe satisfacer el siguiente sistema:

$$\mathbb{H}\begin{bmatrix}\delta_y\\\delta_u\\\delta_q\\\delta_\pi\end{bmatrix}=\begin{bmatrix}0\\-p^1/\Delta t-B^{-1}(u-u^b)-\beta D^Tq\\-q+h_\gamma(Du)\\0\end{bmatrix} \tag{4.24}$$

donde:

$$\mathbb{H}=\begin{bmatrix}\Psi & 0 & 0 & (\Xi)^T\\ 0 & B^{-1} & \beta D^T & \Upsilon^T\\ 0 & -QD & \mathbb{I} & 0\\ \Xi & \Upsilon & 0 & 0\end{bmatrix}$$

con $\Psi = H^T S^T R^{-1} H S - \mathbb{K}$, $\Xi = (\mathbb{E} + \nu\mathbb{A} + \mathbb{Z}\mathbb{U} + \text{diag}(\mathbb{U}\mathbf{y})$ y $\Upsilon = (-1/\Delta t)\,[\mathbb{I}, 0, \ldots, 0]^T$. Entonces el algoritmo para la resolución del problema 4D-VAR con regularización de variación total queda determinado a través de los siguientes pasos:

Algorithm 5 Método de Newton para el problema 4D–VAR con regularización TV

1: Inicializar u^0, $k = 0$
2: **While** : algún criterio de parada sea satisfecho
3: Calcular y^k solución de la ecuación de estado dada en (3.6).
4: Calcular p^k solución de la ecuación adjunta dada en (3.25).
5: Resolver el sistema dado en (4.24).
6: Calcular s usando algún criterio de búsqueda lineal que cumpla las condiciones (4.2) y (4.3).
7: Fijar $u^{k+1} = u^k + s\delta_u$ y $q^{k+1} = q^k + s\delta_q$
8: $k \leftarrow k + 1$
9: **End While**

Realizando un análisis similar para el problema de asimilación de datos 4D-VAR con regularización de variación total generalizada tenemos los siguientes resultados.

$$\begin{aligned}\mathcal{L}(y,u,w,p) = &\frac{1}{2}(SH\mathbf{y}-\mathbf{z})^T R^{-1}(SH\mathbf{y}-\mathbf{z}) + \frac{1}{2}(u-u^b)^T B^{-1}(u-u^b)\\ &+ \alpha\sum_{i=1}^{n} H_\gamma(D_i u - w_i) + \beta\sum_{i=1}^{n-1} H_\gamma(E_i w)\\ &- p^T(\mathbb{E}\mathbf{y} + \nu\mathbb{A}\mathbf{y} + \mathbb{Z}(\mathbf{y})\mathbb{U}\mathbf{y} - \mathbf{f}(u)).\end{aligned}$$

Entonces, el gradiente con respecto a (y,u,w) es el siguiente

$$\nabla_{(y,u,w,p)}\mathcal{L}(y,u,w,p) = \begin{bmatrix} H^T S^T R^{-1}(SH\mathbf{y}-\mathbf{z}) - \mathbb{E}^T p - \nu\mathbb{A}^T p - \mathbb{Z}\mathbb{U}^T p - \text{diag}(\mathbb{U}\mathbf{y})p \\ (1/\Delta t)p^1 + B^{-1}(u-u^b) + \alpha D^T h_\gamma(Du-w) \\ -\alpha h_\gamma(Du-w) + \beta h_\gamma(Ew)\end{bmatrix}$$

Añadiendo las variables duales q_1 y q_2

$$\nabla_{(y,u,w)}\mathcal{L}(y,u,w,p,q_1,q_2) = \begin{bmatrix} H^T S^T R^{-1}(SH\mathbf{y}-z) - \mathbb{E}^T p - \nu\mathbb{A}^T p - \mathbb{Z}\mathbb{U}^T p - \text{diag}(\mathbb{U}\mathbf{y})p \\ (1/\Delta t)p^1 + B^{-1}(u-u^b) + \alpha D^T q_1 \\ -\alpha q_1 + \beta q_2 \\ q_1 - h_\gamma(Du-w) \\ q_2 - h_\gamma(Ew)\end{bmatrix}.$$

Entonces, la segunda derivada está dada por la expresión

$$\nabla^2_{(y,u,w,q_1,q_2)}\mathcal{L}(y,u,w,q_1,q_2,p) = \begin{bmatrix} \Psi & 0 & 0 & 0 & 0 \\ 0 & B^{-1} & 0 & \alpha D^T & 0 \\ 0 & 0 & 0 & -\alpha\mathbb{I} & \beta E^T \\ 0 & -Q_1 D & Q_1 & \mathbb{I} & 0 \\ 0 & 0 & -Q_2 E & 0 & \mathbb{I} \end{bmatrix}$$

donde $\Psi = H^T S^T R^{-1} S H - \mathbb{K}$ y $\mathbb{K}$ está dada en (4.21). Las matrices Q_1 y Q_2 está dadas por

$$Q_1 = \begin{cases} \left[\frac{1}{|Du-w|} - \frac{q_1}{\text{máx}\{1,|q_1|\}} \odot \frac{Du-w}{|Du-w|^2}\right] & \text{si } i \in \mathcal{A}, \\ \gamma\mathbb{I} & \text{si } i \in \mathcal{B}, \\ \left\{\left(1-\frac{\gamma}{2}\theta_\gamma^2\right)\left[\frac{1}{|Du-w|} - \frac{q_1}{\text{máx}\{1,|q_1|\}} \odot \frac{Du-w}{|Du-w|^2}\right] + \gamma^2\theta_\gamma \frac{Du-w}{|Du-w|} \odot \frac{Du-w}{|Du-w|}\right\} & \text{si } i \in \mathcal{I}, \end{cases}$$

y

$$Q_2 = \begin{cases} \left[\frac{1}{|Ew|} - \frac{q_2}{\text{máx}\{1,|q_2|\}} \odot \frac{Ew}{|Ew|^2}\right] & \text{si } i \in \mathcal{A}, \\ \gamma\mathbb{I} & \text{si } i \in \mathcal{B}, \\ \left\{(1-\frac{\gamma}{2}\theta_\gamma^2)\left[\frac{1}{|Ew|} - \frac{q_2}{\text{máx}\{1,|q_2|\}} \odot \frac{Ew}{|Ew|^2}\right] + \gamma^2\theta_\gamma \frac{Ew}{|Ew|} \odot \frac{Ew}{|Ew|}\right\} & \text{si } i \in \mathcal{I}. \end{cases}$$

Usando nuevamente la ecuación (4.23) y de la definición del algoritmo de Newton tenemos que la dirección debe satisfacer el sistema:

$$\mathbb{H}\begin{bmatrix} \delta_y \\ \delta_u \\ \delta_w \\ \delta_{q_1} \\ \delta_{q_2} \\ \delta_\pi \end{bmatrix} = \begin{bmatrix} 0 \\ -p^1/\Delta t - B^{-1}(u-u^b) - \alpha D^T q_1 \\ +\alpha q_1 - \beta E^T q_2 \\ -q_1 + h_\gamma(Du-w) \\ -q_2 + h_\gamma(Ew) \\ 0 \end{bmatrix} \tag{4.25}$$

donde:

$$\mathbb{H} = \begin{bmatrix} \Psi & 0 & 0 & 0 & 0 & \Xi^T \\ 0 & B^{-1} & 0 & \alpha D^T & 0 & \Upsilon^T \\ 0 & 0 & 0 & -\alpha\mathbb{I} & \beta E^T & 0 \\ 0 & -Q_1 D & Q_1 & \mathbb{I} & 0 & 0 \\ 0 & 0 & -Q_2 E & 0 & \mathbb{I} & 0 \\ \Xi & \Upsilon & 0 & 0 & 0 & 0 \end{bmatrix}.$$

y $\Xi = (\mathbb{E} + \nu\mathbb{A} + \mathbb{Z}\mathbb{U} + \text{diag}(\mathbb{U}\mathbf{y})$, $\Upsilon = (-1/\Delta t)\,[\mathbb{I}, 0, \ldots, 0]^T$. Entonces el método de Newton globalizado para el problema 4D-VAR con regularización de variación total generalizada queda completamente determinado a través los pasos que se describen en el Algoritmo 6.

Algorithm 6 Método de Newton para el problema 4D–VAR con regularización TGV

1: Inicializar u^0, $k = 0$
2: **While** : algún criterio de parada sea satisfecho
3: Calcular y^k solución de la ecuación de estado dada en (3.6).
4: Calcular p^k solución de la ecuación adjunta dada en (3.25).
5: Resolver el sistema dado en (4.25).
6: Calcular s usando una regla de búsqueda lineal que cumpla las condiciones (4.2) y (4.3).
7: Fijar $u^{k+1} = u^k + s\delta_u$, $w^{k+1} = w^k + s\delta_w$, $q_1^{k+1} = q_1^k + s\delta_{q_1}$ y $q_2^{k+1} = q_2^k + s\delta_{q_2}$
8: $k \leftarrow k + 1$
9: **End While**

4.2. Análisis de convergencia

Al igual que para los algoritmos anteriores mostraremos en esta sección los resultados teóricos que nos permiten garantizar la convergencia del método. Los siguientes resultados sobre la convergencia del método serán realizados analizando el sistema reducido, siguiendo las ideas desarrolladas en [Hintermüller and Stadler, 2006]. El análisis se realizará para el problema con la regularización de variación total. Sin embargo, los resultados pueden ser extendidos fácilmente al problema con la regularización de variación total generalizada. Comenzamos está sección recordando la estructura que tiene la matriz de segundo orden para el algoritmo tipo Newton:

$$\mathbb{H} = \begin{bmatrix} H^T S^T R^{-1} H S - \mathbb{K} & 0 & 0 & (\mathbb{E} + \nu\mathbb{A} + \mathbb{Z}\mathbb{U} + \operatorname{diag}(\mathbb{U}\mathbf{y}))^T \\ 0 & B^{-1} & \beta D^T & (-1/\Delta t)\,[\mathbb{I}, 0, \ldots, 0] \\ 0 & -QD & \mathbb{I} & 0 \\ (\mathbb{E} + \nu\mathbb{A} + \mathbb{Z}\mathbb{U} + \operatorname{diag}(\mathbb{U}\mathbf{y})) & -(1/\Delta t)\begin{bmatrix} \mathbb{I} \\ 0 \\ \vdots \\ 0 \end{bmatrix} & 0 & 0 \end{bmatrix}$$

Definimos las matrices

$$\begin{aligned} \Xi &= (\mathbb{E} + \nu\mathbb{A} + \mathbb{Z}\mathbb{U} + \operatorname{diag}(\mathbb{U}\mathbf{y})), \\ \Upsilon &= (-1/\Delta t)\,[\mathbb{I}, 0, \ldots, 0]^T, \\ \Psi &= H^T S^T R^{-1} S H - \mathbb{K}. \end{aligned}$$

Además, podemos notar que la matriz Ξ corresponde a la matriz asociada a la ecuación adjunta cuya invertibilidad fue probada en el Corolario 1 y del sistema anterior se tienen las siguientes

igualdades

$$\begin{aligned}\delta_q &= QD\delta_u - q + h_\gamma(Du)\\ \delta_y &= (\Xi)^{-1}\Upsilon\delta_u\\ \delta_\pi &= -(\Xi)^{-T}\Psi(\Xi)^{-1}\Upsilon\delta_u\end{aligned}$$

Con las consideraciones anteriores, el sistema mediante el cual se encuentra la dirección se reduce a:

$$\left(B^{-1} + \beta D^T Q D + \Upsilon^T(\Xi)^{-T}\Psi(\Xi)^{-1}\Upsilon\right)\delta_u = -\frac{1}{\Delta t}p^1 - B^{-1}(u-u^b) - \beta D^T h_\gamma(Du). \tag{4.26}$$

Así, definimos la matriz

$$\Pi = \left(B^{-1} + \beta D^T Q D + \Upsilon^T(\Xi)^{-T}\Psi(\Xi)^{-1}\Upsilon\right). \tag{4.27}$$

y el vector

$$\phi(u) = -\frac{1}{\Delta t}p^1 - B^{-1}(u-u^b) - \beta D^T h_\gamma(Du). \tag{4.28}$$

de donde el sistema que determina la dirección es

$$\Pi d = \phi(u).$$

De ahora en adelante nos concentramos en probar que la matriz Π genera efectivamente direcciones de descenso. Este resultado nos permite deducir que la proyección realizada globaliza el método. Comenzamos por demostrar que la matriz es simétrica.

Proposición 2. *La matriz* Π *dada en* (4.27) *es simétrica.*

Demostración. De la definición tenemos que

$$\begin{aligned}\Pi^T &= \left(B^{-1} + \beta D^T Q D + \Upsilon^T(\Xi)^{-T}\Psi(\Xi)^{-1}\Upsilon\right)^T\\ &= (B^T)^{-1} + \beta D^T Q^T D + \Upsilon^T(\Xi)^{-T}\Psi^T(\Xi)^{-1}\Upsilon\end{aligned}$$

Puesto que B es una matriz de covarianza, es simétrica y su inversa también lo es. Así la demostración se reduce a mostrar que Q y Ψ son simétricas. Para mostrar que Q es simétrica, debemos recordar su definición dada en la ecuación (4.22) y notar que de la forma en la que está definida la matriz, es diagonal y por tanto simétrica. Finalmente analizaremos la matriz Ψ.

$$\begin{aligned}\Psi^T &= \left(H^T S^T R^{-1} S H - \mathbb{K}\right)^T\\ &= H^T S^T (R^T)^{-1} S H - \mathbb{K}^T\end{aligned}$$

Al igual que en el caso anterior puesto que R es una matriz de covarianza es simétrica al igual que su inversa. De la definición de la matriz $\mathbb{K}$ dada en (4.21) se puede notar que son matrices diagonales y por tanto serán simétricas. Así podemos concluir que en general la matriz Ψ es simétrica. □

Finalmente, mostraremos que la matriz Π es definida positiva. Antes de mostrar directamente el resultado que nos garantiza que las direcciones generadas por el sistema (4.26) son efectivamente de descenso, mencionamos algunos resultados preliminares. Comenzamos por recordar el Lema 6 que como se mencionó anteriormente es un resultado análogo al mostrado en [Volkwein, 1997] Lema 3.4 página 83. En el cual se muestra que el estado adjunto está acotado. Ahora mostraremos un resultado intermedio, el cual nos permitirá garantizar que la matriz Π es definida positiva.

Lema 11. *Si el término* $\| H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z}) \|_\infty$ *es suficientemente pequeño, entonces existe* $\kappa > 0$ *tal que para cualquier vector* $\eta \in \mathbb{R}^m$ *se tiene que*

$$\eta^T \mathcal{L}''_{(y,y)}(y,u,p)\eta \geq \kappa \| \eta \|^2 .$$

Demostración. En primer lugar recordamos la estructura que tiene el Lagrangeano del problema (3.17)

$$\begin{aligned}\mathcal{L}(y,u,p) = &\frac{1}{2}(SH\mathbf{y} - \mathbf{z})^T R^{-1}(SH\mathbf{y} - \mathbf{z}) + \frac{1}{2}(u - u^b)^T B^{-1}(u - u^b) + \beta \sum_{i=1}^{n} H_\gamma(D_i u) \\ &- p^T(\mathbb{E}\mathbf{y} + \nu\mathbb{A}\mathbf{y} + \mathbb{Z}(\mathbf{y})\mathbb{U}\mathbf{y} - \mathbf{f}(u)).\end{aligned}$$

Entonces,

$$\mathcal{L}'_y(y,u,p) = H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z}) - \mathbb{E}^T\mathbf{p} - \nu\mathbb{A}^T\mathbf{p} - \mathbb{Z}\mathbb{U}^T\mathbf{p} - \mathrm{diag}(\mathbb{U}\mathbf{y})\mathbf{p}.$$

La segunda derivada con respecto a y es

$$\mathcal{L}''_{(y,y)}(y,u,p) = H^T S^T R^{-1} SH - \mathbb{K}$$

donde $\mathbb{K}$ es la matriz dada por (4.21). Ahora, usando la definición de $\mathbb{K}$ y multiplicando a

ambos lados por el vector $\eta \in \mathbb{R}^{nN_t}$ tenemos que

$$\begin{aligned}
\eta^T \mathcal{L}''_{(y,y)} \eta &= \eta^T H^T S^T R^{-1} SH\eta - \eta^T \mathbb{K}\eta \\
&\geq c \parallel \eta \parallel^2 -2 \sum_{i=1}^{N_t} \eta_i^T (\mathbb{U}_i p^i) \eta_i \\
&= c \parallel \eta \parallel^2 -2 \sum_{i=1}^{N_t} \sum_{j=1}^{n} ((U_p)_{ij}) p_j^i (\eta_j^i)^2 \\
&\geq c \parallel \eta \parallel^2 -2 \parallel \mathbb{U} p \parallel_\infty \sum_{i=1}^{N_t} \sum_{j=1}^{n} (\eta_j^i)^2 \\
&\geq c \parallel \eta \parallel^2 -2 \parallel \mathbb{U} \parallel_\infty \parallel p \parallel_\infty \parallel \eta \parallel^2 \\
&\geq (c - 2\rho \parallel \mathbb{U} \parallel_\infty \parallel H^T S^T R^{-1} (SHy - z) \parallel_\infty) \parallel \eta \parallel^2
\end{aligned}$$

donde c es el valor propio más pequeño de la matriz $H^T S^T R^{-1} SH$ y la última desigualdad se la obtiene al aplicar el Lema 6. Entonces, tomando $\kappa = c - 2\rho \parallel H^T S^T R^{-1}(SHy - z) \parallel_\infty$. Para garantizar que la constante $\kappa > 0$ basta tener que

$$\parallel H^T S^T R^{-1} (SH\mathbf{y} - \mathbf{z}) \parallel_\infty \leq \frac{c}{2\rho}$$

lo que indica que tan pequeño tiene que ser el término $\parallel H^T S^T R^{-1}(SH\mathbf{y} - \mathbf{z}) \parallel_\infty$ para obtener resultado. □

Con los resultados anteriores es posible mostrar que la matriz Π_s es efectivamente una matriz definida positiva, resultado que enunciamos a continuación.

Proposición 3. *La matriz* Π *definida en* (4.27) *es definida positiva.*

Demostración. Vamos a utilizar la definición, entonces sea $\xi \in \mathbb{R}^n \setminus \{0\}$ queremos probar que:

$$\xi^T \Pi \xi > 0$$

Así,

$$\begin{aligned}
\xi^T \Pi \xi &= \xi^T \left(B^{-1} + \beta D^T Q D + \Upsilon^T \Xi^{-T} \Psi \Xi^{-1} \Upsilon \right) \xi \\
&= \xi^T B^{-1} \xi + \beta \xi^T D^T Q D \xi + \xi^T \Upsilon^T \Xi^{-T} \Psi \Xi^{-1} \Upsilon \xi.
\end{aligned}$$

Definiendo los vectores $\zeta = D\xi$ y $\psi = \Xi^{-1}\Upsilon\xi$ tenemos que

$$\xi^T \Pi \xi = \xi^T B^{-1} \xi + \beta \zeta^T Q \zeta + \psi^T \Psi \psi.$$

Notemos en primer lugar que B es una matriz de covarianza definida positiva, $\beta > 0$, y $\Psi =$

$\mathcal{L}''_{(y,y)}$. Así, del Lema 11 se tiene que

$$\psi^T \Psi \psi \geq \kappa \parallel \psi \parallel^2 \geq 0.$$

Por tanto, basta mostrar que la matriz Q dada en (4.22) es semi–definida positiva. Empezamos por recordar que la matriz Q es diagonal por tanto es suficiente analizar el signo de los elementos de la diagonal. Sabemos que

$$\frac{|q_i|}{\text{máx}\{|q_i|, 1\}} \leq 1.$$

Por tanto,

$$\frac{q_i}{\text{máx}\{|q_i|, 1\}} \odot \frac{D_i u}{|D_i u|^2} \leq \frac{D_i u}{|D_i u|^2}. \tag{4.29}$$

Utilizando la definición dada en (4.22) vamos a analizar tres casos:

$i \in \mathcal{A}$:

$$Q_{ii} = \frac{1}{|D_i u|} - \frac{q_i}{\text{máx}\{|q_i|, 1\}} \odot \frac{D_i u}{|D_i u|^2},$$

Utilizando la cota obtenida en (4.29) se tiene que

$$Q_{ii} \geq \frac{1}{|D_i u|} - \frac{D_i u}{|D_i u|^2} = \frac{1}{|D_i u|} \left(1 - \frac{D_i u}{|D_i u|}\right).$$

Sin embargo, de la definición del valor absoluto se tiene

$$\frac{D_i u}{|D_i u|} \leq 1$$

por tanto

$$Q_{ii} \geq \underbrace{\frac{1}{|D_i u|}}_{\geq 0} \underbrace{\left(1 - \frac{D_i u}{|D_i u|}\right)}_{\geq 10} \geq 0.$$

$i \in \mathcal{B}$: De la definición se tiene automáticamente que

$$Q_{ii} = \gamma > 0$$

$i \in \mathcal{I}$: Vamos a comenzar por recordar la forma del conjunto $\mathcal{I} := \{i \colon |\gamma |D_i u| - 1| \leq 1/2\gamma\}$. Para este conjunto el elemento de la diagonal de Q está dado por la siguiente expresión

$$Q_{ii} = \left\{ (1 - \frac{\gamma}{2}\theta_\gamma^2) \left[\frac{1}{|D_i u|} - \frac{q_i}{\text{máx}\{1, \gamma |q_i|\}} \odot \frac{D_i u}{|D_i u|^2} \right] + \gamma^2 \theta_\gamma \frac{D_i u}{|D_i u|} \odot \frac{D_i u}{|D_i u|} \right\},$$

con $\theta_\gamma = (1 - \gamma |D_i u| + 1/2\gamma)$. Del análisis realizado en el primer caso sabemos que

$$\left[\frac{1}{|D_i u|} - \frac{q_i}{\text{máx}\{|q_i|, 1\}} \odot \frac{D_i u}{|D_i u|^2} \right] \geq 0.$$

Además, de la definición del conjunto $\mathcal{I}$ y de θ_γ podemos concluir

$$\theta_\gamma \geq 0,$$
$$1 - \frac{\gamma}{2}\theta_\gamma^2 \geq 1 - \frac{1}{2\gamma} \geq 0.$$

Por tanto, se tiene que

$$Q_{ii} \geq \gamma^2 \theta_\gamma \frac{D_i u}{|D_i u|} \odot \frac{D_i u}{|D_i u|} = \gamma^2 \theta_\gamma \left(\frac{D_i u}{|D_i u|} \right)^2 \geq 0.$$

Así, de los tres casos podemos concluir que efectivamente Q es semi–definida positiva.

Finalizamos la demostración con la siguiente expresión

$$\xi^T \Pi \xi = \underbrace{\xi^T B^{-1} \xi}_{>0} + \underbrace{\beta}_{>0} \underbrace{\zeta^T Q \zeta}_{\geq 0} + \underbrace{\psi^T \Psi \psi}_{\geq \kappa \|\psi\|^2},$$

por lo que se puede concluir que

$$\xi^T \Pi \xi > 0.$$

de donde se obtiene el resultado. □.

Gracias a estos resultados podemos comprobar que la proyección realizada sobre la información de segundo orden nos garantiza que la matriz es definida positiva y por tanto lo podemos entender como el proceso de globalización que se ha mencionado.

Al igual que para los algoritmos anteriores el principal objetivo es aplicar el Teorema 6 que nos garantiza la convergencia del método a puntos estacionarios de la función. Así, empezaremos por probar que la dirección generada por este método satisface la condición del ángulo. De la Proposición 3 la matriz Π dada en (4.27) es definida positiva. Entonces, existen constantes $0 < m_k < M_k$ tales que

$$m_k \parallel q \parallel^2 \leq q^T \Pi_k q \leq M_k \parallel q \parallel^2, \qquad \forall q \in \mathbb{R}^n.$$

A continuación presentamos un resultado que nos garantiza la existencia de constantes $0 < m < M$ independientes de k tal que

$$m \parallel q \parallel^2 \leq q^T \Pi_k q \leq M \parallel q \parallel^2, \qquad \forall q \in \mathbb{R}^n.$$

Proposición 4. *Sea la matriz $\Pi_k = \Pi(u_k)$ definida en* (4.27) *entonces existen constantes $0 < m < M$ tal que*

$$m \parallel q \parallel^2 \leq q^T \Pi_k q \leq M \parallel q \parallel^2, \qquad \forall q \in \mathbb{R}^n.$$

para $u_k \in N_0^\rho$ y Δt suficientemente pequeño.

Demostración. Usando la definición de Π_k tenemos que

$$q^T \Pi_k q = q^T B^{-1} q + \beta q^T D^T Q_k D q + q^T \Upsilon^T \Xi_k^{-T} \Psi_k \Xi_k^{-1} \Upsilon q.$$

donde $Q_k = Q(u_k)$, $\Xi_k = \Xi(u_k)$ y $\Psi_k = \Psi(u_k)$. Entonces,

$$q^T \Pi_k q \geq \lambda_{min}(B^{-1}) \parallel q \parallel^2 + \beta \lambda_{min}(D^T Q_k D) \parallel q \parallel^2 + \lambda_{min}(\Upsilon^T (\Xi_k)^{-T} \Psi_k (\Xi_k)^{-1} \Upsilon) \parallel q \parallel^2,$$

donde $\lambda_{min}(B^{-1})$ es el menor valor propio de la matriz B^{-1}. Así, procedemos a acotar los valores propios más pequeños de las matrices $D^T Q_k D$ y $\Upsilon^T \Xi_k^{-T} \Psi_k \Xi_k^{-1} \Upsilon$. En primer lugar, recordando que Q_k es una matriz semi–definida positiva podemos concluir que $D^T Q_k D$ también es semi–definida positiva y por tanto

$$\lambda_{min}(D^T Q_k D) \geq 0.$$

Por otro lado, analizamos la matriz $\Upsilon^T (\Xi_k)^{-T} \Psi_k (\Xi_k)^{-1} \Upsilon$. Al igual que en el caso anterior, recordando que Ψ_k es definida positiva, podemos concluir que la matriz $\Upsilon^T (\Xi_k)^{-T} \Psi_k (\Xi_k)^{-1} \Upsilon$ también es definida positiva entonces,

$$\lambda_{min}(\Upsilon^T (\Xi_k)^{-T} \Psi_k (\Xi_k)^{-1} \Upsilon) > 0.$$

Así, tenemos que

$$\begin{aligned} q^T \Pi_k q &\geq \lambda_{min}(B^{-1}) \parallel q \parallel^2 + \underbrace{\lambda_{min}(D^T Q_k D)}_{\geq 0} \parallel q \parallel^2 + \underbrace{\lambda_{min}(\Upsilon^T (\Xi_k)^{-T} \Psi_k (\Xi_k)^{-1} \Upsilon)}_{>0} \parallel q \parallel^2, \\ &> \lambda_{min}(B^{-1}) \parallel q \parallel^2 . \end{aligned}$$

Entonces, definiendo $m := \lambda_{min}(B^{-1}) = \lambda_{max}(B)$ se tiene el resultado. Por otro lado, sabemos que

$$q^T \Pi_k q \leq \lambda_{max}(B^{-1}) \parallel q \parallel^2 + \beta \lambda_{max}(D^T Q_k D) \parallel q \parallel^2 + \parallel \Upsilon^T (\Xi_k)^{-T} \Psi_k (\Xi_k)^{-1} \Upsilon \parallel \parallel q \parallel^2 .$$

En este caso vamos a analizar los valores propios de cada matriz. Así, en el caso de la matriz $D^T Q_k D$ puesto que Q_k es diagonal se tiene la siguiente relación

$$D^T Q_k D = Q_k D^T D.$$

La matriz D representa el gradiente discreto es una matriz que no depende de k así que denotaremos por $\lambda_i(D^T D)$ a los valores propios de la matriz $D^T D$. De la definición de los valores propios sabemos que

$$D^T D x = \lambda(D^T D) x,$$

multiplicando a ambos lados por Q_k tenemos que

$$Q_k D^T D x = \lambda(D^T D) Q_k x,$$

entonces,

$$D^T Q_k D x = Q_k D^T D x = \lambda(D^T D) Q_k x = \lambda(D^T D)\lambda(Q_k) x.$$

Así, los valores propios de la matriz $D^T Q_k D$ se pueden caracterizar como

$$\lambda(D^T Q_k D) = \lambda(D^T D)\lambda(Q_k).$$

Por otro lado, puesto que la matriz Q_k es diagonal y sus elementos de la diagonal están acotados de la siguiente manera

$$0 \leq (Q_k)_{ii} \leq \gamma.$$

Por lo tanto, tenemos que

$$\lambda_{max}(D^T Q_k D) = \lambda_{max}(D^T D)\gamma.$$

Finalmente, analizamos la norma de la matriz $\mathrm{Y}^T \Xi_k^{-T} \Psi_k \Xi_k^{-1} \mathrm{Y}$. Usando las propiedades de las normas matriciales inducidas por la norma euclídea tenemos que

$$\| \mathrm{Y}^T \Xi_k^{-T} \Psi_k \Xi_k^{-1} \mathrm{Y} \| \leq \| \Psi_k \| \| \Xi_k^{-1} \|^2 \| \mathrm{Y} \|^2 \tag{4.30}$$

Empezamos analizando el término $\| \Psi_k \|$, así

$$\| \Psi_k \| \leq \| H^T S^T R^{-1} S H \| + \| \mathbb{K} \| .$$

Usando la definición de la matriz $\mathbb{K}$ dada en (4.21) tenemos que

$$\| \mathbb{K}_k \| = 2 \| \mathbb{U} \| \| \mathbf{p}_k \| \leq 2\rho c_\infty \| \mathbb{U} \| \| \| H^T S^T R^{-1} \|_\infty \| S H \mathbf{y}_k - \mathbf{z} \|_\infty,$$

donde c_∞ es la constante de equivalencia entre la norma euclideana y la norma infinito. Además, la última desigualdad se obtuvo usando el Lema 6. Entonces,

$$\begin{aligned} \| \mathbb{K}_k \| &\leq 2\rho c_\infty \| \mathbb{U} \| \| H^T S^T R^{-1} \|_\infty [\| S H \|_\infty \| \mathbf{y}_k \|_\infty + \| \mathbf{z} \|_\infty] \\ &\leq 2\rho c_\infty \| \| \mathbb{U} \| \| H^T S^T R^{-1} \|_\infty [C(f, K_0) \| S H \|_\infty + \| \mathbf{z} \|_\infty] \end{aligned}$$

donde la última desigualdad se la obtuvo al aplicar el Lema 3. Además, puesto que $u_k \in N_0^\rho$ y este conjunto en compacto gracias al Lema 8 lo que implica la existencia de una constante $K_0 > 0$ tal que $\| u_k \| \leq K_0$. Definimos entonces

$$\mu := 2\rho c_\infty \| \| \mathbb{U} \| \| H^T S^T R^{-1} \|_\infty [C(f, K_0) \| S H \|_\infty + \| \mathbf{z} \|_\infty] .$$

Así,

$$\| \Psi_k \| \leq \| H^T S^T R^{-1} S H \| + \mu \tag{4.31}$$

Por otro lado, analizaremos la norma de la matriz Ξ_k^{-1}, para esto usaremos el Lema de Banach para operadores inversos dado en el Lema 9. Entonces, empezamos recordando la estructura de la matriz Ξ_k dada por

$$\Xi_k = \mathbb{E} + \nu\mathbb{A} + \mathbb{Z}_k\mathbb{U} + \text{diag}\,(\mathbb{U}\mathbf{y}_k).$$

En primer lugar recordamos la definición de la matriz $\mathbb{E}$ dada en la ecuación (3.5) la cual tiene la siguiente estructura

$$\mathbb{E} = \frac{1}{\Delta t}\mathbb{F}$$

donde $\mathbb{F}$ es una matriz invertible. Así

$$\mathbb{E}^{-1} = \Delta t\mathbb{F}^{-1} \tag{4.32}$$

Entonces, realizamos la siguiente asociación de la matrices

$$\Xi^{-1} = [\mathbb{E} + (\nu\mathbb{A} + \mathbb{Z}_k\mathbb{U} + \text{diag}\,(\mathbb{U}\mathbf{y}_k))]$$

donde claramente $\mathbb{E}$ es invertible y su inversa está dada en (4.32). Para aplicar el Lema 9 debemos en primer lugar garantizar que $\Delta t \parallel \mathbb{F}^{-1} \parallel\parallel \nu\mathbb{A} + \mathbb{Z}_k\mathbb{U} + \text{diag}\,(\mathbb{U}\mathbf{y}_k) \parallel < 1$. Lo cual se obtiene al tomar Δt suficientemente pequeño de tal manera que

$$\Delta t < \frac{1}{\parallel \mathbb{F}^{-1} \parallel\parallel \nu\mathbb{A} + \mathbb{Z}_k\mathbb{U} + \text{diag}\,(\mathbb{U}\mathbf{y}_k) \parallel}.$$

Así,

$$\parallel \Xi_k^{-1} \parallel \leq \frac{\Delta t \parallel \mathbb{F}^{-1} \parallel}{1 - \Delta t \parallel \mathbb{F}^{-1} \parallel\parallel \nu\mathbb{A} + \mathbb{Z}_k\mathbb{U} + \text{diag}\,(\mathbb{U}\mathbf{y}_k) \parallel}$$

Ahora, acotando el término $\parallel \nu\mathbb{A} + \mathbb{Z}_k\mathbb{U} + \text{diag}\,(\mathbb{U}\mathbf{y}_k) \parallel$ tenemos que

$$\begin{aligned}\parallel \nu\mathbb{A} + \mathbb{Z}_k\mathbb{U} + \text{diag}\,(\mathbb{U}\mathbf{y}_k) \parallel &\leq \nu \parallel \mathbb{A} \parallel + 2 \parallel \mathbb{U} \parallel\parallel \mathbf{y}_k \parallel \\ &\leq \nu \parallel \mathbb{A} \parallel + 2C(f, K_0) \parallel \mathbb{U} \parallel\end{aligned}$$

donde la última desigualdad se obtuvo al aplicar el Lema 3 y puesto que $u_k \in N_0^\rho$ y este conjunto es compacto. De esta manera se obtiene la siguiente cota

$$\parallel \Xi_k^{-1} \parallel \leq \frac{\Delta t \parallel \mathbb{F}^{-1} \parallel}{1 - \Delta t \parallel \mathbb{F}^{-1} \parallel [\nu \parallel \mathbb{A} \parallel + 2C(f, K_0) \parallel \mathbb{U} \parallel} \tag{4.33}$$

la cual es independiente de k.

Reemplazando las desigualdades (4.31) y (4.33) en (4.30) se tiene que

$$q^T\Pi_k q \leq \lambda_{max}(B^{-1}) \parallel q \parallel^2 + \beta\gamma\lambda_{max}(D^TD) \parallel q \parallel^2 + \left(\parallel H^TS^TR^{-1}SH \parallel + \mu\right) \parallel \mathrm{Y} \parallel^2 \frac{\Delta t \parallel \mathbb{F}^{-1} \parallel^2}{(1 - 2C(f, K_0,) \parallel \mathbb{F}^{-1} \parallel^2 \parallel \mathbb{U} \parallel)^2} \parallel q \parallel^2$$

Entonces, definimos

$$M := \lambda_{max}(B^{-1}) + \beta\gamma\lambda_{max}(D^TD) \left(\parallel H^TS^TR^{-1}SH \parallel + \mu\right) \parallel \mathrm{Y} \parallel^2 \frac{\Delta t \parallel \mathbb{F}^{-1} \parallel^2}{(1 - 2C(f, K_0,) \parallel \mathbb{F}^{-1} \parallel\parallel \mathbb{U} \parallel)^2}$$

Resumiendo, hemos encontrado dos constantes $0 < m < M$ independientes de k tal que

$$m \parallel q \parallel^2 \leq q^T\Pi_k q \leq M \parallel q \parallel^2 . \qquad \square$$

Mostramos a continuación que las direcciones d^k generadas por este algoritmo satisfacen la condición del ángulo. Es decir, debemos probar que existe una constante $\eta < 1$ tal que

$$\frac{-\phi(u^k)^T d^k}{\parallel d^k \parallel\parallel \phi(u^k) \parallel} \geq \eta,$$

donde $\phi(u^k)$ esta dada en la expresión (4.28). Así,

$$\frac{-\phi(u^k)^T d^k}{\parallel d^k \parallel\parallel \phi(u^k) \parallel} = \frac{\phi(u^k)\Pi_k^{-1}\phi(u^k)}{\parallel \Pi_k^{-1}\phi(u^k) \parallel\parallel d^k \parallel},$$
$$= \frac{\phi(u^k)\Pi_k^{-1}\Pi_k\Pi_k^{-1}\phi(u^k)}{\parallel \Pi_k\phi(u^k) \parallel\parallel \Pi_k\Pi_k^{-1}\phi(u^k) \parallel}.$$

Definiendo $q = \Pi_k^{-1}\phi(u^k)$ tenemos que

$$\frac{-\phi(u^k)^T d^k}{\parallel d^k \parallel\parallel \phi(u^k) \parallel} = \frac{q^T\Pi_k q}{\parallel q \parallel\parallel \Pi_k q \parallel}.$$

Usando la Proposición 4 y teniendo en cuenta que esta condición implica que $\parallel \Pi_k \parallel \leq M$ entonces:

$$\frac{-\phi(u^k)^T d^k}{\parallel d^k \parallel\parallel \phi(x^k) \parallel} \geq \frac{m \parallel q \parallel^2}{M \parallel q \parallel^2} = \frac{m}{M} =: \eta < 1.$$

Finalmente, para mostrar que los pasos de descenso $\{s_k\}$ satisfacen las condiciones de factibilidad (4.2) y (4.3) usamos el Teorema 7.

La condición (4.6) se satisface automáticamente puesto que de la Proposición 4 las matrices

$\{\Pi_k\}$ satisfacen

$$m \parallel q \parallel^2 \leq q^T \Pi_k q \leq M \parallel q \parallel^2 .$$

Así, tomando la función monótona creciente $\varphi : [0, +\infty) \to [0, +\infty)$ tal que

$$\varphi(x) = \frac{x}{M}$$

se tiene que

$$\begin{aligned} \frac{\phi(u^k)^T d^k}{\parallel d^k \parallel} &= \frac{\phi(u^k)^T \Pi_k^{-1} \Pi_k \Pi_k^{-1} \phi(u_k)}{\parallel d^k \parallel}, \\ &= \frac{(d^k)^T \Pi_k d^k}{\parallel d^k \parallel} \leq \frac{M \parallel d^k \parallel^2}{\parallel d^k \parallel} = M \parallel d^k \parallel, \end{aligned}$$

lo que nos permite concluir que

$$\frac{1}{M} \left(\frac{\phi(u^k)^T d^k}{\parallel d^k \parallel} \right) = \varphi \left(\frac{\phi(u^k)^T d^k}{\parallel d^k \parallel} \right) \leq \parallel d^k \parallel .$$

Usando el Lema 10 tenemos que $\nabla f(u)$ es uniformemente continuo sobre el conjunto de nivel N_0^ρ y por tanto se puede concluir que todo punto de acumulación de la sucesión $\{u_k\}$ generada por el algoritmo es un punto estacionario de f.

Finalmente, con razonamientos similares se puede definir un sistema reducido para el problema con la regularización de variación total generalizada. Usando razonamientos similares se puede probar resultados similares a los mostrados para el problema con variación total y de esta manera poder garantizar la convergencia del método a puntos estacionarios del problema.

5. Búsqueda lineal polinomial

Finalmente, concluimos este capítulo estudiando la técnica de búsqueda lineal polinomial, la que nos garantiza mejores resultados teniendo en cuenta trabajos anteriores en el campo del tratamiento de imágenes. En estos trabajos se muestra que el uso del algoritmo *backtraking* usual para efectuar una búsqueda lineal con la regla de Armijo no es suficiente, puesto que los valores que satisfacen esta regla son muy pequeños, tomándole así al algoritmo muchas más iteraciones de las que de verdad se necesita. En casos extremos, ni siquiera se puede garantizar la existencia de pasos que satisfagan las reglas de Armijo.

La regla de búsqueda lineal polinomial que se utiliza en este trabajo esta detallada en [Dennis Jr and Schnabel, 1996]. A continuación presentamos un resumen de la estrategia y el algoritmo utilizado en este trabajo.

La estrategia de *backtracking* queda determinada a través del parámetro $\alpha < 1$ tal que

$$s_k = \alpha^k.$$

Usualmente, este parámetro α puede ser constante. Sin embargo, conocemos de antemano que usar un valor de α constante puede no ser una buena opción ya que si $\alpha \approx 1$ le tomará muchas iteraciones al algoritmo encontrar el valor tal que se satisfaga la condición de Armijo. Por otro lado, si α es muy pequeño es probable que en la segunda o tercera iteración nos encontremos lejos del valor óptimo y esto puede producir que el algoritmo itere indefinidamente o que los valores para los cuales se satisfaga la condición de Armijo sean muy pequeños. Como solución a este problema se propone el uso de un algoritmo que tome valores variables de α a lo largo de las iteraciones. Específicamente nos vamos a concentrar en el método de interpolación polinomial.

Empecemos por asumir que para cada iteración del método de optimización podemos definir la función $\tilde{f}(\cdot)$ tal que

$$\tilde{f}(s) = f(x_k + sp_k),$$

la cual es la restricción del funcional objetivo a la recta que pasa por x_k en la dirección p_k. Asumimos que el algoritmo de *backtracking* empieza con $s = 1$. En el caso de que para este valor de s no se satisfaga Armijo, es decir se tiene que

$$\tilde{f}(1) > \tilde{f}(0) + c_1\tilde{f}'(0), \tag{4.34}$$

con $c_1 = 1e - 4$. En este caso, debemos realizar una segunda iteración del algoritmo de *backtraking*. La idea principal del método es aprovechar la información que tenemos sobre $\tilde{f}$ para aproximarla mediante algún modelo polinomial y hallar s tal que minimice dicho modelo. De las consideraciones anteriores sabemos que contamos con información de 3 valores sobre la función $\tilde{f}$, es decir

$$\tilde{f}(0) = f(x_k) \quad \text{y} \quad \tilde{f}'(0) = \nabla f(x_k)^T p_k.$$

Además, conocemos que $\tilde{f}(1) = f(x_k + p_k)$. En la primera iteración del algoritmo vamos a buscar un modelo cuadrático que aproxime a $\tilde{f}$ tal que se cumplan las tres condiciones anteriores. Se puede probar que el modelo que satisface estas condiciones está dado por la siguiente expresión.

$$\tilde{m}_q(s) = \left[\tilde{f}(1) - \tilde{f}(0) - c_1\tilde{f}'(0)\right] s^2 + \tilde{f}'(0)s + \tilde{f}(0).$$

Ahora, para calcular el valor de s que minimiza la función cuadrática utilizamos la condiciones de optimalidad de primer y de segundo orden. El punto estacionario de esta función es:

$$\tilde{s} = \frac{-\tilde{f}'(0)}{2[\tilde{f}(1) - \tilde{f}(0) - \tilde{f}'(0)]}, \tag{4.35}$$

además,

$$\tilde{m}_q''(\tilde{\lambda}) = 2[\tilde{f}(1) - \tilde{f}(0) - \tilde{f}'(0)] > 0,$$

esto se obtiene ya que (4.34) es verdadero y $\tilde{f}'(0) < 0$. Por tanto el mínimo del modelo cuadrático es efectivamente $\tilde{s}$ y en el algoritmo del *backtracking* fijamos $s_k = \tilde{s}$. De las propiedades

anteriores sabemos que

$$\tilde{s} = \frac{1}{2(1-\alpha)}.$$

Esto se obtiene de igual manera de la desigualdad (4.34). Sin embargo, los problemas surgen cuando $\tilde{f}(1)$ es mucho más grande que $\tilde{f}(0)$, lo que implica que $\tilde{s}$ sea muy pequeño. Considerando lo anterior el algoritmo propone fijar un valor mínimo de $l = 1/10$.

Una vez obtenido el paso siguiente s_k, existen dos posibilidades: s_k satisface la condición de Armijo en cuyo caso el algoritmo termina o s_k no satisface dicha condición. En caso de que necesitemos realizar una nueva iteración del algoritmo de búsqueda lineal, podemos usar nuevamente un modelo cuadrático. Sin embargo, para poder sacar provecho de toda la información con la que contamos se propone aproximar $\tilde{f}$ por un modelo cúbico que se ajuste a las valores de $\tilde{f}(0)$, $\tilde{f}'(0)$, $\tilde{f}(s_{k-2})$ y s_{k-1}. El polinomio cúbico que satisface estas condiciones es

$$\tilde{m}_{cu}(s) = as^3 + bs^2 + \tilde{f}'(0)s + \tilde{f}(0)$$

con

$$a = \frac{1}{s_{k-2} - s_{k-1}} \left[\frac{1}{s_{k-2}^2}(\tilde{f}(s_{k-2}) - \tilde{f}(0) - \tilde{f}'(0)s_{k-2}) + \frac{1}{s_{k-1}^2}(\tilde{f}(s_{k-1}) - \tilde{f}(0) - \tilde{f}'(0)s_{k-1}) \right]$$

$$b = \frac{1}{s_{k-2} - s_{k-1}} \left[-\frac{s_{k-1}}{s_{k-2}^2}(\tilde{f}(s_{k-2}) - \tilde{f}(0) - \tilde{f}'(0)s_{k-2}) + \frac{s_{k-2}}{s_{k-1}^2}(\tilde{f}(s_{k-1}) - \tilde{f}(0) - \tilde{f}'(0)s_{k-1}) \right]$$

Además, el mínimo de esta función está dado por

$$\tilde{s} = \frac{-b + \sqrt{b^2 - 3a\tilde{f}'(0)}}{3a}. \tag{4.36}$$

Se puede probar que para $c_1 < 1/4$, $\tilde{s}$ nunca será un valor complejo y que $\tilde{s} > 0$. Además, si $\tilde{f}(s_{k-2}) \geq \tilde{f}(0)$, entonces $\tilde{s} < (2/3)\, s_{k-1}$. Al igual que en el caso anterior pueden darse casos en los que el valor de $\tilde{s}$ sea muy pequeño y por tanto se fijan cotas inferiores y superiores para evitar dichos casos patológicos. En el algoritmo se fija la cota inferior $l = 0{,}1$ y la cota superior $u = 0{,}5$.

Además, de los pasos descritos anteriormente, el algoritmo implementado incluye dos condiciones adicionales:

- Se fija un valor mínimo para la longitud del paso de descenso el cual se utiliza cuando la condición de Armijo no ha sido satisfecha, pero $\| s_k p_k \|_2$ es suficientemente pequeño. Entonces el proceso de búsqueda lineal termina. Este criterio previene que el algoritmo itere infinitamente si p_k no es una dirección de descenso.

- Se fija además, un valor máximo para la longitud del paso. Estos casos ocurren cuando la matriz a partir de la cual calculamos la dirección de descenso tiende a ser singular. Este

proceso previene que nos escapemos de la región de convergencia de los algoritmos.

El algoritmo como tal se describe a través de los siguientes pasos:

Algorithm 7 Búsqueda lineal polinomial

Inicializar $s = 1$, $c_1 = 1e - 4$, $k = 1$
While $\tilde{f}(s_k) > \tilde{f}(0) + c_1\tilde{f}'(0)$
If $s_k > 0{,}1$
    If $k = 2$
        Calcular $\tilde{s}$ como en (4.35)
        Fijar $s_k = \tilde{s}$
    Else
        Calcular $\tilde{s}$ como en (4.36)
        If $\tilde{s} < {}^1/_2 s_{k-1}$
            Fijar $s_k = \tilde{s}$
        Else
            Fijar $s_k = {}^1/_2 s_{k-1}$
        End If
    End If
Else $s_k = 0{,}1$
End If
$k \leftarrow k + 1$
End While

Capítulo 5

Experimentos Numéricos

Este capítulo se concentra en la resolución de diferentes experimentos numéricos cuyo principal objetivo es mostrar el desempeño de los algoritmos propuestos, la manera en a que las diferentes regularizaciones recuperan las soluciones de los problemas y la influencia de los datos. Para poder cumplir con estos objetivos se han diseñado cuatro experimentos los cuales se detallan a continuación. El primero tiene como objetivo principal mostrar el desempeño de los algoritmos estudiados en este trabajo al momento de resolver un problema de asimilación de datos variacional de la ecuación de Burgers, además de estudiar la influencia que tienen las reglas de búsqueda lineal con respecto al número de iteraciones que cada algoritmo necesita para satisfacer un criterio de parada. A lo largo de este trabajo hemos mencionado que la regularización de variación total (TV) genera un efecto de escalonamiento en las soluciones, es por eso que el segundo experimento ha sido diseñado con la finalidad de permitirnos visualizar este efecto y además mostrar como la regularización de variación total generalizada (TGV) lo elimina. El tercer experimento analiza la influencia que tiene la cantidad de observaciones que tenemos con respecto a la forma en la que cada problema recupera las soluciones. La principal motivación de este experimento es poder interpretar los resultados aquí obtenidos y aplicarlos al problema de asimilación de datos meteorológicos para de esta manera poder concluir en la necesidad de contar un sistema más completo de recolección de observaciones meteorológicas para poder mejorar los pronósticos considerablemente. El principal objetivo del último experimento es mostrar que el esquema de *discretizar–luego–optimizar* propuesto en este trabajo nos permite resolver problemas que desde el punto de vista funcional son más complejos de analizar. Concretamente, presentamos un experimento cuyo principal objetivo es resolver distintos problemas de asimilación de datos con la ecuación de Burgers cuando su parámetro de viscosidad tiende a cero y analizar el comportamiento del algoritmo de Newton globalizado al resolver dichos problemas.

En cada experimento se detallan los parámetros utilizados. El esquema para la generación de los datos es el siguiente: primero fijamos una función constante a trozos o lineal a trozos la cual será nuestra solución exacta. A partir de esta función resolvemos la ecuación de estado asociada y extraemos la información correspondiente a las observaciones. Dependiendo del

experimento podemos añadir ruido o no a las observaciones a través de la suma de una variable aleatoria que siga una distribución Gaussiana con media 0 y la matriz de covarianza R. Para la generación de la información previa o *background* tomamos la variable de estado exacta correspondiente al tercer instante de tiempo y le sumamos una variable aleatoria que siga una distribución Gaussiana de media 0 y matriz de covarianza B. En todos los experimentos se utilizan $n = 50$ puntos de discretización espacial y $N_t = 100$ puntos en la discretización temporal a menos que se indique lo contrario. Las observaciones serán tomadas cada $N = 10$ puntos en la malla espacial y cada $M = 25$ puntos en la discretización temporal a excepción del experimento que analiza la influencia de la cantidad de observaciones en la reconstrucción de la solución en el cual se indica la cantidad de observaciones que se toman. El horizonte temporal es fijo para todos los experimentos $T = 1$. El parámetro de viscosidad de la ecuación de Burgers es $\nu = 0{,}6$ a excepción del último experimento en el cual se indicará el valor de parámetro escogido. Además, el parámetros de la regularización de Huber, será fijado como $\gamma = 100$ para todos los experimentos, sin hacer distinción en el tipo de regularización que se use.

La resolución de los diferentes experimentos realizados con el problema regularizado TGV se realiza con un esquema de inicialización en caliente, el cual se refiere al uso de la solución obtenida por el problema TV como punto inicial para el problema TGV. Esta técnica fue usada en trabajos pasados, por ejemplo, [Calatroni et al., 2015].

Como criterio de parada usamos la condición

$$\| u^k - u^{k-1} \|_{\mathbb{R}^n} < \eta,$$

con $\eta = 1e-3$.

En algunos experimentos se añadió una regularización elíptica para garantizar la convergencia de los métodos. Esta regularización consiste en añadir el término

$$\mathcal{R}_e(u) = \frac{\mu}{2}(u^T A u + u^T u) \tag{5.1}$$

en el caso del problema con la regularización TV y el término

$$\mathcal{R}_e(u, w) = \frac{\mu}{2}(u^T A u + u^T u + w^T (E^T E) w + w^T w) \tag{5.2}$$

para el problema con la regularización TGV. Esta regularización se utilizó principalmente para los métodos del descenso más profundo y BFGS, garantizando la convergencia de los mismos. En el caso del método Newton globalizado se utilizó únicamente en casos específicos. El valor del parámetro μ utilizado será mencionado al inicio de cada experimento.

Además, se toma como número máximo de iteraciones $iter_{max} = 1000$. En caso de que algún algoritmo no satisfaga el criterio de parada hasta este número de iteraciones se considera que el algoritmo no converge.

1. Análisis de convergencia de los métodos estudiados

Este experimento está diseñado con la finalidad de estudiar el comportamiento de cada uno de los algoritmos presentados en el capítulo anterior. Además, comparamos los resultados obtenidos al usar diferentes reglas de búsqueda lineal. Para efectuar dicha comparación vamos a resolver el problema de asimilación de datos con la ecuación de Burgers usando la regularización de variación total (TV). En este experimento usamos la función exacta dada por la expresión:

$$u_{\text{ex}} = \begin{cases} 2 & \text{si } x \leq 5, \\ 0 & \text{caso contrario.} \end{cases}$$

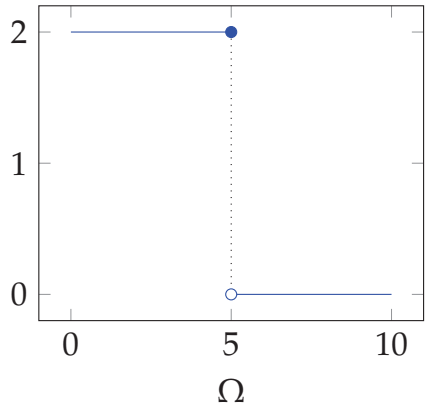


Figura 4. Función exacta para el experimento de análisis de velocidad convergencia de los algoritmos.

En este caso asumimos que las observaciones no tienen errores, por tanto la matriz $R = \mathbb{I}$, por otro lado la matriz $B = (0{,}1)\mathbb{I}$. El parámetro de la regularización TV está dado por $\beta = 0{,}5$, el parámetro de la regularización de Huber es $\gamma = 100$. El problema resuelto con los algoritmos del descenso más profundo y BFGS usará la regularización elíptica definida en (5.1) con el parámetro $\mu = 1e - 10$.

Este problema será resuelto usando el método del descenso más profundo (SDM) definido en el Algoritmo 1, el método BFGS dado por el Algoritmo 3 y el método de Newton Globalizado (NW–G) definido en el Algoritmo 5. Además, usaremos diferentes reglas de búsqueda lineal, usando el algoritmo de *backtracking* con las regla de Armijo y de Wolfe, y finalmente la regla de búsqueda lineal polinomial dada en el Algoritmo 7. Los resultados obtenidos en este experimento son resumidos en la Tabla 1, en donde las siglas "No CV" indican que los algoritmos necesitan más de 1000 iteraciones para satisfacer el criterio de parada. Mientras que las siglas "No s fact" indica que el algoritmo *backtracking* no encontró un paso factible. Así, de estos resultados podemos concluir que el algoritmo Newton Globalizado con la regla de búsqueda lineal polinomial necesita menos iteraciones que los otros algoritmos y además la solución obtenida es más cercana a la solución exacta, que aquellas obtenidas con los otros algoritmos y es por esta razón que para los experimentos siguientes se usará únicamente este

Tabla 1. Comparación de los resultados obtenidos con los algoritmos (SDM), (BFGS) y (NW–G) y usando diferentes reglas de búsqueda lineal

	Armijo		**Wolfe**		**Polinomial**	
	Iter	$\| u - u_{ex} \|$	**Iter**	$\| u - u_{ex} \|$	**Iter**	$\| u - u_{ex} \|$
SDM	61	1.7572	61	1.7572	106	1.8031
BFGS	No *s* fact	No *s* fact	No *s* fact	No *s* fact	87	1.5217
NW–G	No *s* fact	No *s* fact	No *s* fact	No *s* fact	38	0.8853

método.

La Figura 5 muestra las soluciones obtenidas con el algoritmo del descenso más profundo, el método BFGS y el Newton Globalizado usando la regla de búsqueda lineal polinomial. De esta figura podemos concluir nuevamente que el método de Newton globalizado nos permite obtener soluciones más acercadas a la solución exacta y en una menor cantidad de iteraciones.

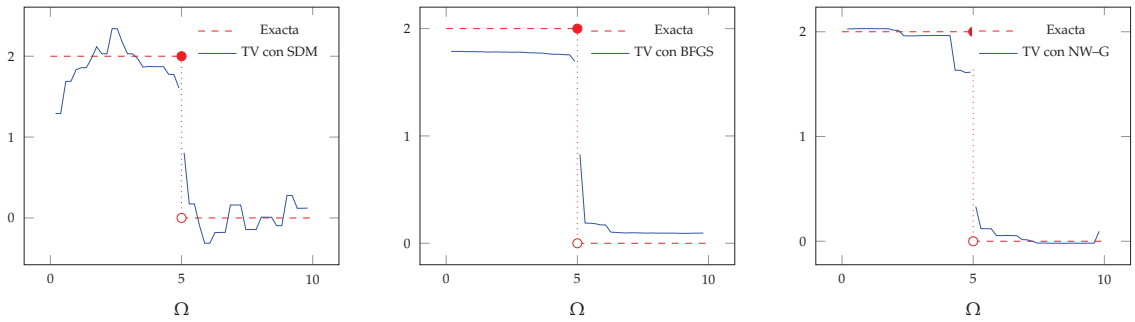
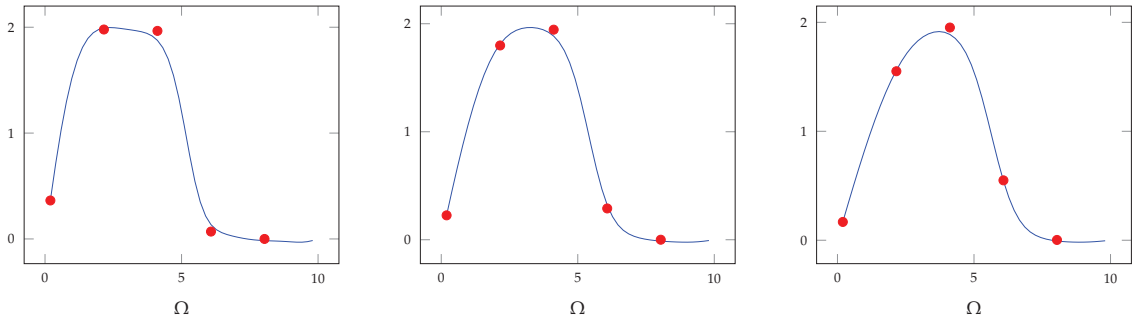


Figura 5. Soluciones obtenidas con la regla de búsqueda lineal polinomial para el método SDM (izquierda), BFGS (centro) y NW–G (derecha)

Además, para este experimento vamos a mostrar la manera en la que la solución obtenida ajusta la variable de estado final a las observaciones introducidas como datos. En la Figura 6 se muestra dicha comparación para tres instantes de tiempo determinados. En cada uno de los instantes se puede apreciar el ajuste de las observaciones con el estado óptimo. Este comportamiento nos permite concluir que si se ajusta de esta manera para un futuro próximo, los pronósticos ampliando la ventana de tiempo serán igual de correctos. De cierta manera podemos extrapolar estos resultados al problema de asimilación de datos meteorológicos justificando de esta manera la importancia de la inclusión de nuevas técnicas de regularización.

2. Comparación entre la regularización TV y TGV

Este experimento está diseñado con la finalidad de mostrar el efecto de escalonamiento producido por la regularización de variación total (TV) y la manera en la que las soluciones del problema con la regularización de variación total generalizada (TGV) eliminan dicho efec-



(a) Solución para $t = 0{,}13$ **(b)** Solución para $t = 0{,}255$ **(c)** Solución para $t = 0{,}38$

Figura 6. Comparación entre las observaciones y el estado final asociado a la solución con el algoritmo NW–G con $\beta = 0{,}5$

to y se obtiene una mejor reconstrucción de la solución. Para el problema con la regularización (TGV) se usa el esquema de inicialización en caliente mencionado al inicio del capítulo. Además, para escoger los parámetros de la regularización TGV se utiliza la siguiente heurística presentada en [Calatroni et al., 2015]

$$\frac{\beta^*}{\alpha^*} \in \frac{1}{n}(0{,}75; 1{,}5).$$

La función exacta utilizada en este experimento está dada por la siguiente expresión:

$$u_{ex} = \begin{cases} \dfrac{3x}{2} & \text{si } x \in [0;2], \\ x - 2 & \text{si } x \in [2;4], \\ 2 & \text{si } x \in [4;5], \\ 3x - 15 & \text{si } x \in [5;6], \\ \dfrac{-3x}{4} + 6 & \text{si } x \in [6;8], \\ 0 & \text{si } x \in [8;10], \end{cases}$$

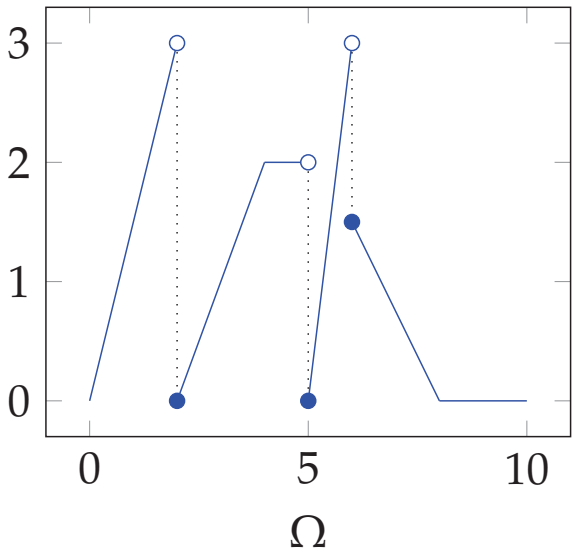
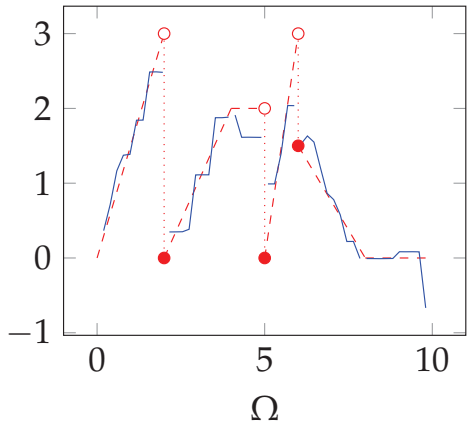
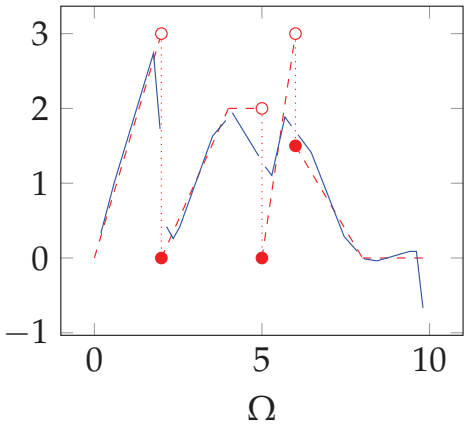


Figura 7. Función exacta para el experimento de comparación de soluciones de los problemas TV y TGV



(a) Solución del problema con la regularización TV con $\beta = 0{,}75$



(b) Solución del problema con la regularización TGV con $\alpha = 5$ y $\beta = 0{,}15$

Figura 8. Soluciones para el experimento de comparación del problema TV y TGV

En las Tablas 2 y 3 se muestran los resultados para el problema con la regularización TV y TGV, respectivamente. En la Figura 8 se muestra los resultados para ambos problemas.

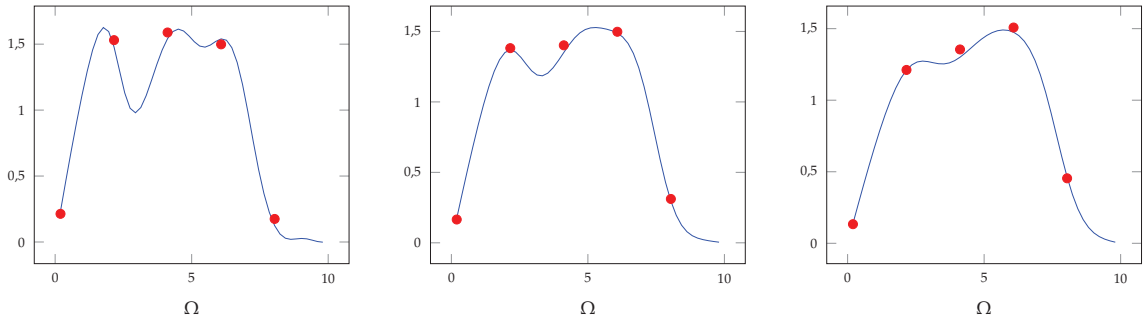
Tabla 2. Resumen experimento para el problema con la regularización TV

β	**Iteraciones**	$\| u - u_{ex} \|$
0.5	37	1.7938
0.75	31	1.9082
1	35	2.0705
5	60	3.5346

Tabla 3. Resumen experimento para el problema con regularización TGV

α	β	**Iteraciones**	$\| \mathbf{u} - \mathbf{u_{ex}} \|$
1	$(0{,}75)\eta$	13	2.0003
1	η	15	1.9522
1	$(1{,}25)\eta$	14	1.9231
1	$(1{,}5)\eta$	12	1.9082
5	$(0{,}75)\eta$	15	1.9794
5	η	15	2.0505
5	$(1{,}25)\eta$	14	2.1398
5	$(1{,}5)\eta$	12	2.2171

Como se mencionó en la introducción de la sección este experimento nos permite comparar



(a) Variable de estado para $t = 0{,}13$

(b) Variable de estado para $t = 0{,}255$

(c) Variable de estado para $t = 0{,}38$

Figura 9. Comparación entre el estado final y las observaciones para el experimento con la regularización TGV con parámetros $\alpha = 1$ y $\beta = 0{,}03$

de manera más concreta las soluciones obtenidas por ambos problemas. De manera concreta, mostramos el efecto de escalonamiento en el caso de la regularización TV y la manera en la que la regularización TGV elimina este efecto y nos permite obtener una mejor reconstrucción de la solución.

Finalmente, presentamos en la Figura 9 el estado óptimo asociado a la solución del problema con la regularización TGV. Podemos apreciar la manera en la que la variable de estado se ajusta a las observaciones de manera óptima, permitiéndonos de esta manera concluir que al aumentar el tamaño de la ventana de tiempo los pronósticos serán mejores que en el caso de no usar esta regularización.

3. Experimentos con mallas más finas

La principal desventaja que presenta el método de Newton Globalizado dado en los Algoritmos 5 y 6 utilizado en este trabajo es el tamaño de la matriz resultante. En el caso del problema con la regularización de variación total el tamaño de la matriz es de $2m + 2n - 1$ donde n es la cantidad de puntos de la discretización espacial y N_t de la discretización temporal y $m = nN_t$. Mientras, en el caso del problema con la regularización TGV el tamaño de la matriz es de $2m + 4n - 3$. En el caso de los experimentos anteriores se tomó $n = 50$ y $N_t = 100$, por tanto las matrices correspondientes eran de tamaño 10099 para el problema TV y 10197 en el caso del problema TGV. A continuación vamos a mostrar que contar con mallas más finas nos garantiza una mejor reconstrucción de la solución. En este caso vamos a resolver un problema de asimilación de datos con la ecuación de Burgers con la regularización TGV únicamente. En este experimento vamos a fijar la solución exacta como la función que se muestra en la figura 10.

Tabla 4. Resumen del experimento para diferentes tamaños de mallas

n	N_t	**Tamaño de la matriz**	α	β	**Iteraciones**	$\| u - u_{ex} \|$	**Tiempo**
50	100	10197	1	0.03	12	1.9082	9.37s.
75	150	22797	5	0.05	14	2.4725	38.35s.
100	200	40397	5	0.0375	12	2.4481	86.79s.
100	300	60397	5	0.0375	11	2.4146	118.76s.
150	300	90597	5	0.0250	11	2.6144	244.59s.

$$u_{ex} = \begin{cases} \frac{3x}{2} & \text{si } x \in [0;2], \\ x-2 & \text{si } x \in [2;4], \\ 2 & \text{si } x \in [4;5], \\ 3x-15 & \text{si } x \in [5;6], \\ \frac{-3x}{4}+6 & \text{si } x \in [6;8], \\ 0 & \text{si } x \in [8;10], \end{cases}$$

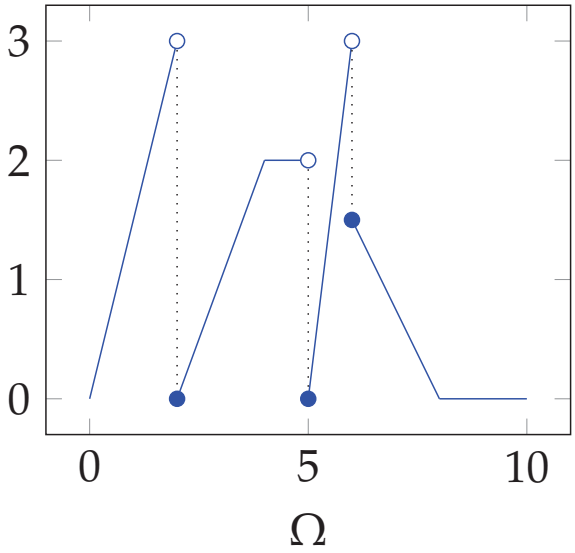
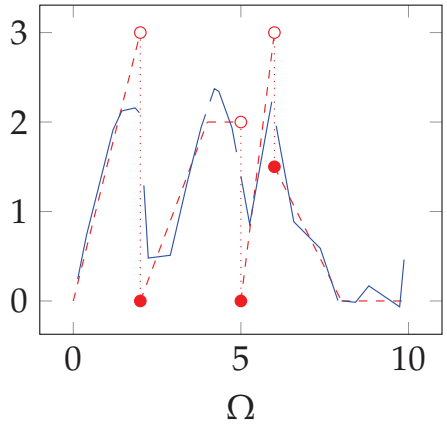


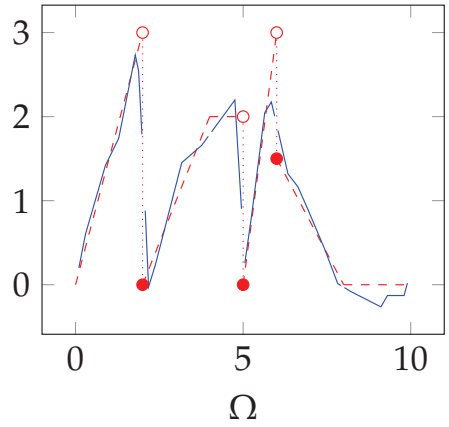
Figura 10. Función exacta para el experimento resolución del problema con distintos tamaños de malla

Debido al tamaño de las matrices este experimento fue realizado usando los recursos del Laboratorio de Cálculo Científico del Centro de Modelización Matemática: ModeMat, Escuela Politécnica Nacional (Quito). Para cada tamaño de malla se realizaron varios experimentos variando los parámetros α y β, sin embargo, en la Tabla 4 se muestran únicamente los mejores resultados.

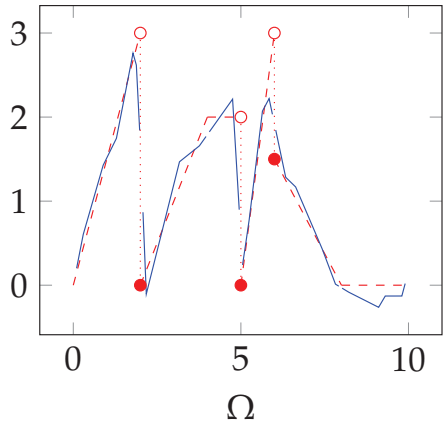
La Figura 11 muestra los gráficos de las soluciones obtenidas para cada tamaño de malla, en el cual se puede apreciar la diferencia en las soluciones cuando tenemos mallas más finas. A pesar de que en la Tabla 4 la diferencia entre las soluciones obtenidas y la exacta va aumentando conforme el tamaño de la malla aumenta, en la Figura 11 se puede apreciar la manera en la que la soluciones obtenidas se van ajustando de mejor manera a la exacta. Este ajuste ocurre particularmente cerca de los puntos de discontinuidad de la función, es decir se recuperan los *sharp fronts* de mejor manera.



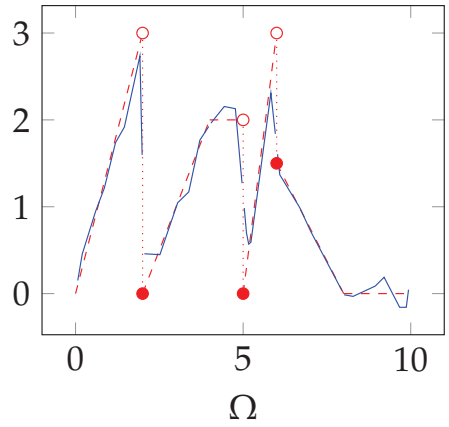
(a) Solución del problema con $n = 75$ y $N_t = 150$



(b) Solución del problema con $n = 100$ y $N_t = 200$



(c) Solución del problema con $n = 100$ y $N_t = 300$



(d) Solución del problema con $n = 150$ y $N_t = 300$

Figura 11. Soluciones obtenidas para el experimento con distintos tamaños de mallas

4. Análisis de la influencia en la cantidad de observaciones

Esta sección está dedicada a mostrar la diferencia en la reconstrucción de las soluciones dependiendo de la cantidad de observaciones que se ingresan como datos en el problema de asimilación de datos. La principal motivación de este experimento es mostrar que los pronósticos o en este caso la reconstrucción de las soluciones puede ser mejorada considerablemente únicamente al incluir más observaciones como datos de entrada. De esta manera, podemos mostrar que en el caso del problema de asimilación de datos meteorológicos si contamos con un sistema de recolección de información de observaciones meteorológicas más amplio a nivel nacional podría mejorar los pronósticos climáticos reales.

Como se mencionó en los experimentos anteriores se utilizaron únicamente 5 observaciones en el espacio en 4 instantes de tiempo, dándonos un total de 20 observaciones de 500 posibles (ya que tenemos 50 puntos de discretización espacial y 100 de discretización temporal). Para este ejemplo se usará como función exacta la descrita por la siguiente expresión:

$$u_{ex} = \begin{cases} \dfrac{3x}{2} & \text{si } x \in [0;2], \\ x-2 & \text{si } x \in [2;4], \\ 2 & \text{si } x \in [4;5], \\ 3x-15 & \text{si } x \in [5;6], \\ \dfrac{-3x}{4}+6 & \text{si } x \in [6;8], \\ 0 & \text{si } x \in [8;10], \end{cases}$$

En la Tabla 5 se muestra las iteraciones y el error ($\| u - u_{ex} \|$) al variar la cantidad de observaciones ingresadas. El problema se resolverá con la regularización TGV con los parámetros $\alpha = 1$ y $\beta = 0{,}03$. Esta tabla muestra además, la cantidad de observaciones en la malla espacial que se tomará (n_o) y cuantos instantes de tiempo se van a considerar (N_o). Como se puede apreciar en esta tabla, para poder reconstruir las soluciones de mejor manera es preferible aumentar la cantidad de observaciones espaciales que los instantes de tiempo. Esto se muestra directamente en la primera parte de la tabla, en donde solo aumentamos la cantidad de instantes de tiempo a ser considerado. Sin embargo, el error no disminuye como se esperaría. Por otro lado, la segunda parte de la tabla nos muestra que a pesar de considerar pocos instantes de tiempo la reconstrucción de la soluciones puede ser mejorada considerablemente únicamente aumentando la cantidad de observaciones espaciales. Finalmente, como se esperaría al contar con información completa se obtiene la mejor reconstrucción; sin embargo, es claro que desde el punto de vista aplicado esto no es posible. En conclusión, gracias a este experimento sabemos que contar con más observaciones distribuidas en la malla espacial mejora considerablemente los pronósticos.

Además, en la Figura 12 se muestran los gráficos de algunas de las soluciones dependiendo

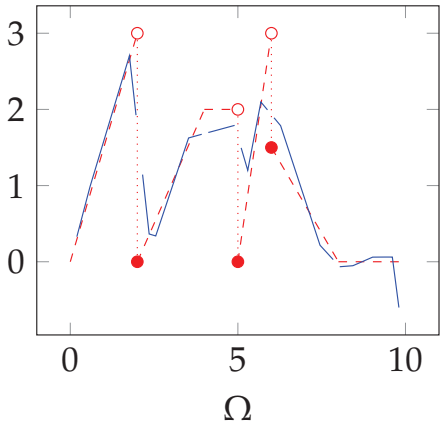
Tabla 5. Resumen del experimento para diferentes cantidades de observaciones

n_o	$(N_t)_o$	**iteraciones**	$\| u - u_{ex} \|$
3	3	11	2.3684
3	4	12	2.3756
3	5	11	2.3809
3	7	13	2.3819
3	10	14	2.3853
3	20	21	2.4042
3	25	23	2.4128
3	50	59	2.4370
5	3	11	2.2572
10	3	15	2.0955
25	3	11	1.9992
50	3	11	0.5530
5	5	13	2.2003
10	10	14	1.8598
25	25	19	1.5750
50	50	19	0.3638
50	100	20	0.2474

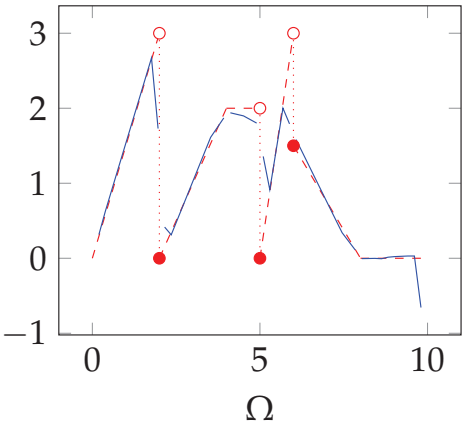
de la cantidad de observaciones que se utilizan.

Otro factor importante dentro de la asimilación de datos es la calidad de las observaciones ingresadas al problema. A lo largo de este capítulo hemos asumido que las observaciones son perfectas, es decir coinciden directamente con la solución de la ecuación de estado asociada con la solución exacta. Sin embargo, desde el punto de vista aplicado es imposible obtener este tipo de observaciones. En particular, para el problema de asimilación de datos meteorológicos los artefactos utilizados para la recolección de observaciones incurren en errores los cuales no pueden ser despreciados. Con la finalidad de reproducir este comportamiento la siguiente parte del experimento muestra que contar con observaciones imperfectas puede afectar notablemente los resultados obtenidos. En el experimento vamos a fijar $n_o = 25$ y $N_o = 25$ y analizaremos las soluciones obtenidas para el problema con la regularización TGV al contar con observaciones perfectas e imperfectas. En la Figura 13 se muestran las soluciones obtenidas al tener observaciones perfectas e imperfectas. Para la generación de la observaciones imperfectas tomamos las observaciones perfectas y le añadimos ruido gaussiano de media cero y matriz de covarianza $R = 10\mathbb{I}$. Ambos problemas se resuelven con la regularización TGV con los parámetros $\alpha = 5$ y $\beta = 0{,}15$.

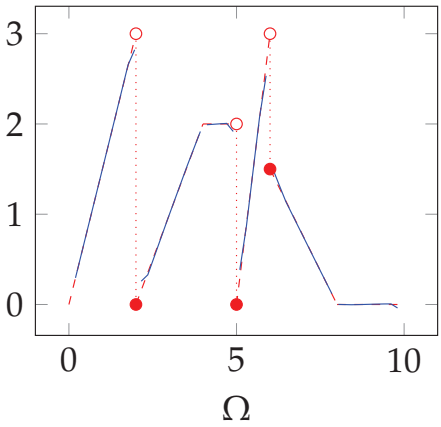
De la Figura 13 se puede concluir la importancia que tiene contar con observaciones que no posean muchos errores ya que estos errores influyen directamente en la correcta reconstrucción de la solución.



(a) Solución con $n_o = 3$ y $N_o = 10$

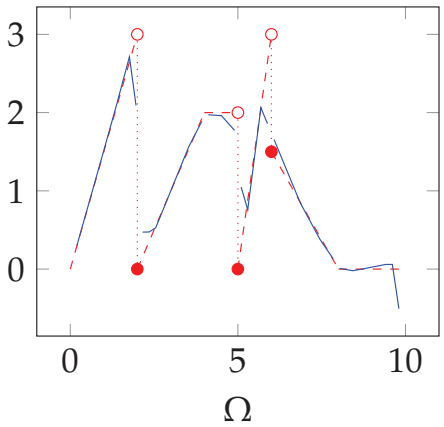


(b) Solución con $n_o = 25$ y $N_o = 3$

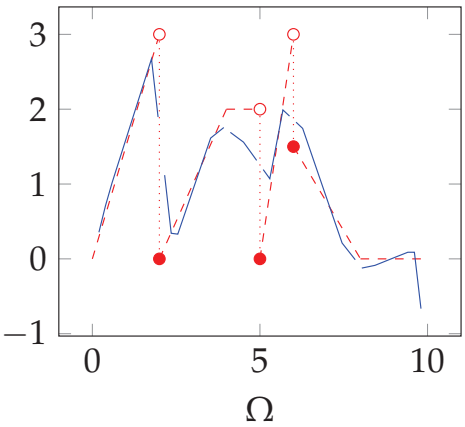


(c) Solución con $n_o = 50$ y $N_o = 100$

Figura 12. Soluciones obtenidas al variar la cantidad de observaciones



(a) Solución con observaciones perfectas



(b) Solución con observaciones imperfectas

Figura 13. Soluciones obtenidas al tener observaciones perfectas e imperfectas

5. Problema de asimilación de datos con el ecuación de Burgers sin viscosidad

De la teoría desarrollada en [Kreiss and Lorenz, 1989] sabemos que si el parámetro de viscosidad en la ecuación de Burgers es cero, no se puede garantizar mediante una formulación débil la existencia de soluciones continuas para el problema. La existencia de soluciones para el problema sin viscosidad se garantiza tomando el limite cuando $\nu \to 0$ para la soluciones del problema clásico de Burgers con viscosidad. Si u_ν es la solución de la ecuación clásica de Burgers asociada al parámetro de viscosidad ν entonces la solución de la ecuación de Bugers sin viscosidad u es $u_\nu \to u$ cuando $\nu \to 0$. El principal objetivo de este experimento es mostrar que usando un esquema de *discretizar–luego–optimizar* nos permite resolver el problema de optimización usando la ecuación de Burgers sin viscosidad. La particularidad de esta ecuación radica en la falta de garantía en la existencia de soluciones a partir de un análisis en espacios funcionales. En nuestro caso, puesto que se ha demostrado la existencia de las soluciones para el problema discretizado en todos los casos, sin que este resultado dependa del valor del parámetro ν podemos resolver los problemas sin mayor inconveniente. A continuación presentamos las soluciones obtenidas cuando hacemos tender a cero el parámetro ν.

Para este experimento vamos a tomar la función exacta definida en la siguiente expresión

$$u_{ex} = \begin{cases} \dfrac{3x}{2} & \text{si } x \in [0;2], \\ x-2 & \text{si } x \in [2;4], \\ 2 & \text{si } x \in [4;5], \\ 3x-15 & \text{si } x \in [5;6], \\ \dfrac{-3x}{4}+6 & \text{si } x \in [6;8], \\ 0 & \text{si } x \in [8;10], \end{cases}$$

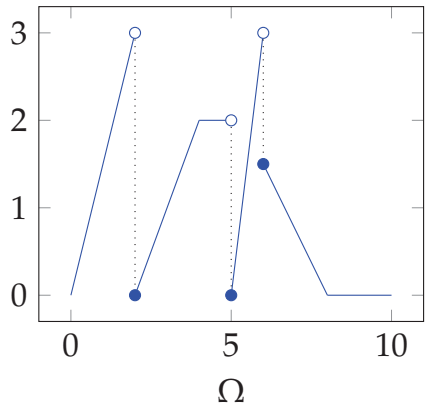
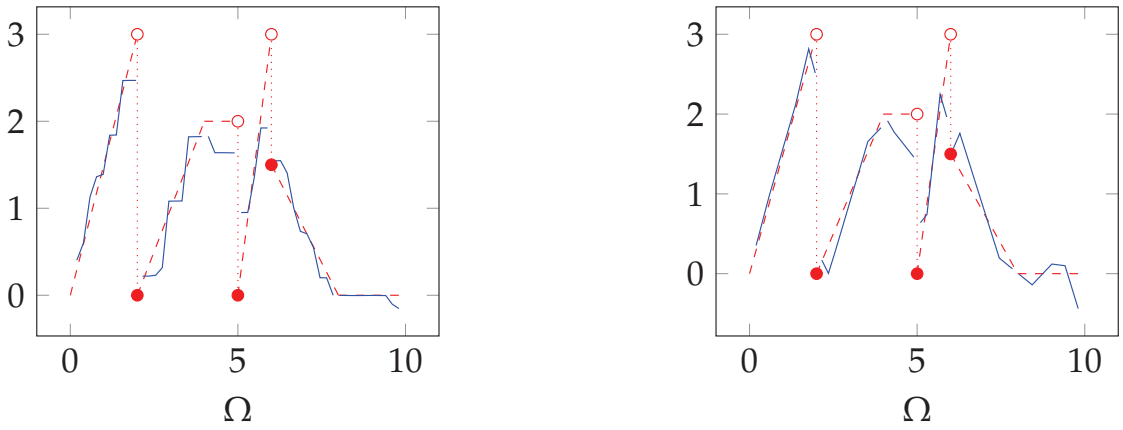


Figura 14. Función exacta para el experimento del problema de asimilación de datos con la ecuación de Burgers sin viscosidad

Resolvemos el problema con la regularización de variación total y la regularización de variación total generalizada. Usando los siguientes parámetros $\beta = 1$ para la regularización TV y $\alpha = 2{,}5$ and $\beta = 0{,}075$ para la regularización TGV. En la Tabla 6 se muestran los resultados obtenidos al variar el parámetro ν. La Figura 15 muestra las soluciones obtenidas al tomar exactamente $\nu = 0$. De la Tabla 6 podemos concluir que el algoritmo resolvió el problema sin ningún tipo de dificultad, y además se obtuvo una muy buena reconstrucción de la solución en

Tabla 6. Resumen del experimento con $\nu \to 0$

	TV		TGV	
ν	**iter**	$\| u - u_{ex} \|$	**iter**	$\| u - u_{ex} \|$
0.6	36	1.9714	14	1.9168
0.01	41	1.6577	31	1.6539
1e-4	41	1.6518	30	1.6205
1e-8	41	1.6518	30	1.6201
0	41	1.6518	30	1.6201



(a) Solución del problema con la regularización TV

(b) Solución del problema con la regularización TGV

Figura 15. Soluciones obtenidas para el experimento con la ecuación de Burgers sin viscosidad

el caso de la regularización TGV. Es decir, estos resultados son consistentes con los anteriores, a pesar de que, como se mencionó al inicio de la sección, la ecuación de estado de este problema no necesariamente tiene solución desde el análisis en espacios funcionales.

Capítulo 6

Conclusiones

El problema de asimilación de datos es un problema inverso mal condicionado, por lo cual necesita de un tratamiento especial para su solución. En el caso de la asimilación de datos meteorológicos se espera que las soluciones sean funciones discontinuas con *sharp fronts*. Teniendo en cuenta esta información a priori se deben proponer regularizaciones que garanticen conservar estas características en las soluciones.

Los problemas de tratamiento de imágenes, otro tipo de problemas inversos mal condicionados, utilizan regularizaciones como la de variación total y variación total generalizada para reconstruir las imágenes obteniendo muy buenos resultados. Con estas ideas en mente en el artículo [Freitag et al., 2010] se propone usar la regularización de variación total para mejorar la reconstrucción de soluciones con *sharp fronts*. El principal aporte de este trabajo fue mostrar que en el caso de tener funciones lineales a trozos en un problema de asimilación de datos el uso de la regularización de variación total generalizada genera mejores soluciones y además elimina el efecto de escalonamiento producido por la regularización de variación total. Este efecto ha sido ampliamente estudiado en el campo del tratamiento de imágenes, y la regularización de variación total generalizada [Bredies et al., 2010] fue presentada como una solución a este efecto.

Los problemas inversos pueden ser tratados desde dos enfoques diferentes, el enfoque variacional y el enfoque estadístico. En el capítulo 2 se mostró que los problemas con regularizaciones de variación total y variación total generalizada pueden ser obtenidos a partir de una técnica bayesiana para obtener el estimador máximo a posteriori (MAP), dando de esta manera un enfoque estadístico a nuestros resultados. Además, se mostró resultados ampliamente conocidos sobre los problemas 3D-VAR y 4D-VAR usuales con técnicas de estimación de máxima verosimilitud y estimación bayesiana. El resto del trabajo se concentró en el enfoque variacional para la solución del problema.

El principal reto que presentan los problemas que involucran regularizaciones de variación total y variación total generalizada es la solución numérica. En primer lugar, debido a que la mayoría de los algoritmos de optimización asumen diferenciabilidad de la función objetivo, es

necesario modificarla usando funciones diferenciables que se aproximen a la función original. Debido a que el término no diferenciable de la función objetivo es el valor absoluto, se propuso usar la regularización de Huber $\mathcal{C}^1$ y $\mathcal{C}^2$ dependiendo del tipo de algoritmo que se usó. La principal ventaja de estas regularizaciones es que son de carácter local y modifican la función únicamente en los puntos de conflicto. Los problemas que involucran la regularización de Huber, en cualquiera de sus dos versiones, se los denomina problemas γ–regularizados. Una vez que hemos modificado las funciones objetivo de los problemas, garantizamos que efectivamente las soluciones de los problemas γ–regularizados convergen a las soluciones de los problemas originales y este resultado se mostró en el capítulo 3 usando técnicas ampliamente conocidas en el campo de la optimización. Finalmente se derivaron las condiciones de optimalidad de primer orden para el problema con la regularización de variación total y la regularización de variación total generalizada.

El capítulo 4 se concentró en el estudio de los diferentes algoritmos iterativos para la solución del problema de optimización. Debido a la naturaleza del problema se pensó que métodos como el del descenso más profundo o BFGS iban a funcionar bien en la solución del problema. Sin embargo, la experimentación numérica reveló un comportamiento totalmente contrario. Estos métodos eran muy inestables con respecto a los parámetros, muchas veces no satisfacían los criterios de convergencia, o las direcciones estaban fueran de la región de convergencia. Debido a este comportamiento fue necesario buscar nuevos métodos que nos garanticen un comportamiento robusto de los algoritmos de optimización. En nuestro caso, optamos por el uso de métodos tipo Newton que aparte de garantizar convergencia más acelerada, nos garantizaban efectivamente la convergencia a puntos estacionarios del problema. La modificación en la matriz de segundo orden denominada proyección es de vital importancia para la globalización del método. Estas ideas fueron tomadas de métodos de optimización desarrollados para la solución de problemas de tratamiento de imágenes. Además, se realizó un análisis de la convergencia de cada uno de los algoritmos presentados en este trabajo. Para esto, usamos resultados presentados en [De los Reyes, 2015] los cuales garantizan la convergencia de la sucesión generada por el método hacia puntos estacionarios del problema. Para dicho efecto se mostró la continuidad uniforme del gradiente del funcional reducido con la finalidad de mostrar la factibilidad de los pasos de descenso, la cual se asume verdadera en el resultado de convergencia.

Finalmente, en el capítulo de experimentos numéricos se mostró que la regularización de variación total efectivamente presenta un efecto de escalonamiento en las soluciones cuando se tratan de soluciones lineales a trozos. Además, se muestra que la regularización de variación total generalizada reconstruyó de mejor manera las soluciones y eliminó el efecto de escalonamiento. En uno de los experimentos mostramos la importancia de contar con mallas más finas para una correcta reconstrucción de las soluciones en especial para recuperar las soluciones cerca de los puntos de discontinuidad de las funciones. El principal inconveniente de tener mallas finas en la solución del problema es el tamaño de los sistemas que tienen que resolverse, sin embargo este problema puede ser evitado usando métodos de paralelización algebráica.

Además, mostramos un experimento en el que analizamos las soluciones obtenidas al variar la cantidad de observaciones con las que contamos. Este experimento fue realizado con la principal motivación de mostrar que la cantidad de observaciones distribuidas en la malla espacial es de vital importancia para la correcta reconstrucción de la solución. Con estos resultados podemos asegurar que en el caso de la asimilación de datos meteorológicos el comportamiento será parecido, y de esta manera alentar a las autoridades competentes en la adquisición de artefactos para la recopilación de observaciones meteorológicas. Finalmente, se mostró también que con estas técnicas se pueden resolver problemas que en espacios de dimensión infinita serían muy difíciles, en particular, nos referimos a la solución del problema de asimilación de datos con la ecuación de Burgers sin viscosidad, mostrando que el desempeño del método es igual que en el caso de la ecuación con viscosidad y cuando dicho parámetro tiende a cero.

Bibliografía

[Bredies et al., 2013] Bredies, K., Dong, Y., and Hintermüller, M. (2013). Spatially dependent regularization parameter selection in total generalized variation models for image restoration. *International Journal of Computer Mathematics*, 90(1):109–123.

[Bredies et al., 2010] Bredies, K., Kunisch, K., and Pock, T. (2010). Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526.

[Bredies and Valkonen, 2011] Bredies, K. and Valkonen, T. (2011). Inverse problems with second-order total generalized variation constraints. *Proceedings of SampTA*, 201.

[Calatroni et al., 2015] Calatroni, L., Chung, C., Reyes, J. C. D. L., Schönlieb, C.-B., and Valkonen, T. (2015). Bilevel approaches for learning of variational imaging models. *arXiv preprint arXiv:1505.02120*.

[De los Reyes, 2015] De los Reyes, J. (2015). *Numerical PDE-Constrained Optimization*. SpringerBriefs in Optimization. Springer International Publishing.

[Dennis Jr and Schnabel, 1996] Dennis Jr, J. E. and Schnabel, R. B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*, volume 16. Siam.

[Elaydi, 2005] Elaydi, S. (2005). *An Introduction to Difference Equations*. Undergraduate texts in mathematics. Springer, 3rd ed edition.

[Freitag et al., 2010] Freitag, M., Nichols, N., and Budd, C. (2010). Resolution of sharp fronts in the presence of model error in variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 139(672):742–757.

[Griebel et al., 1998] Griebel, M., Dornseifer, T., and Neunhoeffer, T. (1998). *Numerical simulation in fluid dynamics: a practical introduction*. SIAM.

[Hintermüller and Stadler, 2006] Hintermüller, M. and Stadler, G. (2006). An infeasible primal-dual algorithm for total bounded variation–based inf-convolution-type image restoration. *SIAM Journal on Scientific Computing*, 28(1):1–23.

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

[Keller, 1976] Keller, J. B. (1976). Inverse problems. *The American Mathematical Monthly*, 83(2):107–118.

[Knoll et al., 2011] Knoll, F., Bredies, K., Pock, T., and Stollberger, R. (2011). Second order total generalized variation (tgv) for mri. *Magnetic resonance in medicine*, 65(2):480–491.

[Kreiss and Lorenz, 1989] Kreiss, H.-O. and Lorenz, J. (1989). *Initial-boundary value problems and the Navier-Stokes equations*, volume 47. Siam.

[Lee and Kitanidis, 2013] Lee, J. and Kitanidis, P. (2013). Bayesian inversion with total variation prior for discrete geologic structure identification. *Water Resources Research*, 49(11):7658–7669.

[Lewis et al., 2006] Lewis, J. M., Lakshmivarahan, S., and Dhall, S. (2006). *Dynamic data assimilation: a least squares approach*, volume 13. Cambridge University Press.

[Quarteroni et al., 2010] Quarteroni, A., Sacco, R., and Saleri, F. (2010). *Numerical mathematics*, volume 37. Springer Science & Business Media.

[Ulbrich and Ulbrich, 2012] Ulbrich, M. and Ulbrich, S. (2012). *Nichtlineare Optimierung*. Springer-Verlag.

[Volkwein, 1997] Volkwein, S. (1997). *Mesh-independence of an augmented Lagrangian-SQP method in Hilbert spaces and control problems for the Burgers equation*. PhD thesis, Technische Universität Berlin.

[Wright and Nocedal, 1999] Wright, S. and Nocedal, J. (1999). *Numerical optimization*. Springer Science.