TOTAL GENERALIZED VARIATION REGULARIZATION IN DATA ASSIMILATION FOR BURGERS’ EQUATION

JUAN CARLOS DE LOS REYES
Research Center of Mathematical Modelling (MODEMAT)
Escuela Politécnica Nacional
Quito, Ecuador

ESTEFANÍA LOAYZA-ROMERO
Faculty of Mathematics
Professorship Numerical Mathematics (Partial Differential Equations)
Chemnitz University of Technology
Chemnitz, Germany

(Communicated by Otmar Scherzer)

ABSTRACT. We propose a second–order total generalized variation (TGV) regularization for the reconstruction of the initial condition in variational data assimilation problems. After showing the equivalence between TGV-regularization and a Bayesian MAP estimator, we focus on the detailed study of the inviscid Burgers’ data assimilation problem. Due to the difficult structure of the governing hyperbolic conservation law, we consider a discretize–then–optimize approach and rigorously derive a first–order optimality condition for the problem. For the numerical solution, we propose a globalized reduced Newton–type method together with a polynomial line–search strategy, and prove convergence of the algorithm to stationary points. The paper finishes with some numerical experiments where, among others, the performance of TGV-regularization compared to TV-regularization is tested.

1. Introduction. Data assimilation can be described as the process in which one aims to find an estimate of the state of a system using observations and background information, in order to obtain improved numerical forecasts of the system. The resulting inverse problems are typically ill–posed, since one seeks a solution using only partial observations of the state. As a remedy, regularization approaches such as filtering, a posteriori regularization, variational regularization, etc., are usually considered [17, 18, 21].

In this paper, we focus our attention on the variational data assimilation framework, whose main idea consists in solving an optimization problem that fits the observations, on the one hand, and uses background information of previous forecasts, on the other. The background term usually plays the role of a regularizer, if one compares with an inverse problems methodology. Classically, variational data assimilation problems arise from a robust Bayesian estimation of the initial condition [14, 17]. If the observations are taken at one sole time instant, one obtains a...
so-called 3D-Var problem, while if a whole time window is considered, the so-called 4D-Var problem is obtained.

A generic mathematical formulation of the 4D-Var data assimilation problem consists in solving the least squares formulation:

\[
\begin{aligned}
& \min_u \frac{1}{2} \sum_{i=1}^{N} (z_i - H_i(y_i))^T R_i^{-1} (z_i - H_i(y_i)) + \frac{1}{2} (u - u_b)^T B^{-1} (u - u_b) \\
\text{subject to:} & \\
& y_i = u, \quad i = 1, \\
& y_{i+1} = \mathcal{M}_{i+1}(y_i), \quad \forall i = \{1, \ldots, N-1\},
\end{aligned}
\]

where \( N \) is the number of time instants at which the observations are taken and each vector \( z_i \) consists of the observations spread over the spatial mesh at the \( i \)-th time instant. Furthermore, \( B \) is the covariance matrix of the background error, \( R_i \) the covariance matrix for the observational error, the operator \( H_i(\cdot) \) is the observation operator that maps the model state to the observation space for each instant of time and \( \mathcal{M}_i(\cdot) \) is the nonlinear state operator for each instant.

Although 4D-Var has been successfully tested and proved to be of operational utility in Numerical Weather Prediction (NWP), there are some shortcomings of the standard approach (1). These are related to the fact that the system state may present discontinuities, which are poorly reconstructed with the standard methodology. In fact, discontinuities in the form of so-called sharp fronts, are of key importance in Numerical Weather Prediction. With this in mind, alternative regularizations such as Total Variation (TV) have been previously considered in [5, 11], where a comparison with respect to more standard regularizations has been carried out with promising results.

Total variation, however, has its downside too, since it performs well mainly when the reconstructed solution has a piecewise constant structure, something that cannot, in general, be expected in, e.g., fluid dynamics. To overcome the piecewise constant reconstruction effect, higher order regularizers have been considered in recent years, with the total generalized variation (TGV) as one of their main representatives [4, 3, 15, 2]. The second–order total generalized variation regularization depends on two positive parameters \( \alpha > 0 \) and \( \beta > 0 \) and has the following structure:

\[
\mathcal{R}(u) = \min_{w \in \mathbb{R}^{n-1}} \alpha \sum_{i=1}^{n-1} |D_i u - w_i| + \beta \sum_{i=1}^{n-1} |E_i w|, \quad u \in \mathbb{R}^n,
\]

where the matrix \( D \) corresponds to the discrete gradient and \( E \) to the symmetrized discrete gradient.

In this paper, we investigate the use of second–order total generalized variation regularization for the solution of data assimilation problems constrained by the inviscid Burgers’ equation. This equation is a first model for turbulence and a standard toy problem for the dynamics of the atmosphere. Due to its hyperbolic conservation law structure, we adopt a discretize–then–optimize approach for the solution of the problem, to avoid complications related to the well–posedness of the equation in function spaces [19]. A similar approach has been previously pursued in [1].

Specifically, we consider the problem
\[
\begin{array}{l}
\min \frac{1}{2} \sum_{i=1}^{N} (z_i - H_i(y_i))^T R^{-1}_i (z_i - H_i(y_i)) + \frac{1}{2} (u - u^b)^T B^{-1} (u - u^b) + R(u)
\end{array}
\]

subject to the discretized Burgers’ equation,

where \( R(\cdot) \) is the regularization term given in (2). For problem (3) we analyze the well-posedness of the discretized equation, existence of optimal solutions, well-posedness of the adjoint equation and the characterization of local minima by means of a first order optimality system of Karush–Kuhn–Tucker type.

The most challenging aspect concerns the numerical solution of the problem. In order to devise a fast superlinear convergent algorithm, we consider a Huber regularization of the total generalized variation and start by verifying the consistency of the approximation. Each regularized problem is solved by means of a modified primal–dual formulation of the optimality system. The modification concerns the second–order iteration matrix, which in its original form does not guarantee descent in each iteration. The global convergence of the resulting algorithm is demonstrated.

The organization of the paper is as follows. Section 2 is devoted to discuss the equivalence of the TGV regularized problem with a Bayesian MAP estimator. In Section 3 we analyze the discretization of the Burgers’ equation as well as the corresponding data assimilation problem, obtaining a necessary first-order optimality condition for the optimization problem. In Section 4 a Huber regularized version of the data assimilation problem is introduced and the consistency of the regularization proved. The next section is devoted to the presentation of a globalized Newton-type method considering dual variables and projections, while its convergence study is carried out in Section 6. In the last section, we present numerical experiments that reveal the improvement in the recovery of the solutions when we use the TGV-regularization and test the performance of the proposed algorithm.

2. Maximum A Posteriori estimation. This section aims to show that problem (3) has a statistical interpretation and can be derived using a Bayesian approach to obtain the Maximum A Posteriori (MAP) estimator. With this purpose, let us start by writing the TGV regularized data assimilation problem in generic form as:

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

with \( x = (u, w), \bar{z} = (z, u^b), E(x) = E(u, w) = (H(M(u)), u) \) and the matrices

\[
G^{-1} = \begin{bmatrix}
R^{-1} & 0 \\
0 & B^{-1}
\end{bmatrix}
\quad \text{and} \quad
\mathbb{D} = \begin{bmatrix}
\alpha D & -\alpha I \\
0 & \beta E
\end{bmatrix}.
\]

As mentioned in the introduction, we expect the solutions of the data assimilation problem to allow discontinuities. This information is in fact what we know a priori about the solution, and we use it for the MAP estimation. The next lines are developed following the ideas in [16].

In the Bayesian method, unlike the maximum likelihood estimation, we assume that the unknown is a random variable, and we know a priori its probability density function \( p(x) \) and the conditional probability density function (c.p.d.f.) of the observations \( \bar{z} \) given \( x \), denoted by \( p(\bar{z}|x) \). The basic idea of this approach is to combine the information from both probability density functions using Bayes’
formula in order to obtain an \textit{a posteriori} probability. In this case, it is assumed that the observations and background information have normally distributed errors with covariance matrices $R$ and $B$, respectively. Therefore, the conditional probability density function (c.p.d.f.) is given by the following Gaussian distribution:

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

where $m_o$ is the total amount of observations.

Now, we need to look for a probability distribution that enables to capture the expected main features of the solutions. Since these solutions are expected to be discontinuous, previous works suggest using the Laplace distribution [5, 16]. Specifically, it is assumed that the solution is a realization of a random process defined by the Laplace prior probability density function for the matrix $D_x$, with parameters $\theta = 1$ and $\mu = 0$, i.e.,

\[
p(x) = \frac{1}{2\theta} \exp \left\{ \frac{1}{\theta} \sum_{i=1}^{2n-1} |D_i x - \mu| \right\}.
\]

Using Bayes’ formula we get that the \textit{a posteriori} probability of $x$ given the observations $\bar{z}$ is

\[
p(x|\bar{z}) = \frac{p(\bar{z}|x)p(x)}{p(\bar{z})}.
\]

Since we aim to find the best “Bayesian estimator” using the maximum a posteriori approach, we maximize the previously obtained \textit{a posteriori} probability

\[
\textup{max } x \ p(x|\bar{z}) = \textup{max } x \ \frac{p(\bar{z}|x)p(x)}{p(\bar{z})}.
\]

The distribution $p(\bar{z})$ is the marginal probability and it does not depend on $x$. Thus, we optimize

\[
\textup{max } x \ p(\bar{z}|x)p(x).
\]

Since the logarithmic function is monotonically increasing and continuous, solving (5) is equivalent to maximize the natural logarithm of the objective function, that is

\[
\textup{max } x \ \ln(p(\bar{z}|x)) + \ln(p(x)).
\]

Replacing the definition of the probability density functions mentioned above, we get

\[
\textup{max } x \ - \frac{1}{2} (\bar{z} - \mathcal{E}(x))^T G^{-1}(\bar{z} - \mathcal{E}(x)) + \sum_{i=1}^{2n-1} |D_i x|
\]

and, consequently, the equivalence of the 4D-VAR problem with TGV-regularization.

3. \textbf{Analysis of the data assimilation problem.} This section is devoted to the analysis of the data assimilation problem for the Burgers’ equation. We start with the discretization of the state system using a finite difference scheme for the spatial discretization and a semi–implicit method for the time discretization. The existence of solutions to the fully discretized equation is justified thereafter using the \textit{implicit function theorem}. The same technique is applied to justify existence of an adjoint state and derive a first-order optimality condition of Karush–Kuhn–Tucker type. This approach is known as \textit{discrete–then–optimize}.
3.1. Discretization of the Burgers’ equation. The Burgers’ equation was developed in 1939 as a simplified version of the Navier–Stokes equation in order to better understand turbulence. The equation has a nonlinear convective term that makes it difficult to solve. The inviscid one–dimensional equation is given by the following model:

\[
\begin{aligned}
\frac{\partial y}{\partial t} + y \frac{\partial y}{\partial x} &= g & \text{in } Q = \Omega \times (0, T), \\
y &= 0 & \text{on } \Gamma \times (0, T), \\
y(x, 0) &= u(x) & \text{in } \Omega.
\end{aligned}
\]

In order to solve the latter, both a spatial and a time discretization scheme have to be utilized. For the spatial discretization several methodologies, like finite differences, finite elements or finite volumes, have been classically considered. Here we focus on a finite differences scheme with a homogeneous partition of the domain \(\Omega = (0, 1)\), whose discretization points are given by 

\[x_i = ih, \quad \forall i = 1, \ldots, n,\]

with \(h = 1/n+1\) and \(n\) the amount of spatial discretization points. For the approximation of the spatial first derivative of \(y_i\) we use a standard backward finite difference scheme, with the corresponding matrix denoted by \(U\).

For the time discretization, in order to guarantee convergence towards a (possibly) discontinuous solution, we focus on semi–implicit schemes. Moreover, to avoid the solution of the non–linearity in each time step, we consider a semi–implicit Euler scheme given by:

\[\frac{y^{j+1} - y^j}{\Delta t} + \text{diag}(y^j) U(y^{j+1}) = g^{j+1},\]

where \(\Delta t = 1/(N_t+1)\) and \(N_t\) is the number of time discretization points. Choosing this kind of discretization, the problem reduces to the solution of a linear system at each time discretization point, with the following structure:

\[y^{j+1} + \Delta t \text{diag}(y^j) U g^{j+1} = \Delta t g^{j+1} + y^j.\]

Considering both the spatial and time discretization, we can rewrite the matrix \(y \in \mathbb{R}^{n \times N_t}\) as a vector with the following structure

\[
y = \begin{bmatrix} y^1 \\ \vdots \\ y^{N_t} \end{bmatrix},
\]

where \(y^i \in \mathbb{R}^n\), thus \(y \in \mathbb{R}^m\) with \(m = n \cdot N_t\). In this context, we define the new matrices

\[E = \frac{1}{\Delta t} \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & U & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & U \end{bmatrix},
\]

\[Z = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & \text{diag}(y^1) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \text{diag}(y^{N_t-1}) \end{bmatrix}, \quad g(u) = \begin{bmatrix} u/\Delta t \\ g_2 \end{bmatrix}.
\]
The system that we have to solve is then given by:
\begin{equation}
\label{eq:state_equation}
e(y, u) = 0,
\end{equation}
where
\begin{align}
\label{eq:projection}
e & : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m \\
\label{eq:linearized_system}
y, u & \to Ey + Z(y)\bar{y} - g(u).
\end{align}

In order to prove the local existence of solutions of the state equation we will use the implicit function theorem (see, e.g., [7]). For this purpose, we consider hereafter the following assumption.

**Assumption 1.** For each \( j = 1, \ldots, N_t - 1 \),
\[ y_j^1 U_{ii} \geq 0, \quad \text{for all } i = 1, \ldots, n. \]

**Remark 1.** Let us notice that Assumption 1 is of importance not only for the analysis, but also for the numerical solution of both the state equation and the data assimilation problem. This hypothesis may be constructively guaranteed by using an upwinding scheme.

Computing the partial derivative of the state equation operator \( e \) with respect to the state, in direction \( \varphi \), we can write the linearized equation as follows:
\begin{align}
\label{eq:linearized_adjoint}
\varphi^1 & = \phi^1 \\
\label{eq:linearized_state}
\frac{1}{\Delta t} \varphi^{j+1} + \text{diag}(y^j)U(\varphi^{j+1} + \text{diag}(Uy^{j+1}))\varphi^j & = \phi^j, \quad \text{for } j = 1, \ldots, N_t - 1,
\end{align}
for \( \phi \in \mathbb{R}^m \). The equation can also be written in short form as
\begin{equation}
\label{eq:linearized_equation}
[E + ZU + \text{diag } (Uy)]\varphi = \phi.
\end{equation}
where \( \varphi \) is the vectorized form of \( \varphi \in \mathbb{R}^n \times N_t \).

**Proposition 1.** Under Assumption 1, the linearized state equation (10a)-(10b) has, for each \( \phi \in \mathbb{R}^m \), a unique solution \( \varphi \in \mathbb{R}^m \).

**Proof.** From equation (10a) \( \varphi^1 \) is uniquely defined. Given \( \varphi^j \), for some \( j \geq 1 \), the equation for \( \varphi^{j+1} \) is given by
\begin{equation}
\label{eq:recursive_equation}
\frac{1}{\Delta t} \varphi^{j+1} + \text{diag}(y^j)U\varphi^{j+1} = \phi^j - \text{diag}(Uy^{j+1})\varphi^j.
\end{equation}
Since by assumption \( y_j^1 U_{ii} \geq 0 \), for all \( i = 1, \ldots, n \), the matrix \( \left( \frac{1}{\Delta t} I + \text{diag}(y^j)U \right) \) is strictly diagonally dominant and, therefore, nonsingular. Hence, \( \varphi^{j+1} \) is uniquely determined. Proceeding recursively, the result is obtained. \[\square\]

**Theorem 3.1.** Let Assumption 1 hold. Let \( e(y, u) \) be the discretized Burgers' equation operator given by (9) and \( (y, \bar{y}) \) such that \( e(\bar{y}, \bar{u}) = 0 \). Then there exists a neighborhood \( \mathcal{U} \) of \( \bar{u} \) and a continuously differentiable implicit function \( y(u) \) such that \( e(y(u), u) = 0 \), for all \( u \in \mathcal{U} \).

**Proof.** Due to the definitions of the matrices \( E, Z, U \) we know that \( e(y, u) \) is continuously differentiable and its derivatives are given by
\[ e_y(y, u) = E + ZU + \text{diag } (Uy) \]
and \( e_u(y, u) = [-\frac{1}{\Delta t}I, 0, 0, \ldots, 0]^T \). Let \( (\bar{y}, \bar{u}) \) be such that \( e(\bar{y}, \bar{u}) = 0 \). From Lemma 1 we know that \( e_y(\bar{y}, \bar{u}) \) is invertible and, from the implicit function theorem, there exists an open neighborhood \( \mathcal{U} \) of \( \bar{u} \) and a continuously differentiable function \( y(u) \) such that \( e(y(u), u) = 0 \), for all \( u \in \mathcal{U} \). \[\square\]
In the next Lemma we prove the boundedness of $y$ with respect to the data.

**Lemma 3.2.** Let Assumption 1 hold and let $y$ be the solution of the state equation $e(y, u) = 0$. The following inequality holds

$$
\|y^j\| \leq \|u\| + \Delta t \sum_{i=1}^{j} \|g^i\| \quad \forall j = 1, \ldots, N_t.
$$

**Proof.** We start by recalling the state equation given by

$$
e(y, u) = Ey + Z(y)Uy - g(u) = 0,$$

where $y = (y_1, \ldots, y_{N_t})^T$ and $g(u) = (u/\Delta t, g^2, \ldots, g^{N_t})^T$. Performing a component-wise analysis we distinguish two cases:

For $j = 1$:

$$y^1 = u.$$

For $j = 2, \ldots, N_t$

$$y^j + \Delta t \text{diag}(y^{j-1})(U y^j) = \Delta t g^j + y^{j-1}.$$

Multiplying the previous equality with $(y^j)^T$ and using the hypothesis $y_i^T U_{ii} \geq 0$, for each $j = 1, \ldots, N_t - 1$ and all $i = 1, \ldots, n$, we get that

$$\|y^j\|^2 \leq (y^j)^T (I + \Delta t \text{diag}(y^{j-1})(U)) y^j \leq \|y^{j-1}\|^2 + \Delta t \|g^j\|^2 \|y^j\|,$$

which implies that

$$\|y^j\| \leq \|y^{j-1}\| + \Delta t \|g^j\|, \quad \text{for all } j = 1, \ldots, N_t.$$

Using the discrete Gronwall’s inequality we finally obtain

$$\|y^k\| \leq \|u\| + \Delta t \sum_{j=2}^{k} \|g^j\|, \quad \text{for all } k = 1, \ldots, N_t. \quad \square$$

### 3.2. Discretized data assimilation problem.

For the discretization of problem (3), we assume that the observation operator $H(\cdot)$ is linear and, therefore, there exist matrices $S \in \mathbb{R}^{m_r \times (n_o N_o)}$ and $H \in \mathbb{R}^{n_o N_o \times m}$ such that $H(x) = SHx$, with $m = n N_t$, $n_o$ the amount of observations taken in space, $N_o$ the number of instants at which we take the observations and $m_o = n_o N_o$. Thus, the problem is given by:

$$
\min_{(y,u,w) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^{n-1}} J(y, u, w) = \frac{1}{2} (SHy - z)^T R^{-1} (SHy - z) + \frac{\mu}{2} w^T w
$$

$$+ \frac{1}{2} (u - u^b)^T B^{-1} (u - u^b) + \alpha \sum_{i=1}^{n-1} |D_i u - w_i| + \beta \sum_{i=1}^{n-1} |E_i w|$$

subject to:

$$Ey + Z(y)Uy - g(u) = 0,$$

where the matrix $D$ corresponds to the discrete gradient associated with a forward finite differences scheme and $E$ is the one associated with backward finite differences. It is worth to remark that since we work with 1D–spatial functions the symmetrized discrete gradient match with the usual one. The Tikhonov term $\frac{\mu}{2} w^T w$ is further added to obtain well–posedness of the inverse problem.
3.3. Adjoint state. This subsection is devoted to discuss the main features of the adjoint equation. We start by noticing that the adjoint equation is given by the following expression:

\begin{equation}
E^T p + ZU^T p + \text{diag}(Uy)p - H^T S^T R^{-1}(SHy - z) = 0
\end{equation}

or, equivalently,

\[ e_y(y, u)^T p = H^T S^T R^{-1}(SHy - z). \]

Thanks to Proposition 1, which establishes the invertibility of the matrix \( e_y(y, u) \), existence of a unique solution to the adjoint equation (15) follows (see, e.g., [8]).

In the next Lemma we prove the boundedness of the adjoint state, by making use of the discrete Gronwall’s inequality [20].

**Lemma 3.3.** Let Assumption 1 hold and let \((y, p)\) be the state and adjoint state associated to some control \(u\). Let us moreover assume that there exists a constant \(c > 0\) such that

\begin{equation}
1 - \Delta t\|U\|\|u\| - |\Delta t|^2\|U\|\|g\| \geq c.
\end{equation}

Then the following inequality holds

\[ \|p\| \leq \rho\|H^T S^T R^{-1}(SHy - z)\|, \]

with \(\rho > 0\) a constant independent of \(u\) and \(y\).

**Proof.** Defining \(p = P\vec{v}\), with \(P\) a permutation matrix given by

\begin{equation}
P = \begin{bmatrix}
0 & 0 & \ldots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 1 & \ldots & 0 \\
1 & 0 & \ldots & 0 \\
\end{bmatrix}
\end{equation}

and \(I\) the identity matrix, we get that

\[ p^j = v^{N_t - j + 1} \]

and therefore the adjoint system (15), in the variable \(v\), is given by

\[ E^T P\vec{v} + ZU^T P\vec{v} + \text{diag}(Uy)P\vec{v} - H^T S^T R^{-1}(SHy - z) = 0. \]

Performing a component–wise analysis in time we distinguish the following cases:

- **For** \(j = 1:\)
  \[ \left( \frac{1}{\Delta t} I + \text{diag}(y^1) \right) v^{N_t} = \frac{1}{\Delta t} v^{N_t - 1} + \left[ H^T S^T R^{-1}(SHy - z) \right]_1 \]

- **For** \(j = 2, \ldots, N_t - 1:\)
  \[ \left( \frac{1}{\Delta t} I + \text{diag}(y^{N_t - j - 1})U + \text{diag} (Uy^j) \right) v^{N_t - j + 1} = \frac{1}{\Delta t} v^{N_t - j} + \left[ H^T S^T R^{-1}(SHy - z) \right]_j \]

- **For** \(j = N_t:\)
  \[ \left( \frac{1}{\Delta t} I + \text{diag}(y^{N_t - 1})U + \text{diag} (Uy^{N_t}) \right) v^1 = \left[ H^T S^T R^{-1}(SHy - z) \right]_{N_t}. \]
Multiplying by $v^{N_t-j+1}$ it then follows, for all $j = 1, \ldots, N_t - 1$, that
\[
\left( \frac{1}{\Delta t} \| + \text{diag}(y^{j-1}) U + \text{diag} (U y^{N_t}) \right) v^{N_t-j+1} \geq \frac{1}{\Delta t} \| v^{N_t-j+1} \|^2 - \| U \| \| y \| \| v^{N_t-j+1} \|^2 \geq \frac{1}{\Delta t} \| v^{N_t-j+1} \|^2 - \| U \| (\| u \| + \Delta t \| g \|) \| v^{N_t-j+1} \|^2,
\]
where the last inequality was obtained thanks to the bound (13).

Since by hypothesis there exists a constant $C > 0$ such that (16) holds, we get that
\[
\frac{c}{\Delta t} \| v^{N_t-j+1} \| \leq \frac{1}{\Delta t} \| v^{N_t-j} \| + \| \left[ H^T S^T R^{-1} (S H y - z) \right] \|.
\]
Finally, applying the discrete Gronwall’s inequality the result follows with $\rho = \frac{\Delta t}{c} \exp (1/c - 1)$. \hfill \Box

3.4. Optimality system. Next we derive formally a necessary optimality condition for problem (14) in form of a Karush–Kuhn–Tucker system.

**Theorem 3.4.** Let $\bar{u}$ be an optimal solution for problem (14) and let Assumption 1 hold. Then there exists an adjoint state $\bar{p} \in \mathbb{R}^m$ such that the following optimality system is satisfied:

\[
\begin{align*}
\mathbb{E} \bar{y} + Z \bar{U} \bar{y} - g(u) &= 0 \quad \text{(19a)} \\
\mathbb{E}^T \bar{p} + Z \mathbb{U}^T \bar{p} + \text{diag} (U \bar{y}) \bar{p} - H^T S^T R^{-1} (S H y - z) &= 0 \quad \text{(19b)} \\
\left( B^{-1} (\bar{u} - u^b) + \frac{1}{\Delta t} \beta^j, u - \bar{u} \right) + \alpha \sum_{i=1}^{n-1} |D_i u - w_i| - \alpha \sum_{i=1}^{n-1} |D_i \bar{u} - \bar{w}_i|
\end{align*}
\]

\[
+ (\mu \bar{w} - w - \bar{w}) + \beta \sum_{i=1}^{n-1} |E_i w| - \beta \sum_{i=1}^{n-1} |E_i \bar{w}| \geq 0, \quad \forall (u, w) \in \mathbb{R}^n \times \mathbb{R}^{n-1}.
\]

**Proof.** Due to the nondifferentiability of the objective function, we use [8, Thm. 6.1] in order to compute the optimality system of the problem. Thus, we define the functions
\[
\begin{align*}
j_1(u, w) &= \frac{1}{2} (S H y - z)^T R^{-1} (S H y - z) + \frac{1}{2} (u - u^b)^T B^{-1} (u - u^b) + \frac{\mu}{2} w^T w, \\
j_2(u, w) &= \sum_{i=1}^{n-1} |D_i u - w_i| + \sum_{i=1}^{n-1} |E_i w|.
\end{align*}
\]
We notice that the function $j_1(u, w)$ is differentiable and $j_2(u, w)$ is continuous and convex. Thus, using [8, Thm 6.1] we know that the optimality condition of the problem is given by the following variational inequality:

\[
(\nabla_j^1(\bar{u}, \bar{w}), (u, w) - (\bar{u}, \bar{w})) + j_2(u, w) - j_2(\bar{u}, \bar{w}) \geq 0, \quad \forall (u, w) \in \mathbb{R}^n \times \mathbb{R}^{n-1}.
\]
For the first term we have
\[
(\nabla j_1(\bar{u}, \bar{w}), (u, w) - (\bar{u}, \bar{w})) = (SH\bar{y} - z)R^{-1}SHy'(\bar{u})(u - \bar{u})
\]
\[
+ (\bar{u} - u)B^{-1}(u - \bar{u}) + \mu \bar{w}^T(w - \bar{w})
\]
\[
= (H^T S^T R^{-1}(SH\bar{y} - z), y'(\bar{u})(u - \bar{u}))
\]
\[
+ (B^{-1}(\bar{u} - u^b), u - \bar{u}) + (\mu \bar{w}, w - \bar{w}).
\]
\[
(21)
\]
Introducing now the adjoint state \(\bar{p}\) as the unique solution to (15) and replacing this in (21), we get
\[
(\nabla j_1(\bar{u}, \bar{w}), (u, w) - (\bar{u}, \bar{w})) = (B^{-1}(\bar{u} - u^b), u - \bar{u}) + (e_y(\bar{y}, \bar{u})^T \bar{p}, y'(\bar{u})(u - \bar{u}))
\]
\[
+ (\mu \bar{w}, w - \bar{w}).
\]
\[
(22)
\]
Furthermore, by differentiating (8), in the direction \(h\), we get the linearized equation
\[
e_y(y(\bar{u}), \bar{u})y'(\bar{u})h + e_u(y(\bar{u}), \bar{u})h = 0
\]
Using the adjoint operator of \(e_y(\bar{y}, \bar{u})\), the linearized equation (23) and the fact that
\[
e_u(y, u) = \left(-\frac{1}{\Delta t} \mathbb{I}, 0, \ldots, 0 \right)^T
\]
we get, together with (20), that
\[
\left( B^{-1}(\bar{u} - u^b) + \frac{1}{\Delta t} \bar{p}, u - \bar{u} \right) + (\mu \bar{w}, w - \bar{w}) + \alpha \sum_{i=1}^{n-1} |D_i u - w_i|
\]
\[
- \alpha \sum_{i=1}^{n-1} |D_i \bar{u} - \bar{w}_i| + \beta \sum_{i=1}^{n-1} |E_i w| - \beta \sum_{i=1}^{n-1} |E_i \bar{w}| \geq 0,
\]
\[
\forall (u, w) \in \mathbb{R}^n \times \mathbb{R}^{n-1}. \quad \Box
\]

4. Regularized data assimilation problem. As a preparatory step for the solution of the data assimilation problem (14) we consider next a properly regularized version, which consists in replacing the non-differentiable part of the objective function, by a differentiable one. Since we are going to use second order optimization methods we choose a \(C^2\) Huber regularization, which, for \(a \in \mathbb{R}\), is given by the following expression:
\[
H_\gamma(a) = \begin{cases}
|a| + C_1 + K & \text{if } a \in \mathcal{A}, \\
\frac{\gamma}{2} a^2 & \text{if } a \in \mathcal{B}, \\
F|a| + \frac{G}{2} |a|^2 + \frac{C}{3} |a|^3 + D & \text{if } a \in \mathcal{I},
\end{cases}
\]
\[
(24)
\]
where
\[
\mathcal{A} := \left\{ a \in \mathbb{R}: \gamma |a| \geq 1 + \frac{1}{2\gamma} \right\},
\]
\[
\mathcal{B} := \left\{ a \in \mathbb{R}: \gamma |a| \leq 1 - \frac{1}{2\gamma} \right\},
\]
\[
\mathcal{I} := \left\{ a \in \mathbb{R}: |\gamma |a| - 1| \leq \frac{1}{2\gamma} \right\}
\]
and
\[
F = 1 - \frac{(2\gamma + 1)^2}{8\gamma}, \quad G = \frac{\gamma}{2}(2\gamma + 1)
\]
\[ C = -\frac{\gamma^3}{2} \quad D = \left(\frac{\gamma}{2} - \frac{G}{2}\right) l_1^2 - F |l_1| - \frac{C}{3} |l_1|^3 \]

\[ C_1 = \frac{\gamma}{2} l_1^2 - l_2 \quad K = F(l_2 - l_1) + \frac{G}{2} (l_2^2 - l_1^2) + \frac{C}{3} (l_2^3 - l_1^3) \]

with constants \( l_1 = \frac{1}{\gamma} \left(1 - \frac{1}{2\gamma}\right) \) and \( l_2 = \frac{1}{\gamma} \left(1 + \frac{1}{2\gamma}\right) \).

Therefore, the regularized version of the \( \ell^1 \)-norm is given by

\[ \sum_{i=1}^{n} |\xi_i| = \sum_{i=1}^{n} H_\gamma(\xi_i) \quad \text{for} \quad \xi \in \mathbb{R}^n. \]

The first derivative of the regularized \( \ell^1 \)-norm is the vector whose components are given by

\[ (h_\gamma(\xi))_i = \begin{cases} \frac{\xi_i}{|\xi_i|} & \text{if } \xi_i \in A, \\ \frac{\xi_i}{|\xi_i|} \left(1 - \frac{\gamma}{2} \left(1 - \gamma|\xi_i| + \frac{1}{2\gamma}\right)^2\right) & \text{if } \xi_i \in I, \end{cases} \]

and the second derivative is a diagonal matrix with elements:

\[ (h_\gamma'(\xi))_i = \begin{cases} \frac{1}{|\xi_i|} - \frac{\xi_i \cdot \xi_i}{|\xi_i|^3} & \text{if } \xi_i \in A, \\ \frac{1}{|\xi_i|} - \frac{\xi_i \cdot \xi_i}{|\xi_i|^3} + \gamma^2 \theta_\gamma(\xi_i) \frac{\xi_i \cdot \xi_i}{|\xi_i|^2} & \text{if } \xi_i \in I, \end{cases} \]

and \( \theta_\gamma(\xi) = 1 - \gamma|\xi_i| + \frac{1}{2\gamma} \).

Therefore, the regularized objective function takes the following form

\[ J_\gamma(y, u, w) = \frac{1}{2} (SHy - z)^T R^{-1} (SHy - z) + \frac{1}{2} (u - u^b)^T B^{-1} (u - u^b) \]

\[ + \frac{\mu}{2} w^T w + \alpha \sum_{i=1}^{n-1} H_\gamma(D_i u - w_i) + \beta \sum_{i=1}^{n-1} H_\gamma(E_i w). \]

Furthermore, the TGV–regularized problem can be written in the following way

\[ \min_{(y, u, w)} J_\gamma(y, u, w) \]

subject to: \( Ey + Z(y)Uy - g(u) = 0. \)

4.1. **Consistency of the regularization.** In this section, we analyze the convergence of the regularized solutions towards the solution of the original data assimilation problem. Let us start by noticing that, for \( a \in \mathbb{R} \), the following holds:

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

In order to prove that the solutions of the regularized problem converge to the original, we need to show the uniform boundedness of the solutions to the regularized data assimilation problems with respect to the Huber regularization parameter \( \gamma \).

The next theorem shows the convergence of the solutions of the regularized problems to the solution of the original one.
Theorem 4.1. Let \((y_\gamma, u_\gamma, w_\gamma)\) be a sequence of global solutions to problem (28), where \(y_\gamma\) satisfies \(e(y_\gamma, u_\gamma) = 0\). Furthermore, let \((y, u, w)\) be a global solution of problem (14). Then, there exists a subsequence \((y_{\gamma_k}, u_{\gamma_k}, w_{\gamma_k})\) such that
\[
(y_{\gamma_k}, u_{\gamma_k}, w_{\gamma_k}) \to (y, u, w), \quad \text{for } k \to \infty.
\]

Proof. We start by recalling that the matrix \(B\) is a covariance symmetric and positive definite matrix. Hence,
\[
(u_\gamma - u_b)^TB^{-1}(u_\gamma - u_b) \geq \lambda_{\min}(B^{-1})\|u_\gamma - u_b\|^2,
\]
where \(\lambda_{\min}(B^{-1})\) is the minimum eigenvalue of the matrix \(B^{-1}\). From the definition of the objective function it follows that
\[
J_\gamma(y_\gamma, u_\gamma, w_\gamma) \geq (u_\gamma - u_b)^TB^{-1}(u_\gamma - u_b) \geq \lambda_{\min}(B^{-1})\|u_\gamma - u_b\|^2.
\]
Using the optimality of \((y_\gamma, u_\gamma, w_\gamma)\) and the fact that \(H_\gamma(a) \leq |a|\), for all \(a \in \mathbb{R}\),
\[
J_\gamma(y_\gamma, u_\gamma, w_\gamma) \leq J_\gamma(y, u, w) \leq J(y, u, w).
\]
Thus, we can conclude that
\[
\|u_\gamma - u_b\|^2 \leq \frac{J(y, u, w)}{\lambda_{\min}(B^{-1})},
\]
and, therefore, the sequence \(\{u_\gamma\}\) is bounded. By the Bolzano–Weierstrass Theorem there exists a convergent subsequence \(\{u_{\gamma_k}\}\) whose limit is denoted by \(\bar{u}\). Using similar arguments we can conclude that
\[
\|w_\gamma\| \leq \sqrt{\frac{2J(y, u, w)}{\mu}},
\]
which is a uniform bound for \(\{w_\gamma\}\). Using again the Bolzano–Weierstrass Theorem, there exists a convergent subsequence denoted by \(\{w_{\gamma_k}\}\) whose limit will be denoted by \(\bar{w}\). Moreover, using the definition of the vector \(y_\gamma\) and by Lemma 3.2, the sequence \(\{y_\gamma\}\) is also bounded and there exists a convergent subsequence \(\{y_{\gamma_k}\}\) with its limit denoted by \(y\). Thus, the triplet \((\bar{y}, \bar{u}, \bar{w})\) is a candidate for solution of problem (14).

Next we will prove the feasibility of \((\bar{y}, \bar{u}, \bar{w})\), that is, \(e(\bar{y}, \bar{u}) = 0\). Recalling the state equation given in (8) we know that
\[
e(\gamma_{\gamma_k}, u_{\gamma_k}) = E_{\gamma_{\gamma_k}} + Z_{\gamma_{\gamma_k}}u_{\gamma_k} - g(u_{\gamma_k}) = 0,
\]
where \(Z_{\gamma_{\gamma_k}} = Z(y_{\gamma_k})\). Taking the limit as \(k \to \infty\) we have
\[
E\bar{y} + \lim_{\gamma \to \infty} (Z_{\gamma_{\gamma_k}}u_{\gamma_k}) - g(\bar{u}) = 0.
\]
In order to guarantee that \(\bar{y}\) satisfies the state equation we have to prove that
\[
\lim_{k \to \infty} Z_{\gamma_{\gamma_k}}u_{\gamma_k} = \bar{Z}U\bar{y},
\]
where \(\bar{Z} = Z(\bar{y})\). This follows from the continuity of \(Z(y)\) with respect to the argument. Consequently,
\[
E\bar{y} + \bar{Z}U\bar{y} - g(\bar{u}) = 0.
\]
Finally, since \(H_\gamma(a) \to |a|\) as \(\gamma \to \infty\), we obtain
\[
J(\bar{y}, \bar{u}, \bar{w}) = \lim_{k \to \infty} J_\gamma(y_{\gamma_k}, u_{\gamma_k}, w_{\gamma_k}) \leq J(y, u, w).
\]
Since \((y, u, w)\) is a global solution of problem (14) we conclude that \((\bar{y}, \bar{u}, \bar{w})\) is also a global solution.

In a similar manner as in the proof of Theorem 3.4, we obtain existence of a Lagrange multiplier \(p_\gamma\) (adjoint state) such that the following optimality system characterizing the optimal solutions of (28) holds:

\begin{align*}
(29a) & \quad E y_\gamma + U Z y_\gamma - g_\gamma = 0, \\
(29b) & \quad E^T p_\gamma + U^T p_\gamma + \text{diag}(U y_\gamma) p_\gamma - H^T S T R^{-1} (H S y_\gamma - z) = 0, \\
(29c) & \quad B^{-1} (u_\gamma - u^b) + \alpha D^T h_\gamma (D u_\gamma - w_\gamma) + \frac{1}{\Delta t} p^1_\gamma = 0, \\
(29d) & \quad \mu w_\gamma - \alpha h_\gamma (D u_\gamma - w_\gamma) + \beta E^T h_\gamma (E w_\gamma) = 0,
\end{align*}

where \(h_\gamma(\cdot)\) is given in (25). Differently from system (19), the regularized optimality condition (29) does not involve variational inequalities. This fact facilitates the solution of the resulting nonlinear system, although enough care has still to be taken in order to get a robust and fast convergent method for its solution.

Furthermore, due to the continuity of the solution mapping \((u, w) \mapsto y(u, w)\) (see Lemma 3.1), we can define the reduced functional as:

\begin{align*}
(30) & \quad f(u, w) = \frac{1}{2} (S H y(u, w) - z)^T R^{-1} (S H y(u, w) - z) + \frac{1}{2} (u - u^b)^T B^{-1} (u - u^b) \\
& \quad + \mu w^T w + \alpha \sum_{i=1}^{n-1} H_\gamma (D_i u - w_i) + \beta \sum_{i=1}^{n-1} H_\gamma (E_i w),
\end{align*}

with its gradient given by

\begin{align*}
(31) & \quad \nabla f(u, w) = \begin{bmatrix}
\frac{1}{\Delta t} p^1_\gamma + B^{-1} (u - u^b) + \alpha D^T h_\gamma (D u - w) \\
\mu w - \alpha h_\gamma (D u - w) + \beta E^T h_\gamma (E w)
\end{bmatrix},
\end{align*}

where \(p_\gamma = [p^1_\gamma, p^2_\gamma, \ldots, p^{N_t}_\gamma]^T\) satisfies equation (29b).

5. Reduced primal–dual Newton type method. The numerical solution of the TGV data assimilation problem can in principle be carried out by solving the optimality system or, alternatively, by using iterative methods for the solution of the optimization problem. In this paper, we focus our attention on the use of an iterative Newton–type method, with the additional use of auxiliary dual multipliers for improving the robustness of the approach. The proposed algorithm has the following ingredients:

1. Use of Newton’s method for the regularized optimization problems;
2. Use of both primal and dual variables for the solution of a saddle point system in order to get search directions;
3. Projection of the dual variables to get modified second–order matrices that guarantee descent directions;
4. Use of polynomial line–search rule for the step length.

We start by recalling Newton’s method for our equality constrained optimization problem, which is given through the following steps:
where \( e'(y,u,w)(v_1, v_2, v_3) = e_y(y,u,w)v_1 + e_u(y,u,w)v_2 + e_w(y,u,w)v_3 \). The Lagrangian is given by \( \mathcal{L}(y,u,w,p) = \frac{1}{2}(SHy - z)^T R^{-1}(SHy - z) + \frac{1}{2}(u - u^b)^T B^{-1}(u - u^b) + \frac{\mu}{2} w^T w \). The second derivative of this extended Lagrangian is then given by \( \mathcal{L}''(y,u,w,q_1,q_2,p) = \begin{bmatrix} 
abla_{y}^2 \mathcal{L} & \nabla_{y} \mathcal{L} & \nabla_{u} \mathcal{L} & \nabla_{w} \mathcal{L} & \nabla_{p} \mathcal{L} \\ \nabla_{y} \mathcal{L}^T & \nabla_{y}^2 \mathcal{L} & \nabla_{u} \mathcal{L} & \nabla_{w} \mathcal{L} & \nabla_{p} \mathcal{L} \\ \nabla_{u} \mathcal{L}^T & \nabla_{u} \mathcal{L} & \nabla_{w}^2 \mathcal{L} & \nabla_{w} \mathcal{L} & \nabla_{p} \mathcal{L} \\ \nabla_{w} \mathcal{L}^T & \nabla_{w} \mathcal{L} & \nabla_{w} \mathcal{L} & \nabla_{w}^2 \mathcal{L} & \nabla_{p} \mathcal{L} \\ \nabla_{p} \mathcal{L}^T & \nabla_{p} \mathcal{L} & \nabla_{p} \mathcal{L} & \nabla_{p} \mathcal{L} & \nabla_{p}^2 \mathcal{L} \\ \end{bmatrix} \).
TGV-regularization for data assimilation

\[ h'(\gamma(y)) \] is the second derivative of the Huber regularization given in (26) and (33)

\[ \Psi = H^T S^T R^{-1} HS - \mathcal{K}, \]

with

\[ \mathcal{K} = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 \\
0 & \text{diag}(U^T p^2) + \text{diag}(U p^2) & 0 & 0 & 0 \\
0 & 0 & \text{diag}(U^T p^2) + \text{diag}(U p^2) & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \text{diag}(U^T p^N) + \text{diag}(Up^N) 
\end{bmatrix}. \]

By analyzing the elements of the diagonal of the second derivative given by (26), it can be verified that there are a lot of zero entries, which causes ill-conditioning of the matrix, with the corresponding algorithmic robustness issues. Moreover, the positive definiteness of the matrix \( \mathcal{L}'' \) does not necessarily hold. For these reasons, we modify the matrix in such a way that better conditioning properties and positive definiteness are obtained.

To do so, we consider the primal dual formulation of the problem and perform a projection of the dual variables (see [12]). Specifically, this process consists in replacing the term

\[ (Du - w) \odot (Du - w) \]

which appears in the second derivative of the Huber regularization, by the following term

\[ \frac{q_1}{\max\{1, |q_1|\}} \odot \frac{Du - w}{|Du - w|^2}, \]

where \((u \odot w)\) is the component-wise product of the vectors \(u\) and \(w\). Thus, \(h'(Du - w)\) is replaced by the diagonal matrix \(Q_1\) given by

\[ \text{diag}(Q_1) = \begin{cases}
\frac{1}{|\zeta|} - \frac{q_1}{\max\{1, |q_1|\}} \odot \frac{\zeta}{|\zeta|^2} & \text{if } \zeta_i \in A, \\
\gamma & \text{if } \zeta_i \in B, \\
\left(1 - \frac{\gamma}{2} \vartheta_i^2\right) \left[\frac{1}{|\vartheta|} - \frac{q_2}{\max\{1, |q_2|\}} \odot \frac{\vartheta}{|\vartheta|^2}\right] + \gamma^2 \theta_i \frac{\zeta}{|\zeta|} \odot \frac{\zeta}{|\zeta|} & \text{if } \zeta_i \in \mathcal{I},
\end{cases} \]

where \(\zeta_i = D_i u - w_i\). Moreover, \(h'(E w)\) is replaced by the diagonal matrix \(Q_2\) given by

\[ \text{diag}(Q_2) = \begin{cases}
\frac{1}{|\vartheta|} - \frac{q_2}{\max\{1, |q_2|\}} \odot \frac{\vartheta}{|\vartheta|^2} & \text{if } \theta_i \in A, \\
\gamma & \text{if } \theta_i \in B, \\
\left(1 - \frac{\gamma}{2} \vartheta_i^2\right) \left[\frac{1}{|\vartheta|} - \frac{q_2}{\max\{1, |q_2|\}} \odot \frac{\vartheta}{|\vartheta|^2}\right] + \gamma^2 \theta_i \frac{\vartheta_i}{|\vartheta_i|} \odot \frac{\vartheta_i}{|\vartheta_i|} & \text{if } \theta_i \in \mathcal{I},
\end{cases} \]

where \(\vartheta_i = E_i w\) and the division, max and \(|\cdot|\) are component-wise operations. It is worth to remark that in the last term of the third case of the \(Q_1\) and \(Q_2\) matrices we do not perform the projection.

In this way, the modified second derivative of the Lagrangian used in the method is

\[ \mathcal{L}''(y, u, w, q_1, q_2) = \begin{bmatrix}
\Psi & 0 & 0 & 0 & 0 \\
0 & B^{-1} & 0 & \alpha D^T & 0 \\
0 & 0 & \mu \mathbb{I} & -\alpha \mathbb{I} & \beta E^T \\
0 & -Q_1 D & Q_1 & \mathbb{I} & 0 \\
0 & 0 & -Q_2 E & 0 & \mathbb{I}
\end{bmatrix}. \]
On the other hand, the derivatives of the state equation are given by
\[
e_y(y, u) = E + ZU + \text{diag}(UY) =: \Xi
\]
and
\[
e_u(y, u) = -(1/\Delta t) \begin{bmatrix}
0 & 0 & \cdots & 0 \\
1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
0 & 0 & \cdots & 1
\end{bmatrix} =: \Upsilon
\]
Furthermore, since the state equation does not depend on \( p, q_1 \) nor \( q_2 \), its derivatives with respect to these variables are equal to zero.

Now, the system that defines the method is given by:
\[
\begin{bmatrix}
\Psi & 0 & 0 & 0 & 0 & \Xi^T \\
0 & B^{-1} & 0 & \alpha D^T & 0 & \Upsilon^T \\
0 & 0 & \mu I & -\alpha I & \beta E^T & 0 \\
0 & -Q_1 D & Q_1 & \mathbb{I} & 0 & 0 \\
\Xi & \Upsilon & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
\delta_y \\
\delta u \\
\delta w \\
\delta q_1 \\
\delta q_2 \\
\delta \pi
\end{bmatrix}
= \begin{bmatrix}
0 \\
-p^1/\Delta t - B^{-1}(u - u^b) - \alpha D^T q_1 \\
-\mu w + \alpha q_1 - \beta E^T q_2 \\
-q_1 + h_\gamma(Du - w) \\
-q_2 + h_\gamma(Ew) \\
0
\end{bmatrix}
\]

5.1. Reduced method and matrix properties. Before presenting the convergence result of the proposed algorithm, we start by analyzing the reduced version of it.

From Lemma 1 we know that the matrix \( \Xi \) in (37) is invertible and, therefore, the following equalities hold:
\[
\begin{align*}
\delta q_1 &= Q_1 D \delta u - Q_1 \delta w - q_1 + h_\gamma(Du - w), \\
\delta q_2 &= Q_2 E \delta w - q_2 + h_\gamma(Du - w), \\
\delta y &= -(\Xi)^{-1} \Psi \delta u, \\
\delta \pi &= (\Xi)^{-T} \Psi (\Xi)^{-1} \Psi \delta u.
\end{align*}
\]
System (39) can then be rewritten as follows:
\[
(B^{-1} + \alpha D^T Q_1 D + \Upsilon^T (\Xi)^{-T} (\Xi)^{-1} \Psi) \delta u - \alpha D^T Q_1 \delta w
= -1/\Delta t p^1 - B^{-1}(u - u^b) - \alpha D^T h_\gamma(Du - w),
\]
\[
-\alpha Q_1 D \delta u + (\mu I + \alpha Q_1 + \beta E^T Q_2 E) \delta w = -\mu w + \alpha h_\gamma(Du - w) - \beta E^T h_\gamma(Ew),
\]
or, equivalently,
\[
\Pi \begin{bmatrix}
\delta u \\
\delta w
\end{bmatrix}
= -\nabla f(u, w),
\]
where \( \nabla f(u, w) \) was defined in (31) and
\[
\Pi :=
\begin{bmatrix}
B^{-1} + \alpha D^T Q_1 D + \Upsilon^T (\Xi)^{-T} (\Xi)^{-1} \Psi
& -\alpha D^T Q_1 \\
-\alpha Q_1 D & \mu I + \alpha Q_1 + \beta E^T Q_2 E
\end{bmatrix}.
\]
Once we have defined the reduced matrices we present the reduced method given in Algorithm 2.

**Algorithm 2 Reduced Newton–type method**

1: Initialize $k = 0, u_0$

2: repeat
   3: Compute $y^k$ the solution of the state equation given in (8).
   4: Compute $p^k$ the solution of the adjoint equation given in (15).
   5: Compute the descent direction by solving system (40)
   6: Compute $s^k$ using Algorithm 3 below.
   7: Set $u^{k+1} = u^k + s^k \delta u$, $w^{k+1} = u^k + s^k \delta w$ and $x^k = (u^k, w^k)$.
   8: $k \leftarrow k + 1$.
   9: until Stopping criteria

We focus next on the main features of the matrix $\Pi$ that can guarantee that $d^k = [\delta u, \delta w]^T$ is indeed a descent direction. We start by proving its symmetry.

**Proposition 2.** The matrix $\Pi$ given in (41) is symmetric.

*Proof.* Since $\Pi$ is a block–matrix we know that

$$
\Pi^T = \begin{bmatrix}
\Pi_{11}^T & \Pi_{12}^T \\
\Pi_{12} & \Pi_{22}
\end{bmatrix}.
$$

We analyze each of the components starting with $\Pi_{11}$, which is given by

$$
\Pi_{11}^T = (B^{-1} + \beta D^T Q_1 D + \Upsilon \Xi^{-1} \Upsilon^{-T})^T
= (B^T)^{-1} + \beta D^T Q_1^T D + \Upsilon \Xi^{-1} \Upsilon^T \Xi^{-1} \Upsilon.
$$

Since $B$ is a covariance matrix, it is symmetric and its inverse as well. It remains to prove that the matrices $Q_1$ and $\Psi$ are also symmetric. For the matrix $Q_1$, we know by definition that it is a diagonal matrix. Finally, we are going to analyze the matrix $\Psi$ given by

$$
\Psi^T = (H^T S^T R^{-1} S H - \Sigma)^T = H^T S^T R^{-T} - \Sigma^T.
$$

We notice that $R$ is also a covariance matrix and, therefore, symmetric. Since $\Sigma$ is a diagonal matrix (see eq. (35)), $\Psi$ is also symmetric and therefore $\Pi_{11}$ is symmetric.

We now focus on

$$
\Pi_{22}^T = \mu I + \alpha Q_1^T + \beta E^T Q_2^T E.
$$

Since $Q_1$ and $Q_2$ are diagonal matrices, it immediately follows that $\Pi_{22}$ is symmetric.

By definition we have

$$
\Pi^T = \begin{bmatrix}
\Pi_{11}^T & (-\alpha Q_1 D)^T \\
(-\alpha D^T Q_1) & \Pi_{22}^T
\end{bmatrix} = \begin{bmatrix}
\Pi_{11} & (-\alpha D^T Q_1) \\
(-\alpha Q_1 D) & \Pi_{22}
\end{bmatrix} = \Pi
$$

and the proof is complete. \qed

Now we are going to prove that $\Pi$ is positive definite and, therefore, that the directions given in (40) are indeed descent directions for the data assimilation problem. Before stating the main result, we prove some required lemmata.

**Lemma 5.1.** If the term $\|H^T S^T R^{-1}(S H y - z)\|$ is small enough, then there exists a constant $\kappa > 0$ such that for all $\xi \in \mathbb{R}^m \setminus \{0\}$ we have

$$
\xi^T \mathcal{L}^n_{(y, x)}(y, u, w, p) \xi \geq \kappa \|\xi\|^2.
$$
Proof. The first and second partial derivatives of the Lagrangian (32), with respect to \( y \), are given by
\[
\mathcal{L}'_y(y, u, w, p) = H^T S^T R^{-1} (SHy - z) - E^T p - Z U^T p - \text{diag}(Uy)p.
\]
and
\[
\mathcal{L}''_y(y, u, w, p) = H^T S^T R^{-1} SH - K,
\]
respectively, where \( K \) is given in (34). Using the definition of the matrix \( K \) and multiplying both sides by \( \xi \in \mathbb{R}^m \setminus \{0\} \) we get
\[
\xi^T \mathcal{L}''_y = \xi^T H^T S^T R^{-1} SH \xi - \xi^T K \xi \geq c\|\xi\|^2 - 2 \sum_{i=1}^{N} \xi_i^T (U_i p^i) \xi_i
\]
\[
= c\|\xi\|^2 - 2 \sum_{i=1}^{N} \sum_{j=1}^{n} (U_{ij})^2 p_i^j \xi_i^2 \geq c\|\xi\|^2 - 2\|Up\| \sum_{i=1}^{N} \sum_{j=1}^{n} (\xi_i^j)^2
\]
\[
\geq c\|\xi\|^2 - 2\|U\|\|\xi\|\|\xi\|^2
\]
\[
\geq (c - 2\rho\|U\|\|H^T S^T R^{-1} (SHy - z)\|) \|\xi\|^2,
\]
where \( c \) is the smallest eigenvalue of the matrix \( H^T S^T R^{-1} SH \). The last inequality was obtained by applying Lemma 3.3. Defining
\[
\kappa := c - 2\rho\|U\|\|H^T S^T R^{-1} (SHy - z)\|.
\]
and setting
\[
\|H^T S^T R^{-1} (SHy - z)\| \leq \frac{c}{2\rho},
\]
we guarantee that \( \kappa > 0 \) and thus we get the desired result. \( \square \)

Proposition 3. The matrix \( \Pi \) defined in (41) is positive definite.

Proof. Multiplying on both sides of \( \Pi \) by a block vector \((x, y)\) we obtain that
\[
(x, y) \Pi \begin{pmatrix} x \\ y \end{pmatrix} = (x, y) \begin{pmatrix} B^{-1} + \alpha D^T Q_1 D + \Upsilon^T (\Xi)^{-1} \Psi (\Xi)^{-1} \Upsilon & -\alpha D^T Q_1 \\ -\alpha Q_1 D & \mu I + \alpha Q_1 + \beta E T Q_2 E \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}
\]
\[
\geq x^T B^{-1} x + \alpha x^T D^T Q_1 D x - 2\alpha x^T D^T Q_1 y + \mu y^T y + \alpha y^T Q_1 y + \beta y^T E^T Q_2 E y,
\]
since the matrix \( \Upsilon^T (\Xi)^{-1} \Psi (\Xi)^{-1} \Upsilon \) is positive semi-definite thanks to Lemma 5.1. To proof the statement we have to verify that both \( Q_1 \) and \( Q_2 \) are positive semi-definite.

The matrix \( Q_1 \) is diagonal, thus we just need to prove that each element of its diagonal is positive. We start noticing that \( \frac{\|(q_1)_i\|}{\max \{\|(q_1)_i\|, 1\}} \leq 1 \) and, therefore,
\[
\frac{(q_1)_i}{\max \{\|(q_1)_i\|, 1\}} \circ \frac{D_i u - w_i}{|D_i u - w_i|^2} \leq \frac{D_i u - w_i}{|D_i u - w_i|^2}.
\]
Now using definition (35) we analyze the following three cases.
\[
D_i u - w_i \in \mathcal{A}:
\]
\[
(Q_1)_{ii} = \frac{1}{|D_i u - w_i|} - \frac{(q_1)_i}{\max \{\|(q_1)_i\|, 1\}} \circ \frac{D_i u - w_i}{|D_i u - w_i|^2}.
\]
Using the bound given in (42) we have
\[(Q_1)_{ii} = \frac{1}{|D_i u - w_i|} - \frac{(q_1)_i}{\max\{|(q_1)_i|, 1\}} \odot \frac{D_i u - w_i}{|D_i u - w_i|^2},\]
\[\geq \frac{1}{|D_i u - w_i|} \left(1 - \frac{D_i u - w_i}{|D_i u - w_i|}\right) \geq 0.\]

**D_i u - w_i ∈ B:** From the definition we have
\[(Q_1)_{ii} = \gamma > 0.\]

**D_i u - w_i ∈ I:** We start by recalling the structure of the set
\[I := \{i : |\gamma|D_i u - w_i| - 1| \leq 1/2\}.\]

The elements of the diagonal of Q1 are given by
\[(Q_1)_{ii} = \left(1 - \frac{\gamma}{2} \theta_i^2\right) \left[\frac{1}{|D_i u - w_i|} - \frac{(q_1)_i}{\max\{|(q_1)_i|, 1\}} \odot \frac{D_i u - w_i}{|D_i u - w_i|^2}\right]
+ \gamma^2 \theta_i \frac{D_i u - w_i}{|D_i u - w_i|} \odot \frac{D_i u - w_i}{|D_i u - w_i|}.
\]
with \(\theta_i = (1 - \gamma|D_i u - w_i| + 1/2\gamma).\) Using the same argument as in the first case we have
\[\left[\frac{1}{|D_i u - w_i|} - \frac{(q_1)_i}{\max\{|(q_1)_i|, 1\}} \odot \frac{D_i u - w_i}{|D_i u - w_i|^2}\right] \geq 0.\]

By the definition of the set \(I\) it follows that \(\theta_i \geq 0\) and \(1 - \frac{\gamma}{2} \theta_i^2 \geq 1 - \frac{1}{2\gamma} \geq 0.\) Therefore,
\[(Q_1)_{ii} \geq \gamma^2 \theta_i \left(\frac{D_i u - w_i}{|D_i u - w_i|} \odot \frac{D_i u - w_i}{|D_i u - w_i|}\right) = \gamma^2 \theta_i \left(\frac{D_i u - w_i}{|D_i u - w_i|}\right)^2 \geq 0.\]

Since each element of the diagonal of Q1 is non-negative we can conclude that the matrix is in fact a positive semi definite matrix. Furthermore, we justify why the last term on this expression was not projected, because otherwise we can not guarantee that this term is positive.

In a similar manner it can be proved that Q2 is also positive semi-definite. Altogether we obtain that
\[(x, y) \Pi \left[\begin{array}{c} x \\ y \end{array}\right] \geq \lambda_{\min}(B^{-1})\|x\|^2 + \mu\|y\|^2 + \alpha\|Q_1^{1/2} D x - Q_1^{1/2} y\|^2.\]

6. **Convergence analysis.** This section is devoted to the study of the convergence properties of the proposed algorithm. Our aim is to use the general results developed for the convergence analysis of descent methods (see [8, Theorem 4.2]).

In order to apply these results, we need to verify that the angle condition and the efficiency of the line-search, i.e.,
\[\eta \|\nabla f(x^k)\|\|d^k\|\]
\[f(x^k + s_k d^k) < f(x^k)\] (descent)
where \( x = (u, w) \) and \( f \) stands for the reduced cost function, i.e., \( f(x) = J_\gamma(y(u), u) \).

These conditions are required to hold on the level set, which is defined next.

**Definition 6.1.** Let \( \rho > 0 \) be a constant, the level set of the problem is given by 
\[
N_0^\rho := \{(u, w) + (d_u, d_w) : f(u, w) \leq f(u_0, w_0), \|(d_u, d_w)\| \leq \rho \},
\]
for a given \( u_0, w_0 \).

Due to the continuous differentiability of \( y(u, w) \) (see Proposition 3.1) and the structure of the objective function it is easy to prove that \( f(x) \) is radially unbounded and continuous, therefore, the level set \( N_0^\rho \) is compact.

### 6.1. Angle condition

We start by proving that the angle condition (43) holds.

**Proposition 4.** Let \( \Pi_k = \Pi(x_k) \) be defined by (41), then there exist constants \( 0 < \zeta < \bar{c} \) independent of \( k \) such that
\[
\zeta \|q\|^2 \leq q^T \Pi_k q \leq \bar{c} \|q\|^2, \quad \text{for all } q \in \mathbb{R}^n, \quad \text{for all } x_k \in N_0^\rho.
\]

**Proof.** We start recalling the structure of the matrix \( \Pi \).

\[
\Pi = \begin{bmatrix}
\Pi_{11} & \Pi_{12} \\
\Pi_{21} & \Pi_{22}
\end{bmatrix} = \begin{bmatrix}
B^{-1} + \alpha D Q_1 D + \gamma T \Xi^{-T} \Psi \Xi^{-1} \gamma \\
-\alpha Q_1 D & \mu + \alpha Q_1 + \beta E^T Q_2 E
\end{bmatrix}
\]

From Proposition 3 it immediately follows that
\[
q^T \Pi_k q \geq \zeta \|q\|^2, \quad \forall q \in \mathbb{R}^n, \quad \text{with } \zeta = \min \{\lambda_{\min}(B^{-1}), \mu\}.
\]

Concerning the upper bound, we get that
\[
(q_1, q_2) \Pi \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = q_1^T B^{-1} q_1 + \alpha q_1^T D^T Q_1 D q_1 - 2 \alpha q_1^T D^T Q_1 q_2 + \alpha q_2^T Q_1 q_2 + \mu q_2^T q_2 \\
+ \beta q_2^T E^T Q_2 E q_2 + q_1^T \gamma T \Xi^{-T} \Psi \Xi^{-1} \gamma q_1
\]
\[
\leq C_B \|q_1\|^2 + \alpha \|Q_1^{1/2} D q_1 - Q_1^{1/2} q_2\|^2 + \mu \|q_2\|^2 \\
+ \beta \gamma \|E\|^2 \|q_2\|^2 + \frac{1}{(\Delta t)^2} \|\Psi\| \|\Xi^{-1}\|^2 \|q_1\|^2.
\]

For the last term in the previous inequality we first get a bound on \( \Psi_k \), which is given by
\[
\|\Psi_k\| \leq \|H^T S^T R^{-1} S H\| + \|K_k\|.
\]

Using the definition of \( K_k \) given in (34) and Lemma 3.3,
\[
\|K_k\| = 2 \|U\| \|p_k\| \leq 2 \rho \|U\| \|H^T S^T R^{-1}\| \|S H y_k - z\| \\
\leq 2 \rho C_0 \|U\| \|H^T S^T R^{-1}\|,
\]

where the last inequality was obtained using Lemma 3.2 and the fact that \( u_k \in N_0^\rho \).

**Defining**
\[
\xi := 2 \rho C_0 \|U\| \|H^T S^T R^{-1}\|
\]

we get that
\[
\|\Psi_k\| \leq \|H^T S^T R^{-1} S H\| + \xi =: C_\Psi.
\]

For the norm of \( \Xi^{-1} \) we consider an a-priori estimate of the adjoint equation. Specifically, let \( v \) be solution of
\[
\Xi v = \Xi v + Z_k \Xi v + \text{diag}(\Xi y_k)v = g.
\]
Proceeding as in the proof of Lemma 3.3 we get that
\[ \|v\| \leq \rho \|g\|, \]
with \( \rho \) independent of \( y \) and \( u \), and therefore \( \|\Xi^{-1}\| \leq \rho \).

Inserting the last results in (46) and taking into account that \( \|Q_i\| \leq \gamma \), for \( i = 1, 2 \), we then obtain
\[ q^T \Pi Q \leq \left( C_B + \gamma C_D + \tilde{C} \right) \|q_1\|^2 + \| \mu + \gamma (\alpha + \beta C_E) \|q_2\|^2 \]
\[ \leq \max \left\{ \left( C_B + \gamma C_D + \tilde{C} \right), \mu + \gamma (\alpha + \beta C_E) \right\} \|q\|^2, \]
where \( \tilde{C} := \frac{\rho^2 C_1}{(\Delta t)^2} \). Therefore, the constant that we were looking for is
\[ \bar{c} = \max \left\{ C_B + \gamma C_D + \tilde{C}, \mu + \gamma (\alpha + \beta C_E) \right\} \]
\[ \Box \]

It is worth to remark that, from Proposition 4 it follows immediately that the descent directions \( \{d_k\} \) generated by Algorithm 2 satisfy the angle condition (see, e.g., [8]).

6.2. Uniform continuity of the gradient. In order to present the result that guarantees the uniform continuity of the reduced gradient, we start by proving the following lemmata that allow us to write, in a shorter way, the main theorem of this subsection.

**Lemma 6.2.** Let \( u_1, u_2 \in N_0^\rho \). Furthermore, let \( y(u_1) \) and \( y(u_2) \) be solutions of the state equations \( e(y(u_1), u_1) = 0 \) and \( e(y(u_2), u_2) = 0 \), respectively. Then the following bound holds
\[ \|y^j(u_1) - y^j(u_2)\| \leq C\|u_1 - u_2\|, \quad j = 1, \ldots, N_t, \]
where \( C > 0 \) is a constant independent of \( u_1, u_2 \) and \( y \).

**Proof.** From Proposition 3.1 we know that \( y(u) = [y^1(u), y^2(u), \ldots, y^{N_t}(u)] \) is a continuously differentiable function on \( \mathbb{R}^m \), which means that \( y^j(u) \) is continuously differentiable for all \( j = 1, \ldots, N_t \). Using the mean value theorem we can conclude that
\[ \|y^j(u_1) - y^j(u_2)\| \leq \|(y^j)'(\xi)\|\|u_1 - u_2\| \]
for some \( \xi \) that belongs to a neighborhood which contains \( u_1, u_2 \). Since \( (y^j)' \) is continuous and \( u_1, u_2 \) belong to the level set \( N_0^\rho \) and this is compact, there exists \( C > 0 \) that satisfies the required bound. \( \Box \)

**Lemma 6.3.** Let \( u_1, u_2 \in N_0^\rho \). Furthermore, let \( y(u_1) \) and \( y(u_2) \) be solutions of the state equations \( e(y(u_1), u_1) = 0 \) and \( e(y(u_2), u_2) = 0 \), respectively. In addition let \( p(u_1), p(u_2) \) be the associated adjoint states. Then the following bound holds
\[ \|p^j(u_1) - p^j(u_2)\| \leq \tilde{C}\|u_1 - u_2\|, \quad j = 1, \ldots, N_t, \]
where \( \tilde{C} > 0 \) is a constant independent of \( u_1, u_2 \) and \( y \) or \( p \).

**Proof.** From the analysis component–wise in time of two adjoint equations (15) in the variables \( u_1, u_2 \in N_0^\rho \) and denoting
\[ A := \left( \frac{1}{\Delta t} I + \text{diag}(y^{-1}(u_1))U^T + \text{diag}(Uy^j(u_1)) \right), \]
Theorem 6.4. The gradient of the reduced cost functional, given by
\[
\nabla f(u, w) = \left[ \frac{1}{\Delta t} p^1 + B^{-1}(u - u^h) + \alpha D^T h_\gamma(Du - w), \right. \\
\left. \mu w - \alpha h_\gamma(Du - w) + \beta E^T h_\gamma(Ew) \right]
\]
is uniformly continuous on the level set \( N_0^p \).

Proof. The proof follows directly from the Heine Theorem since, thanks to Lemmas 6.2 and 6.3, \( \nabla f(u, w) \) is continuous over the level set \( N_0^p \), which is compact. \( \square \)

6.3. Polynomial line–search. Concerning the line–search procedure, we consider the polynomial backtracking algorithm proposed in [10]. In the following lines we provide a short presentation of the algorithm.

It is known that the usual backtracking is defined by a parameter \( \lambda \leq 1 \) such that in each iteration the new step length is computed as \( s_k = \lambda^k \). Nevertheless, there are some other methods that instead of taking a constant value of \( \lambda \), they make it variable throughout the iterations, and this is the case of the polynomial line–search rule.
The main feature of the polynomial backtracking is to choose the step length \( s_k \) by using a quadratic or cubic approximation of the objective function. Indeed, if the Armijo condition is not satisfied in the first iteration, i.e.,

\[
\tilde{f}(1) > \tilde{f}(0) + c_1 \tilde{f}'(0),
\]

where \( \tilde{f}(s) = f(x_s) \), \( x_s = x + sd \) and \( 0 < c_1 \ll 1 \) is a given constant, the algorithm uses the known function information and the quadratic model

\[
\tilde{m}_q(s) = [\tilde{f}(1) - \tilde{f}(0) - c_1 \tilde{f}'(0)]s^2 + \tilde{f}'(0)s + \tilde{f}(0),
\]

whose exact minimizer is given by

\[
(48) \quad \tilde{s} = \frac{-\tilde{f}'(0)}{2[\tilde{f}(1) - \tilde{f}(0) - c_1 \tilde{f}'(0)]}.
\]

The algorithm checks if the Armijo condition is fulfilled or not. If the answer is yes, the algorithm stops with \( s_k = \tilde{s} \). Otherwise the new information is used in order to build a cubic model of the function given by the expression

\[
\tilde{m}_c(s) = as^3 + bs^2 + \tilde{f}'(0)s + \tilde{f}(0),
\]

with

\[
a = \frac{1}{s_{k-2} - s_{k-1}} \left[ \frac{1}{s_{k-2}^2} (\tilde{f}(s_{k-2}) - \tilde{f}(0) - \tilde{f}'(0)s_{k-2}) \right] + \frac{1}{s_{k-2} - s_{k-1}} \left[ \frac{1}{s_{k-1}^2} (\tilde{f}(s_{k-1}) - \tilde{f}(0) - \tilde{f}'(0)s_{k-1}) \right],
\]

\[
b = \frac{1}{s_{k-2} - s_{k-1}} \left[ -\frac{s_{k-1}}{s_{k-2}} (\tilde{f}(s_{k-2}) - \tilde{f}(0) - \tilde{f}'(0)s_{k-2}) \right] + \frac{1}{s_{k-2} - s_{k-1}} \left[ \frac{s_{k-2}}{s_{k-1}^2} (\tilde{f}(s_{k-1}) - \tilde{f}(0) - \tilde{f}'(0)s_{k-1}) \right],
\]

whose minimizer is given by

\[
(49) \quad \tilde{s} = \frac{-b + \sqrt{b^2 - 3a \tilde{f}'(0)}}{3a}.
\]

This method can be summarized through the following steps:

It is worth to remark that despite the different ways of computing the value of the step–length, the Armijo condition still has to hold, that is

\[
(50) \quad f(x_s) \leq f(x) - c_1 s \nabla f(x)^T d.
\]

Using [13, Lemma 2.2] we can guarantee that there exist steps \( s_k \) that satisfy the Armijo condition, no matter which backtracking rule we are using.

For the polynomial backtracking, the following bounds for the step lengths hold. If in the first iteration the Armijo condition is not fulfilled, the previous step–length is bounded by

\[
(51) \quad \tilde{s} < \frac{1}{2(1 - c_1)}.
\]

If the Armijo condition is not fulfilled in the next iterations, the following bound holds

\[
(52) \quad \tilde{s} < \frac{2}{3} s_{\text{prev}},
\]

where \( s_{\text{prev}} \) is the previous step–length.
Algorithm 3 Polynomial Backtracking

1: Initialize $s = 1, c_1 = 1e - 4, k = 1$
2: While $f(s_k) > f(0) + c_1 f'(0)$
3: If $s_k > 0.1$
4: If $k = 2$
5: Compute $\tilde{s}$ given by (48)
6: Set $s_k = \tilde{s}$
7: Else
8: Compute $\tilde{s}$ given by (49)
9: If $\tilde{s} < \frac{1}{2}s_{k-1}$
10: set $s_k = \tilde{s}$
11: Else
12: Set $s_k = \frac{1}{2}s_{k-1}$
13: End If
14: End If
15: Else $s_k = 0.1$
16: End If
17: $k \leftarrow k + 1$
18: End While

Proposition 5. Let $\nabla f$ be the gradient of the reduced functional given in (31). If the iterations generated by Algorithm 2, with $\{s_k\}$ given by Algorithm 3, satisfy

\begin{equation}
\|d^k\| \geq \frac{-\nabla f(x^k)^T d^k}{M\|d^k\|},
\end{equation}

with $M > 0$, then $\{s_k\}$ satisfies the efficiency of the line search.

Proof. Arguing by contradiction, we assume that there is an infinite set $K$ and $\varepsilon > 0$ such that

\begin{equation}
f(x^k + s_k d^k) - f(x^k) \xrightarrow{k \to \infty} 0 \quad \text{and} \quad \frac{\nabla f(x^k)^T d^k}{\|d^k\|} \leq -\varepsilon
\end{equation}

for all $k \in K$.

Now, since (53) holds, it follows that, for all $k \in K$,

\begin{equation}
\|d^k\| \geq \frac{-\nabla f(x^k)^T d^k}{M\|d^k\|} \geq \frac{\varepsilon}{M} =: \delta > 0
\end{equation}

On the other hand, using the mean value Theorem, we know that for all $k \in K$, there exists $\tau_k \in [0, s_k]$ such that

\begin{equation}
\frac{f(x^k + s_k d^k) - f(x^k)}{\|s_k d^k\|} = \frac{\nabla f(x^k + \tau_k d^k)^T d^k}{\|d^k\|} - \frac{c_1 \nabla f(x^k)^T d^k}{\|d^k\|} \leq \frac{\|\nabla f(x^k + \tau_k d^k) - \nabla f(x^k)\|}{\|d^k\|} + (1 - c_1) \frac{\nabla f(x^k)^T d^k}{\|d^k\|}
\end{equation}

\begin{equation}
\leq \|\nabla f(x^k + \tau_k d^k) - \nabla f(x^k)\| - (1 - c_1) \frac{\varepsilon}{M}.
\end{equation}
Since $\nabla f$ is uniformly continuous on the compact level set $N_\rho^0$ (see Theorem 6.4), there exists $\eta$ independent of $x^k$ such that, if $\|d\| \leq \eta$, then

$$\|\nabla f(x^k + d) - \nabla f(x^k)\| < (1 - c_1) \frac{\varepsilon}{M},$$

for any $k \in K$.

Jointly with (55) and the result above it is clear that the Armijo condition holds whenever $s_k \|d^k\| \leq \eta$. Using the bounds obtained for the polynomial backtracking (51) and (52) we know that $s_k = 1$ and the Armijo condition is fulfilled in the first iteration or we have two possible scenarios:

- Armijo condition is fulfilled in the second iteration:

  $$s_k \leq \frac{1}{2(1 - c_1)} =: \sigma < 1 \text{ and } s_k \|d^k\| > \eta$$

- Armijo condition is fulfilled in a later iteration:

  $$s_k \leq \frac{2}{3} s_{\text{prev}} < \left(\frac{2\sigma}{3}\right)^n \text{ and } s_k \|d^k\| > \eta$$

Therefore, for all $k \in K$ it follows that

$$s_k \|d^k\| \geq \min \left\{ \eta, \frac{2\sigma \eta}{3}, \delta \right\} =: \theta > 0.$$

Using again the Armijo condition we have

$$f(x^k) - f(x^k + s_k d^k) \geq -s_k c_1 \nabla f(x^k)^T d^k$$

$$= c_1 \left(\frac{-\nabla f(x^k)^T d^k}{\|d^k\|}\right) (s_k \|d^k\|) \geq c_1 M \delta \theta > 0.$$

Consequently, $f(x^k) - f(x^k + s_k d^k) \rightarrow 0$ and we get the result.

Condition (53) is automatically fulfilled for our algorithm since the matrices $\{\Pi_k\}$ satisfy the condition given in Proposition 4 and $\nabla f$ is uniformly continuous thanks to Lemma 6.4.

6.4. Convergence result. Summarizing, thanks to the satisfaction of the angle condition, the uniform continuity of the gradient and the efficiency of the line search, we arrive at the following global convergence result for the proposed algorithm.

**Theorem 6.5.** Let $f$ be defined as in (30) and let $\{x^k\}$, $\{s_k\}$ and $\{d^k\}$ be sequences generated by Algorithm 2. Then,

$$\lim_{k \rightarrow +\infty} \nabla f(x^k) = 0$$

and every accumulation point of $\{x^k\}$ is a stationary point of $f$.

Although we do not provide any proof of superlinear convergence, in the tests below this convergence rate can be experimentally verified.

7. Numerical experiments. In this section, we report on numerical experiments whose primary goal is to show the behavior of TGV compared to other variational regularizations, for the solution of the data assimilation problem of the inviscid Burgers’ equation. Moreover, the performance of the proposed solution algorithm for the TGV data assimilation problem is experimentally verified.

Three experiments have been designed for these purposes using the spatial domain $\Omega = (0, 10)$ and the time interval $[0, T] = [0, 1]$. The first one aims to verify the well-known staircase effect of total variation (TV) and also to show how the
total generalized variation regularization (TGV) behaves in such cases. The second experiment has been constructed with the purpose of illustrating the performance of the proposed algorithm regarding its convergence and its behavior with respect to the regularization parameters. The last experiment aims to study the influence of the number of observations and the background covariance matrix in the performance of the algorithm.

We perform the following steps for the construction of the dataset:

1. Set an exact solution (initial condition);
2. Solve the state equation with the exact solution and extract perfect observations. The observations will be extracted every $M$ instants in time and $N$ nodes in the space. That means, we will take the instants of time $1, M+1, 2M+1, \ldots$ and the nodes $1, N+1, 2N+1, \ldots$;
3. Add a Gaussian noise with zero mean and a covariance matrix $B = \sigma I$ to the exact solution for the background information;
4. Fix the parameters $\alpha, \beta, \gamma$ and $\mu$.

The parameters for the TGV–regularization were chosen following the heuristic described in [9], where $\frac{\beta}{\alpha} \in \left(0.75, 1.5\right)$. For the background matrix we set $\sigma = 0.1$, and assume perfect observations, i.e., $R = I$. Finally, as stopping criteria we use $\|u^k - u^{k-1}\| < 1e^{-3}$.

7.1. **Comparison between TV and TGV regularization.** This experiment is designed with the purpose of showing the staircase effect of the TV regularization on the data assimilation problem and the way how the TGV regularization eliminates it. We solve problem (29) with 50 spatial discretization points and 150 points for the time discretization. Furthermore, we take 25 observations from 7500 possible choices, using $M = 25, N = 10$.

The exact solution for this experiment is given by the expression in Figure 1.

![Figure 1. Exact solution](image-url)

As quality measure we use the SSIM (structural similarity index) [22], typically considered in image and signal processing. This index is given by the following expression

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)},$$

where $\mu_x, \mu_y$ are the mean and $\sigma_x^2, \sigma_y^2$ are the variances of the vectors $x$ and $y$, respectively. Furthermore, $\sigma_{xy}$ is the covariance between $x$ and $y$, and the constants
TGV-regularization for data assimilation

\[ C_1 = k_1 L^2, \quad C_2 = k_2 L^2, \] with \( k_1 = 0.01, \quad k_2 = 0.03 \) and \( L = 2 \). It is easy to note that \( \text{SSIM} = 1 \) if the two functions are identical. Therefore, we will look for a set of parameters that guarantee the solution to be as close as possible to 1. Figure 2 shows the value of the SSIM index as we vary the parameters for the TV and TGV regularizations.

Now we choose the maximum value of SSIM for each problem and the solutions obtained for the best cases in terms of the SSIM index are presented in Figure 3. The best TV solution was obtained by setting \( \beta = 0.85 \) and the algorithm converges after 21 iterations with \( \gamma = 1e5 \) and \( \text{SSIM} = 0.9495 \). On the other hand, the best TGV solution was obtained with \( \alpha = 23.5 \) and \( \beta = 0.611 \) and the algorithm converges after 15 iterations with \( \gamma = 1e4, \mu = 1e-10 \) and the value of \( \text{SSIM} = 0.9581 \). Furthermore, Figure 4 shows the way the optimal state obtained with the TGV regularization fits with the given observations for three specific snapshots.

7.2. Performance of the algorithm. This experiment is devoted to the study of the performance of the algorithm concerning its rate of convergence, global convergence, behavior with respect to the parameters and generation of descent directions.
We use different exact solutions, shown in Figure 5. Furthermore, for each experiment we present the best solution (in terms of the SSIM) obtained. For Experiment 1, we used the parameters $\alpha = 5$ and $\beta = 0.075$, after 15 iterations the best solution obtained with a SSIM factor of 0.9592 is presented in the lower left part of Figure 5. The best solution showed in the lower middle part of Figure 5, is the same showed in the previous subsection. Finally, the best solution for Experiment 3 was obtained by using the parameters $\alpha = 2.5$ and $\beta = 0.075$ and the solution obtained has a SSIM factor of 0.9594.

For the global convergence of the algorithm, we use different exact solutions and provide the algorithm with different starting points. In particular, we use
5 different initial points which are \( u^0 \equiv 0.5, 1, 2 \), \( u^0 \) randomly generated with a uniform distribution in the interval \([0, 1]\) and \( u^0 \) as the solution of the problem with the TV regularization. Figure 6 shows that, in each case, the same value of the objective function is obtained when the algorithm stops. Moreover, this experiment allows us to show numerically that in all the cases the algorithm generates in fact descent directions. We use a fixed value of the parameters \( \alpha, \beta \) as follows: for experiments 1 and 2 we use \( \alpha = 10 \) and \( \beta = 0.2 \) and for experiment 3, \( \alpha = 5, \beta = 0.1 \). Furthermore, we use the parameters \( \gamma = 1e4, \mu = 1e - 10 \).

Another critical aspect in the study of the performance of the algorithm is the analysis of the rate of convergence. Since the proposed method is a second order method, it is natural to expect a local superlinear order, and that is what is analyzed in the following experiment. For this purpose, we solve the same problems as in the previous case but setting the initial value as \( u^0 \equiv 1 \), and we vary the parameters \( \alpha, \beta \). Figure 7 shows the decreasing of the residual computed as

\[
\text{RES} = \frac{\|u_k - u^*\|}{\|u_{k-1} - u^*\|},
\]

where we use \( u^* \) as the final value to which the algorithm converges. Furthermore, this figure suggests that the algorithm converges locally superlinearly, and since these results have been obtained by varying the values of \( \alpha \) and \( \beta \), this behavior appears to be independent of the value of the regularization parameters.
We study the behavior of the algorithm when we vary the value of $\mu$. From the theoretical point of view, the addition of this parameter is important to guarantee the generation of descent directions. However, we are also aware that adding this term is only an artifice. We solve two experiments using as exact solutions the ones depicted in Figure 8. Both experiments were solved with $\alpha = 2.5$, $\beta = 0.075$ and $\gamma = 1e4$. The results are presented in Table 1, where NaN means that the algorithm could not find a descent direction due to the singularity of the matrix. Adding the term involving the quadratic norm of $w$ let us obtain a positive definite matrix and, therefore, nonsingular. In any iteration, the algorithm actually finds a descent direction. Furthermore, Table 1 also shows that we can use small values for the parameter $\mu$ without significantly changing the optimal value of the objective function. Finally, we also present the time of execution for these experiments. The computations were performed in a CPU Intel Core i7-6700, with 3.40 GHz, with a storage of 8MB.

![Experiment 3](image1)

![Experiment 4](image2)

**Figure 8.** Exact solutions for the experiment

### 7.3. Influence of the number of observations and the background covariance matrix

The last experiment aims to investigate the influence of the number
Table 1. Results of the experiments with different values of $\mu$

| Experiment | $\mu$   | iter | SSIM  | $J_{\gamma}$ | time (s) |
|------------|---------|------|-------|--------------|----------|
| 3          | 0       | NaN  | NaN   | 35.8330      | 7.2      |
|            | 1e-6    | 13   | 0.9594| 35.8330      | 6.9      |
|            | 1e-8    | 12   | 0.9594| 35.8330      | 6.9      |
|            | 1e-10   | 12   | 0.9594| 35.8333      | 7.1      |
|            | 1e-12   | 13   | 0.9594| 35.8329      | 7.1      |
| 4          | 0       | NaN  | NaN   | 35.8330      | 7.6      |
|            | 1e-6    | 14   | 0.9592| 39.2436      | 7.6      |
|            | 1e-8    | 13   | 0.9592| 39.2434      | 7.0      |
|            | 1e-10   | 13   | 0.9592| 39.2434      | 7.0      |
|            | 1e-12   | 14   | 0.9591| 39.2437      | 7.7      |
|            | 1e-14   | 16   | 0.9591| 39.2439      | 9.1      |

Table 2. Summary of the experiment of observations and the value of the background covariance matrix in the quality of the solutions, and how these factors affect the performance of the algorithm. The results are presented in Table 2, where a weak dependence of the number of iterations with respect to the number of observations can be observed. For the experiment we use the exact solution from Experiment 2 and the parameters $\alpha = 23.5$, $\beta = 0.611$ and $\mu = 1e-10$. We vary the number of time instants and spatial points at which we obtain the observations. Finally, in order to reduce the value of the noise we vary the value $\sigma$ within the background covariance matrix $B = \sigma I$.

Acknowledgments. This research has been partially supported by the Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) through grant PIC-015-INAMHI-001: “Sistema de pronóstico del clima y el tiempo para todo el territorio ecuatoriano: modelización numérica y estadística”. E. Loayza-Romero acknowledges partial support by the German Academic Exchange Service (DAAD).

REFERENCES

[1] A. Apte, D. Auroux and M. Ramaswamy, Variational data assimilation for discrete Burgers equation, Proc. Electron. J. Differential Equations, 19 (2010), 15–30, Available form http://ejde.math.txstate.edu
[2] K. Bredies, Y. Dong and M. Hintermüller, Spatially dependent regularization parameter selection in total generalized variation models for image restoration, *Int. J. Comput. Math.*, 90 (2013), 109–123.

[3] K. Bredies, K. Kunisch and T. Pock, Total generalized variation, *SIAM J. Imaging Sci.*, 3 (2010), 492–526.

[4] K. Bredies and T. Valkonen, Inverse problems with second-order total generalized variation constraints, *Proc. SampTA*, 201 (2011), Available from: https://imsc.uni-graz.at/bredies/papers/SampTA2011.pdf

[5] C. J. Budd, M. A. Freitag and N. K. Nichols, Regularization techniques for ill-posed inverse problems in data assimilation, *Comput. & Fluids*, 46 (2011), 168–173.

[6] T. F. Chan, G. H. Golub and P. Mulet, A nonlinear primal-dual method for total variation-based image restoration, *SIAM J. Sci. Comput.*, 20 (1999), 1964–1977.

[7] P. G. Ciarlet, *Linear and Nonlinear Functional Analysis with Applications*, SIAM, 2013.

[8] J. C. De los Reyes, *Numerical PDE-constrained Optimization*, Springer, 2015.

[9] J. C. De los Reyes, C. B. Schönlieb and T. Valkonen, Bilevel parameter learning for higher-order total variation regularisation models, *J. Math. Imaging Vision*, 57 (2017), 1–25.

[10] J. E. Dennis Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, 1996.

[11] M. A. Freitag, N. K. Nichols and C. J. Budd, Resolution of sharp fronts in the presence of model error in variational data assimilation, *Q. J. Royal Meteor. Soc.*, 139 (2013), 742–757.

[12] M. Hintermüller and G. Stadler, An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration, *SIAM J. Sci. Comput.*, 28 (2006), 1–23.

[13] M. Hinze, R. Pinnau, M. Ulbrich and S. Ulbrich, *Optimization with PDE Constraints*, Mathematical Modelling: Theory and Applications, 23. Springer, New York, 2009.

[14] E. Kalnay, *Atmospheric Modeling, Data Assimilation, and Predictability*, Cambridge University Press, 2003.

[15] F. Knoll, K. Bredies, T. Pock and R. Stollberger, Second order total generalized variation (TGV) for MRI, *Magn. Reson. Med.*, 65 (2011), 480–491.

[16] J. Lee and P. K. Kitanidis, Bayesian inversion with total variation prior for discrete geologic structure identification, *Water Resour. Res.*, 49 (2013), 7658–7669.

[17] J. M. Lewis, S. Lakshminarahan and S. K. Dhall, *Dynamic Data Assimilation: A Least Squares Approach*, Cambridge University Press, 2006.

[18] J. L. Mueller and S. Siltanen, *Linear and Nonlinear Inverse Problems with Practical Applications*, SIAM, 2012.

[19] S. Pfaff and S. Ulbrich, Optimal boundary control of nonlinear hyperbolic conservation laws with switched boundary data, *SIAM J. Control Optim.*, 53 (2015), 1250–1277.

[20] A. Quarteroni, R. Sacco and F. Saleri, *Numerical Mathematics*, Springer-Verlag Italia, Milan, 1998.

[21] C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, 2002.

[22] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, *IEEE Trans. Image Process.*, 13 (2004), 600–612.

Received May 2018; revised November 2018.

E-mail address: juan.delosreyes@epn.edu.ec
E-mail address: estefania.loayza@mathematik.tu-chemnitz.de