Spring 2022

# Variational data assimilation for two interface problems

Xuejian Li



Department: Mathematics and Statistics

# VARIATIONAL DATA ASSIMILATION FOR TWO INTERFACE PROBLEMS

by

XUEJIAN LI

A DISSERTATION

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

MATHEMATICS WITH COMPUTATIONAL AND APPLIED MATHEMATICS EMPHASIS

2022

Approved by:

Xiaoming He, Advisor
Daozhi Han
John R. Singler
Mingzhen Wei
Yanzhi Zhang

# ABSTRACT

Variational data assimilation (VDA) is a process that uses optimization techniques to determine an initial condition of a dynamical system such that its evolution best fits the observed data. In this dissertation, we develop and analyze the variational data assimilation method with finite element discretization for two interface problems, including the Parabolic Interface equation and the Stokes-Darcy equation with the Beavers-Joseph interface condition. By using Tikhonov regularization and formulating the VDA into an optimization problem, we establish the existence, uniqueness and stability of the optimal solution for each concerned case. Based on weak formulations of the Parabolic Interface equation and Stokes-Darcy equation, the dual method and Lagrange multiplier rule are utilized to derive the first order optimality system (OptS) for both the continuous and discrete VDA problems, where the discrete data assimilations are built on certain finite element discretization in space and the backward Euler scheme in time. By introducing auxiliary equations, rescaling the optimality system, and employing other subtle analysis skills, we present the finite element convergence estimation for each case with special attention paid to recovering the properties missed in between the continuous and discrete OptS. Moreover, to efficiently solve the OptS, we present two classical gradient methods, the steepest descent method and the conjugate gradient method, to reduce the computational cost for well-stabilized and ill-stabilized VDA problems, respectively. Furthermore, we propose the time parallel algorithm and proper orthogonal decomposition method to further optimize the computing efficiency. Finally, numerical results are provided to validate the proposed methods.

## ACKNOWLEDGMENTS

I give my sincere thanks to my grandparents and my parents for their persistent love and support. You have shaped me into who I am and enabled me to stay tough at every age.

I would like to thank my advisor, Dr. Xiaoming He, for introducing me to the academic world with immense support and guiding. I especially appreciate your great patience and personality and for helping to keep me on track. I appreciate the infinite academic freedom you gave to me which prepared me to stay open for pursuing any research possibilities. It was an unique experience working with you.

Thanks to Dr. John R. Singler for listening to my questions, bringing clarity to my confusions, and bringing me into a new research field. I am grateful for all discussions and support from you. Thanks to Dr. Yanzhi Zhang and Dr. Daozhi Han for all discussions about my research and daily life. Thanks to Dr. Mingzhen Wei for listening to my presentations and giving me meaningful advises. A special thank to Dr. Wei Gong for many valuable discussions on optimizations, optimal control, numerical methods and analysis.

I wish I express my sincere gratitude to my friend and roommate, John Carter, and his family. Thank you for providing me with memorable moments and for putting up with my monochromatic lifestyle.

I would also like to extend a thank you to my friends Louis Steinmeister and Austin Vandegriffe for their great conversations and for exposing me to new things. I appreciate all of the times we had together.

Finally, This work is partially supported by National Science Foundation grants DMS-1722647 and DMS-2111421.

## TABLE OF CONTENTS

Page

## LIST OF TABLES

Table Page

# 1. INTRODUCTION

Data assimilation seeks to incorporate observations into a physics model for attaining the best possible estimate of the state forecast, subject to initial conditions. This process results in data driven initial conditions for PDEs, yielding more accurate forecasts. It can improve models in numerous fields, such as weather forecast [1–3], geoscience [4–7], ocean flow prediction [8–10], and biology transport [11–13]. Currently there are two main categories of data assimilation techniques. The first category includes the statistical methods based on the Bayes' Theorem and the Kalman filtering approach, which evolve the state vector along with time according to error statistics [14–20]. The second category includes the variational methods based on the optimal control theory, which minimizes a cost functional measuring the discrepancy between the state variable and the observed data [21–27]. Besides, some other data assimilation techniques are also very popular for certain problems in last few decades, such as nudging method [28, 29] and continuous data assimilation method [30–32].

In this dissertation we discuss the variational method to solve the data assimilation problem for interface problems. The variational methods are first introduced by J L. Lions in optimal control to estimate relevant model parameters for parabolic equations [33]. Afterwards, a vast amount of literature employing variational methods has been contributed to investigate the data assimilation problem for different physical models both theoretically and computationally, see, e.g., [34–37]. However, as far as we are aware, there is not a thorough consideration of the data assimilation for the interface problems. Therefore, a primary interest in this dissertation is to investigate the variational data assimilation for the interface problems, such as the Parabolic Interface equation and Stokes-Darcy equation. We will focus on identifying a faithful initial condition for the physical model such that the target state can be better predicted. Our approach to achieve this goal is by incorporating the deterministic or noisy measurement into interface equation through an appropriately designed cost functional. The optimization theories and techniques play key roles dealing with the challenges throughout the simulation process.

In Section 2, we introduce fundamental concepts, notations, and theories often used in the discussion of the variational data assimilation.

In Section 3, we start our data assimilation journey from a PDE constrained optimization problem, and introduce basic ingredients for the well-posedness analysis [33]. We recall two powerful approaches to derive the optimal condition, i.e., dual method and Lagrange multiplier rule. Meanwhile, the finite element methods are proposed to discretize the optimization problems, along with a mathematic derivation of its optimality system. To handle the extreme computational cost arising from the variational data assimilation, we discuss two classical gradient methods, steepest descent method and conjugate gradient method, to address the large-scale computing difficulties. Secondly, based on a multiple shooting method, we present the time parallel algorithm to further improve the efficiency and increase the solving flexibility. Finally, the proper orthogonal decomposition methods are utilized to overcome the memory challenges and optimize the computational resource in variational data assimilation.

In Section 4, a VDA method is proposed and analyzed for the Parabolic Interface equation [38, 39] that models important physical phenomena when two or more distinct materials or fluids with different conductivities or diffusions are involved. By using Tikhonov regularization and optimization formulation, the existence, uniqueness, and stability of the optimal solution are established. The optimality system for both continuous and discrete VDA are derived, where the discrete VDA is built on a finite element discretization. The finite element convergence with the optimal error estimate is proved with special attention paid to the recovery of Galerkin orthogonality. Lastly, we provide the detail of implementing different numerical algorithms and use numerical experiments to validate the proposed methods.

In Section 5, due to numerous attention and potential applications received by the Stokes-Darcy model in recent decades [40–47], we investigate the variational data assimilation for the Stokes-Darcy equation with Beavers-Joseph interface condition. Based on a proper mathematical interpretation of the Stokes-Darcy equation, we formulate the VDA into a constrained optimization problem and present the well-posedness analysis. We derive the first order optimality system using Lagrange multiplier rule for both the continuous and discrete VDA problems. Again, the discrete VDA utilizes a finite element method. By rescaling the optimality system and analyzing its fundamental operator properties, we prove the optimal finite element convergence rate via introducing relevant axillary equations and recovering uncertainties missed in the OptS. The numerical algorithms are presented in detail and numerical experiments are provided at the end.

In Section 6, we draw conclusions and look forward to the future works.

This dissertation consists of material partially from one submitted paper [48] and another to be submitted paper [49]. Some minor changes to the papers are added in this dissertation in order to provide more details and increase the readability.

# 2. MATHEMATICAL PREMILINARIES

In this section, we are going to provide the necessary mathematics preliminaries for discussion of partial differential equations, numerical methods, optimization, and data assimilation.

## 2.1. BASIC CONCEPTS

Let $\|\cdot\|_X$ denote the norm for a normed vector space $X$, $(\cdot,\cdot)$ denote the inner product for Hilbert spaces, $\langle\cdot,\cdot\rangle$ represents a general duality paring between a Banach space and its dual space (see definition 2). Then, we introduce preliminaries for this dissertation as follows.

**Definition 1** *Let $X$ and $Y$ be normed vector spaces, a linear operator $T : X \to Y$ is said to be continuous if there exists a constant C such that*

$$\|Tx\|_Y \leq C\|x\|_X.$$

We use $\mathscr{L}(X,Y)$ to denote a set of bounded linear operators from $X$ to $Y$, with which the operator norm is defined as:

$$\|T\|_{\mathscr{L}(X,Y)} = \sup_{0\neq x\in X} \frac{\|Tx\|_Y}{\|x\|_X} = \sup_{\|x\|_X=1} \|Tx\|_Y = \sup_{\|x\|_X\leq 1} \|Tx\|_Y.$$

Definition 1 indicates that boundedness and continuity are equivalent for a linear operator. In addition, the continuity at $x = 0 \in X$ of a linear operator gives the boundedness as well. Furthermore, one can verify that $\mathscr{L}(X,Y)$ is a Banach space if $Y$ is a Banach spaces.

For a special set of linear operator $T \in \mathscr{L}(X,\mathbb{R})$, $T$ is also called a linear functional on $X$, which gives the definition of the dual space of $X$.

**Definition 2** *Let X be a normed vector space, we define X'* [1] $:= \mathscr{L}(X, \mathbb{R})$ *as the dual space of* $X$*. For* $T \in X'$ *and* $x \in X$*, we also write*

$$Tx = \langle T, x \rangle.$$

In Banach space, the dual space usually can be identified by another Banach space. For example,

$$L^p(\Omega) \cong L^q(\Omega), \ \text{ for } 1 < p < \infty, \ \frac{1}{p} + \frac{1}{q} = 1,$$

note that $(L^1(\Omega))' = L^\infty(\Omega)$, but $L^1(\Omega) \neq (L^\infty(\Omega))'$. Another example is that the continuous function space $C(\Omega)$ can be represented by the bounded variation function space $BV(\Omega)$, i.e., $C(\Omega) \cong BV(\Omega)$. Specifically, in Hilbert space, we can observe even nicer representation.

**Theorem 1 (Riesz representation )** *[[50]] Let* $H$ *be a Hilbert space and let* $f : H \to R$ *be a bounded linear functional. Then there exists a unique element* $y \in H$ *such that*

$$\langle f, x \rangle = (y, x), \ \ \forall x \in H.$$

*This element* $y \in H$ *satisfies*

$$\|f\|_{H'} = \|y\|_H = \sup_{x \in H/\{0\}} \frac{(y, x)}{\|x\|_H}.$$

*Thus the map* $H \to H' : y \mapsto (y, \cdot)$ *is an isometry of normed vector spaces.*

[1] In this dissertation, X′ refers the dual of X if X is a Banach space, X′ also refers to the Gâteaux or Fréchet derivative operator when X is a mapping.

**Definition 3** *Let $A : X \to Y$ be a bounded linear operator, where $X$ and $Y$ are normed vector spaces. The adjoint operator of A is defined as $A^* : Y' \to X'$:*

$$\langle g, x\rangle = \langle A^*T, x\rangle = \langle T, Ax\rangle, \quad \forall T \in Y'.$$

**Definition 4** *Let $A : X \to Y$ be a bounded linear operator, where $X$ and $Y$ are Hilbert spaces. The adjoint operator of A is defined as $A_H^* : Y \to X$:*

$$(A_H^* y, x) = (y, Ax), \quad \forall x \in X, y \in Y.$$

The adjoint operator in Hilbert space is, to some extent, identical with the adjoint definition in general normed vector space. This can be seen by the isometric relation:

$$A^* : Y' \cong Y \to X' \cong X.$$

For this reason, we will not distinguish $A_H^*$ and $A^*$ from now on, only using $A^*$ to denote the adjoint operator of a bounded linear operator $A$ either in Hilbert space or Banach space. One can further verify some basic properties of the adjoint operator:

- $(\alpha A)^* = \alpha A^*$.
- $(A + B)^* = A^* + B^*$.
- $(AB)^* = B^* A^*$.
- $(A^{-1})^* = (A^*)^{-1}$.

In the following, we define the convergence and continuity in a weak sense, which coincide with the strong convergence and continuity in the finite dimension space. However, they behave differently in the infinite dimension space.

**Definition 5** *Let $X$ be a normed vector space and $\{x_n\}_{n\in N^+}$ be a sequence in $X$, $x_n$ weakly converges to $x \in X$ if and only if*

$$\langle f, x_n \rangle = \lim_{n\to\infty} \langle f, x \rangle, \quad \forall f \in X'.$$

In Hilbert space $H$, the weak convergence $x_n \rightharpoonup x$ is equivalent to

$$(y, x_n) = (y, x), \quad \forall y \in H,$$

which is a consequence of the Riesz representation, i.e., $H \cong H'$.

Similarly, we need weak* convergence, this is useful when weakly convergence fails in some considerations.

**Definition 6** *Let $X$ be a normed vector space and $\{f_n\}_{n\in N^+}$ be a sequence in $X'$. Then $f_n$ converges to $f \in X'$ in a weak* sense if and only if*

$$\langle f_n, x \rangle = \lim_{n\to\infty} \langle f, x \rangle, \quad \forall x \in X.$$

**Definition 7** *Let $U$ be a subset of a Banach space. We say that the functional $f : U \to \mathbb{R}$ is (sequentially) weakly/weak* lower-semicontinuous on $U$ if for every sequence $\{u_k\}_{k\in N^+} \subset U$ converging weakly/weak* to $u \in U$, we have that*

$$f(u) \leq \liminf_{k\to\infty} f(u_k).$$

**Remark 1** *The weakly lower semi-continuouity is more often used in optimization problem. We recall the following useful results.*

- *Convex, (lower) continuous functional is weakly lower semi-continuous.*

- *Norm is convex and continuous.*

It is worth mentioning that the dual space $X'$ of a normed vector space is a Banach space, no matter if $X$ is complete or not. It is then helpful to conceive a bidual space of $X$ for purpose of investigating the structure of $X$.

**Definition 8** *Let $X$ be a real normed vector space, the bidual space of $X$ is the dual space of $X'$ and denoted by*

$$X'' = (X')' = \mathscr{L}(X', \mathbb{R}).$$

There is a natural mapping $\mathscr{C} : X \mapsto X''$ defined as

$$\langle \mathscr{C}_x, f \rangle = \langle f, x \rangle,$$

where $\mathscr{C}$ is called the canonical mapping. By the definition of dual space, we may not be able to expect that there exists a one-to-one onto correspondence between $X$ and $X'$ in general. However, this bijection can be preserved for some normed vector spaces.

**Definition 9** *A real normed vector space $X$ is called reflexive if the canonical mapping $\mathscr{C}$ is bijective.*

We recall some useful conclusions in reflexive normed vector spaces [50].

**Corollary 1** *The following statement is true for the reflexive normed vector space:*

*(i) Reflexive space is complete.*

*(ii) $X$ is reflexive iff $X'$ is reflexive.*

*(iii) Finite dimensional space is reflexive.*

*(iv) $L^p$ is reflexive for $1 < p < \infty$.*

*(v) Hilbert space $H$ is reflexive, i.e., $H \cong H' \cong H''$.*

**Theorem 2 (Eberlin-Šmulian Theorem)** *Let $X$ be a real Banach space and*

$$B := \{x \in X : \|x\| \leq 1\}$$

*be the closed unit ball. Then the following are equivalent [50].*

*(i) X is reflexive.*

*(ii) B is weakly compact.*

*(iii) B is sequentially weakly compact.*

*(iv) Every bounded sequence in X has a weakly convergent subsequence.*

**Theorem 3 (Banach–Alaoglu Theorem)** *[50] Let X be a separable real normed vector space. Then every bounded sequence in the dual space $X'$ has a weak$^*$ convergent subsequence.*

Note that the result in Theorem 3 can be extended to no-separable normed vector space.

At the end of this section, we introduce the bilinear form that will be often used in the quadratic optimization and partial differential equation.

**Definition 10** *Let X and Y be normed vector spaces over the field $\mathbb{R}$. A bilinear form is a function $b : X \times Y \to \mathbb{R}$ such that*

$$b(\alpha_1 x_1 + \alpha_2 x_2, y) = \alpha_1 b(x_1, y) + \alpha_2 b(x_2, y),$$

$$b(x, \alpha_1 y_1 + \alpha_2 y_2) = \alpha_1 b(x, y_1) + \alpha_2 b(x, y_2),$$

*where $\alpha_1, \alpha_2 \in \mathbb{R}, x_1, x_2 \in X$, and $y_1, y_2 \in Y$.*

The bilinear form $b(\cdot,\cdot)$ is bounded if satisfying the following condition:

$$|b(x, y)| \leq C\|x\|_X\|y\|_Y, \quad \forall x \in X, y \in Y,$$

and its norm is defined as

$$\|b\|_b = \sup_{0\neq x\in X, 0\neq y\in Y} \frac{|b(x, y)|}{\|x\|_X\|y\|_Y} = \sup_{\|x\|_X=1, \|y\|_Y=1} |b(x, y)|.$$

**Corollary 2 (Reisz Representation)** *[50, Chapter 3] Let X and Y be Hilbert spaces and $b : X \times Y \to \mathbb{R}$ be a bounded bilinear form. Then there exists a unique $S \in \mathscr{L}(X, Y)$ such that*

$$b(x, y) = (Sx, y) \quad \textit{and} \quad \|b\|_b = \|S\|_{\mathscr{L}(X,Y)}.$$

## 2.2. DERIVATIVES IN BANACH SPACES

Let $X$ and $Y$ be Banach spaces and $F : X \to Y$ be a mapping from $X$ to $Y$.

**Definition 11** *For an element $x \in X$, if the limit*

$$\delta F(u, h) = \lim_{s \to 0} \frac{1}{s}(F(x + sh) - F(x))$$

*exists for $h \in X$. We call $\delta F(u, h)$ a directional derivative of F at x in direction h. If this limit exists for all directions $h \in X$, then we call the mapping $h \mapsto \delta F(x, h)$ the first variation of F at x.*

**Definition 12** *F is said to be Gâteaux differentiable at x if it is directional differentiable and*

$$\delta F(x, h) = Ah, \quad \textit{with } A \in \mathscr{L}(X, Y)$$

*exists for all $h \in X$. We refer A as the Gâteaux derivative of F at $x \in X$.*

If $f : X \to R$ is a Gâteaux differentiable functional, then the Gâteaux derivative is an element of the dual space $X'$.

**Definition 13** *A mapping $F : X \to Y$ is Fréchet-differentiable if there is an operator $A \in \mathscr{L}(X,Y)$ and a mapping $r : X \times X \to Y$ such that*

$$F(x+h) = F(x) + Ah + r(x,h) \quad \text{and} \quad \frac{\|r(x,h)\|_Y}{\|h\|_X} \to 0$$

*for all $h \in X$. We refer A as the Fréchet derivative of F at x and write $F'(x) = A$.*

**Theorem 4 (Chain rule)** *Let $X, Y$ and $Z$ be Banach spaces and let $F : X \to Y$ and $G : Y \to Z$ be Fréchet-differentiable at $x \in X$, then the composition*

$$E(x) := G \circ (F(x))$$

*is also Fréchet-differentiable at $x \in X$ and*

$$E'(u) = G'(F(u)) \circ F'(u).$$

**Remark 2**

- *Gâteaux differentiable can not imply continuity.*
- *Gâteaux differentiable can not imply Fréchet-differentiable.*
- *Chain rule is not generally true for Gâteaux derivative.*
- *If the Fréchet derivative exists, so does the Gâteaux derivative and they coincide.*
- *If $F$ is Fréchet-differentiable at $x \in X$, then $F$ is continuous at $x$.*
- *If $F$ is Fréchet-differentiable at $x$, then the derivative $F'(x)$ is unique.*
- *If $F$ is continuously differentiable at $x$, the $F$ is Fréchet-differentiable at $x$.*
- *If $F' \in \mathscr{L}(X,Y)$ is also differentiable at $x$, then $F''(x) \in \mathscr{L}(X, \mathscr{L}(X,Y))$.*

Next, we provide a few examples to concretely show the calculation of derivative for a mapping in a general Banach space.

**Example 1** *The bounded linear operator $T : X \to Y$ is Fréchet-differentiable with derivative*

$$T'(x)h = Th.$$

*It is simply true by the linearity $T(x + h) = T(x) + T(h)$.*

**Example 2** *The functional $\frac{1}{2}\|u\|_{L^2(\Omega)}$ is Fréchet-differentiable and its derivative at $u$ is $\int_\Omega uhdx$ and the gradient is $u$.*

**Proof** :

*We set $J(u) = \frac{1}{2}\|u\|_{L^2(\Omega)} = \frac{1}{2}(u, u) = \int_\Omega uudx$. Then*

$$\begin{aligned} J(u + h) - J(u) &= \frac{1}{2}(u + h, u + h) - \frac{1}{2}(u, u) \\ &= \frac{1}{2}(u, u) + \frac{1}{2}(u, h) + \frac{1}{2}(h, u) + \frac{1}{2}(h, h) - \frac{1}{2}(u, u) \\ &= (u, h) + \frac{1}{2}(h, h). \end{aligned}$$

*Obviously, $\frac{\frac{1}{2}\|h\|^2}{\|h\|} \to 0$ as $h \to 0$. Therefore, $J'(u)h = (u, h)$. By Riesz representation, we obtain $J'(u) = u$.*

**Example 3** *The bilinear and continuous mapping $B(u, u) : X \times X \to Y$ is Fréchet-differentiable with derivative $B(u, h) + B(h, u)$ at $u$.*

**Proof** :

*We set $f(u) = B(u, u)$ and do the calculus variation:*

$$\begin{aligned} f(u + h) - f(u) &= B(u + h, u + h) - B(u, u) \\ &= B(u, u) + B(u, h) + B(h, u) + B(h, h) - B(u, u) \\ &= B(u, h) + B(h, u) + B(h, h). \end{aligned}$$

*Since $B(u,u) : X \times X \to Y$ is continuous, we have $\frac{\|B(h,h)\|}{\|h\|} \to 0$ as $h \to 0$. Therefore, the result holds.*

**Example 4** *The next example is an application of the Chain Rule. Find the Fréchet derivative of functional $J(u) = \int_\Omega (\hat{u} - u^2)^2 : X \to \mathbb{R}$, where $\hat{u}$ is a known function in $X$.*

**Solution** :

*First, it is not difficult to verify that $J(u)$ is Fréchet differentiable. We consider $J(u) = E \circ G(u)$, where $G(u) = u^2$ and $E(z) = \int_\Omega (\hat{u} - z)^2 dz$. By applying the chain rule, we obtain*

$$f'(u)h = \langle E'(G(u)), G'(u)h \rangle = 4 \int_\Omega (\hat{u} - u^2) u h dx,$$

*since $G'(u), E'(z)$ can be easily calculated as*

$$G'(u)h = 2uh, \quad \langle E'(z), v \rangle = 2 \int_\Omega (\hat{u} - z) v dz.$$

Let $U \in X$ be a nonempty subset of a real normed vector space $X$ and $J : U \to \mathbb{R}$ be a given functional. Define a minimization problem:

$$\min_{x \in U} J(u). \tag{2.1}$$

**Definition 14** *For $x \in U \subset X$ the direction $y - x \in U$ is called admissible if there exists a sequence $\{t_n\}_{n \in N}$ with $0 < t_n \to 0$ as $n \to \infty$, such that $x + t_n(y - x) \in U$ for every $n \in N$.*

**Theorem 5** *Suppose that $x \in U \subset X$ is a local minimum of (2.1) and that $y - x$ is an admissible direction. If $f$ is directionally differentiable at $x$, in direction $y - x$, then*

$$\delta J(x)(y - x) > 0.$$

Theorem 5 indicates the following Corollary immediately.

**Corollary 3** *If $U = X$ and $J(x)$ is Gâteaux differentiable at $x$ where the local optimal solution is obtained, then*

$$J'(x)h = 0, \quad \forall h \in X.$$

**Theorem 6 (Lagrange Multiplier Rule)** *[[51]] Let $X$ and $Y$ be real Banach spaces. Let $U$ be an open subset of $X$ and let $J : U \to \mathbb{R}$ be a continuously differentiable function. Let $g : U \to Y$ be another continuously differentiable function, the constraint: $g(u) = 0$. Suppose also that the Fréchet derivative $g'(u) : X \to Y$ of $g$ at $u$ is a surjective linear map. Then there exists a Lagrange multiplier*

$$\lambda : Y \to \mathbb{R}$$

*such that*

$$J'(u) = \lambda \circ g'(u).$$

## 2.3. MINIMIZATION PROBLEM

We conclude this Section by applying the above definitions and theories to two prototypical minimization problems.

Let $X$ be a Banach space, $U_{ad} \in X$ be a nonempty set of optimal variables, and $J : U \to \mathbb{R}$ be a functional, bounded from below. Consider the problem:

$$\min_{x \in U} J(u). \tag{2.2}$$

A urgent question to be answered is whether problem (2.2) is well-posed, i.e.,

- For all data, does there exist a solution of the problem?

- For all data, is the solution unique?

- Does the solution depend continuously on the data?

We start from the existence discussion. Since $J(u)$ is bounded below, therefore infimum exists. We can construct a minimizing sequence $\{u_n\} \in U_{ad}$ such that

$$J(u_n) \to \inf_{u \in U_{ad}} J(u).$$

If $J(u)$ is coercive, i.e., $J(u) \to \infty$ as $\|u\| \to \infty$ or $J(u)+Q \geq C\|u\|$, $Q$ is a known quantity. Then $\{u_n\}$ is a bounded sequence in $U_{ad}$. Based on Theorem 2, there exists a subsequence still denoted as $\{u_n\}$ converging weakly to $u^\star$ in $X$ if $X$ is a reflexive Banach space. Assume further that the functional $J(u)$ is weakly lower semi-continuous, we can deduce

$$J(u^\star) \leq \liminf_{k \to \infty} J(u_n) = \inf_{u \in U_{ad}} J(u),$$

which concludes that $u^\star$ is a minimizer once $U_{ad}$ is weakly closed, or convex and closed. Second, in order to show the uniqueness, assume that $J(u)$ is strictly convex and $u_1$ and $u_2$ are two optimal solutions, we will have

$$J(\frac{u_1 + u_2}{2}) < \frac{J(u_1)}{2} + \frac{J(u_2)}{2} = \inf_{u \in U_{ad}} J(u),$$

which contradicts the definition of infimum. Therefore, solution is unique.

Aside from concerning the stability, we can summarize a few necessary ingredients when modeling an optimization problem.

**Theorem 7** *Let $U_{ad}$ be a closed and convex subset of a reflexive Banach space, if a functional $J(u)$ is bounded below, coercive, weakly lower semi-continuous, and strictly convex, then there exists a unique minimizer for $J(u)$ [33].*

Next, we look at an optimization problem with a constraint. Let $Z$ be a reflexive Banach space and $X, Y$ be Banach spaces. We state the problem as

$$\min_{u,z \in X \times Z_{ad}} J(u, z) \qquad \text{subject to} \qquad e(u, z) = 0. \tag{2.3}$$

Here, $J : X \times Z_{ad} \to \mathbb{R}$, $e : X \times Z_{ad} \to Y$, and $Z_{ad}$ is convex and closed. Before preceding the wellposdness discussion, we clarify a few understandings in (2.3). The functional $J(u, z)$ usually consists of two components:

$$J(u, z) = J_1(u) + J_2(z),$$

where $J_1 : X \to \mathbb{R}$ is the objective and $J_2 : Z_{ad} \to \mathbb{R}$ is a penalty or regularization for optimal variable. For $e(u, z) = 0$, we assume $e : X \times Z_{ad} \to Y$ is smooth enough and the condition for implicit function theorem holds, i.e., $e(u, z) = 0$ is differentiable and is able to define a mapping $u(z) : Z_{ad} \to X$. Problem (2.3) is then rewritten as:

$$\min_{z \in Z_{ad}} J(z) \qquad \text{subject to} \qquad e(z) = 0.$$

**Theorem 8** *The problem (2.3) admits an unique solution if the following conditions hold:*
*(i) $Z_{ad}$ is convex and closed.*
*(ii) The equation $e(u, z) = 0 : X \times Z_{ad} \to Y$ defines a mapping $u(z)$ that is continuous with respect to $z$.*
*(iii) $J$ is continuous in term of variables $u$ and $z$, respectively.*
*(iv) $J$ is strictly convex, coercive, and bounded below.*

**Proof:**

Conditions (ii) and (iii) indicate that $J$ is continuous, the convexity further implies the weakly lower semi-continuity of $J$. Then the rest of proof is using Theorem 7.

**Remark 3** • *Coercivity is contained in convexity for some contexts, for instance, if $J$ is proper, convex, lower semi-continuous in a Hilbert space, then $J$ is also coercive.*

- *The continuity of $u(z)$ could happen in a weak sense, i.e., $z_n \rightharpoonup u$ as $u(z_n) \rightharpoonup u(z)$.*

For simplicity of the following presentation, we assume $Z_{ad} = Z$. We next committee to characterize the optimal solution via calculating the first oder variation of the functional $J(u, z)$:

$$\begin{aligned}\langle J'(u,z), h\rangle &= \langle J_u(u,z), u'(z)h\rangle + \langle J_z(u,z), h\rangle \\ &= \langle (u'(z))^* J_u(u,z), h\rangle + \langle J_z(u,z), h\rangle.\end{aligned}$$

To further understand the structure of $(u'(z))^*$, we do calculus variation with respect to $z$ of the constraint equation $e(u, z) = 0$:

$$e_u(u,z)u'(z)h + e_z(u,z)h = 0,$$

which gives

$$u'(z)h = -(e_u(u,z))^{-1} e_z(u,z)h,$$

and

$$(u'(z))^* = -\left(e_z(u,z)\right)^* \left((e_u(u,z))^{-1}\right)^* = \left(e_z(u,z)\right)^* \left((e_u(u,z))^*\right)^{-1}. \tag{2.4}$$

Substituting (2.4) into the first order variation of the functional $J$, we obtain

$$\begin{aligned}\langle J'(u,z), h\rangle &= \langle (u'(z))^* J_u(u,z), h\rangle + \langle J_z(u,z), h\rangle \\ &= \langle -\left(e_z(u,z)\right)^* \left((e_u(u,z))^*\right)^{-1} J_u(u,z), h\rangle + \langle J_z(u,z), h\rangle.\end{aligned}$$

It is now natural to introduce the adjoint equation

$$(e_u(u,z))^* u^* = J_u(u,z)$$

to simplify the first order variation as:

$$\langle J'(u,z),h\rangle = \langle -(e_z(u,z))^* u^*, h\rangle + \langle J_z(u,z),h\rangle.$$

The first order optimal condition is concluded by sufficing

$$\langle -(e_z(u,z))^* u^* + J_z(u,z),h\rangle = 0.$$

# 3. VARIATIONAL DATA ASSIMILATION AND ALGORITHMS

## 3.1. FORMULATION OF VARIATIONAL DATA ASSIMILATION

Let $U_{ad}$ denote an admissible solutions set that could be either a Hilbert space $H(\Omega)$ or a closed convex subset of $H(\Omega) \subset L^2(\Omega)$, where $\Omega$ is an open bounded domain with regular boundary $\partial\Omega$. Given $T > 0$, $\gamma > 0$, and a distributed observation $\hat{u} \in L^2(0,T;\mathcal{H}(\Omega)) \subset L^2(0,T;V'(\Omega))$, the data assimilation for a given dynamical system is considered as [33]:

$$\min_{u_0 \in U_{ad}} J(u_0) = \frac{1}{2}\int_0^T \|\hat{u} - u(u_0)\|^2_{\mathcal{H}} dt + \frac{\gamma}{2}\|u_0\|^2_H, \tag{3.1}$$

subject to

$$\begin{cases} \dfrac{\partial u}{\partial t} + Au = f \in L^2(0,T;V'(\Omega)), \\ u(0) = u_0 \in L^2(\Omega), \end{cases} \tag{3.2}$$

where $\mathcal{H}(\Omega)$ and $V(\Omega)$ are appropriate Hilbert spaces, $A$ is a linear operator describing some physical phenomena, and the mapping $u(u_0)$ is defined as the solution of (3.2) with the initial value $u_0$. By incorporating the $\hat{u}$ with the output of the dynamical system (3.2) through the cost functional (3.1), our purpose is to identify a reliable initial condition $u_0$ for a better state predictions via minimizing the cost functional (3.1).

Since (3.1)-(3.2) can be viewed as an optimization problem, we use the argument developed in Chapter 2 to investigate whether (3.1)-(3.2) is well-posed.

- $U_{ad}$ is a closed and convex, or weakly closed.
- Cost functional (3.1) is non-negative, therefore bounded below .
- Cost functional (3.1) is obviously coercive due to the term $\frac{\gamma}{2}\|u_0\|^2$.
- Assuming the wellposdeness of (3.2), the mapping $u(u_0)$ is a continuous mapping.

- By the continuity of norm, cost functional (3.1) is continuous.
- From the linearity of (3.2) and the strict convexity of norm for Hilbert space, we can verify the cost functional (3.1) is strictly convex.

Apparently, the data assimilation problem (3.1)-(3.2) satisfies all ingredients as an optimization problem for the Theorem 8, hence admits an unique optimal solution. Furthermore, by calculating the first order derivative of the cost functional (3.1), the optimal solution $u_0$ can be characterized by

$$\begin{aligned}\langle F'(u_0), v - u_0\rangle = \int_0^T ((u(u_0) - \hat{u}, u(v) - u(u_0))dt \\ + \gamma(u_0, v - u_0) \geq 0, \qquad \forall v \in U_{ad}.\end{aligned} \tag{3.3}$$

Next, we show that the solution of problem (3.1)-(3.2) is stable with respect to the perturbation on the distributed observations and the regularization parameter $\gamma$.

**Theorem 9** *The solution of problem* (3.1)*-*(3.2) *continuously depends on the observational data* $\hat{u}$ *and the parameter* $\gamma$*.*

**Proof:** Introducing perturbations $\epsilon_1 \in \mathbb{R}$ on $\gamma$ and $\epsilon_2 \in L^2(0, T; L^2(\Omega))$ on $\hat{u}$ respectively, and letting $\bar{u}_0$ denote the perturbed optimal solution, we then have

$$\begin{aligned}&\int_0^T (u(\bar{u}_0) - \hat{u} - \epsilon_2, u(v) - u(\bar{u}_0))dt \\ &+ (\gamma + \epsilon_1)(\bar{u}_0, v - \bar{u}_0) \geq 0 \qquad \forall v \in U_{ad}.\end{aligned} \tag{3.4}$$

Taking $v = u_0$ in (3.4) and $v = \bar{u}_0$ in (3.3) gives us

$$\int_0^T (u(\bar{u}_0) - \hat{u} - \epsilon_2, u(u_0) - u(\bar{u}_0))dt + (\gamma + \epsilon_1)(\bar{u}_0, u_0 - \bar{u}_0) \geq 0,$$

$$\int_0^T (u(u_0) - \hat{u}, u(\bar{u}_0) - u(u_0))dt + \gamma(u_0, \bar{u}_0 - u_0) \geq 0.$$

Adding the two inequalities together leads to

$$\begin{aligned}&\int_0^T \|u(u_0) - u(\bar{u}_0)\|_{\mathcal{H}}^2 dt + (\gamma + \epsilon_1)\|(u_0 - \bar{u}_0)\|_H^2 \\ &\leq \int_0^T (\epsilon_2, u(\bar{u}_0) - u(u_0))dt + \epsilon_1(u_0, u_0 - \bar{u}_0).\end{aligned} \tag{3.5}$$

Applying the Cauchy-Schwarz inequality and Young's inequality for the right hand side terms in (3.5), we have

$$\int_0^T (\epsilon_2, u(\bar{u}_0) - u(u_0))dt \leq \int_0^T \|u(u_0) - u(\bar{u}_0)\|_{\mathcal{H}}^2 dt + \frac{1}{4}\|\epsilon_2\|_{L^2(0,T;\mathcal{H}(\Omega))}^2, \tag{3.6}$$

$$\epsilon_1(u_0, u_0 - \bar{u}_0) \leq \frac{|\epsilon_1|}{2}\|u_0\|_H^2 + \frac{|\epsilon_1|}{2}\|u_0 - \bar{u}_0\|_H^2. \tag{3.7}$$

Combining (3.5)-(3.7) and setting $|\epsilon_1| \leq \frac{\gamma}{3}$, we have the inequality

$$\frac{\gamma}{2}\|u_0 - \bar{u}_0\|_H^2 \leq \frac{1}{4}\|\epsilon_2\|_{L^2(0,T;\mathcal{H}(\Omega))}^2 + \frac{|\epsilon_1|}{2}\|u_0\|_H^2, \tag{3.8}$$

which implies that the solution of problem (3.1)-(3.2) continuously depends on the observational data $\hat{u}$ and $\gamma$.

Moreover, inequality (3.8) indicates that small $\gamma$ will degrade the stability of the data assimilation problem.

Since the arguments above deeply rely on optimal control theory [33], we refer such approach as the variational data assimilation method.

The next step is to find out the optimal solution, we will present two main techniques, dual method and Lagrange multiplier rule, to derive the optimality system.

## 3.2. DERIVATION OF THE OPTIMALITY SYSTEM

**3.2.1. Dual Method.** To begin with, based on equation (3.2), we define the constraint operator $M : L^2(\Omega) \times L^2(0,T;V(\Omega)) \to L^2(\Omega) \times L^2(0,T;V'(\Omega))$:

$$M\,(u_0, u) = \begin{pmatrix} \frac{\partial u}{\partial t} + Au - f \\ u(0) - u_0 \end{pmatrix} \in L^2(0,T;V'(\Omega)) \times L^2(\Omega).$$

We then do standard calculus variation for the cost functional $J(u_0, u)$ with respect to $u_0$

$$\langle \frac{\partial J(u_0,u)}{\partial u_0}, h\rangle = \langle J_{u_0}(u_0,u), h\rangle + \langle J_u(u_0,u), \frac{\partial u}{\partial u_0} h\rangle, \quad \forall h \in U_{ad}. \tag{3.9}$$

A calculus variation of the constraint equation $M\,(u_0, u) = 0$ with respect to $u_0$ leads to

$$M_{u_0} h + M_u \tfrac{\partial u}{\partial u_0} h = 0, \tag{3.10}$$

where

$$M_{u_0} h = \begin{pmatrix} 0 \\ -h \end{pmatrix}, \quad M_u \tfrac{\partial u}{\partial u_0} h = \begin{pmatrix} (\frac{\partial}{\partial t} + A) \frac{\partial u}{\partial u_0} h \\ \frac{\partial u(0)}{\partial u_0} h \end{pmatrix}.$$

Recall that

$$M_{u_0} : L^2(\Omega) \to L^2(0,T;V(\Omega)) \times L^2(\Omega),$$
$$M_u : L^2(0,T;V(\Omega)) \to L^2(0,T;V'(\Omega)) \times L^2(\Omega),$$
$$J_u(u_0,u) : L^2(0,T;V(\Omega)) \to \mathbb{R}.$$

In order to connect (3.9) and (3.10), it is natural to introduce the adjoint operator, i.e.,

$$M_u^* : L^2(0,T;V''(\Omega)) \times (L^2(\Omega))' \to L^2(0,T;V'(\Omega)),$$

then there exists a element $z \in L^2(0,T;V''(\Omega)) \times (L^2(\Omega))'$ such that

$$M_u^* z = J_u(u_0, u). \tag{3.11}$$

Using the equation (3.10), we deduce the following operator calculation

$$\langle M_u^* z, \tfrac{\partial u}{\partial u_0} h\rangle = \langle z, M_u \tfrac{\partial u}{\partial u_0} h\rangle = \langle z, -M_{u_0} h\rangle = \langle z, -\begin{pmatrix} 0 \\ -h \end{pmatrix}\rangle = \langle u^*(0), h\rangle, \tag{3.12}$$

where $z \in L^2(0,T;V(\Omega)) \times L^2(\Omega)$ defined as $z = \begin{pmatrix} u^* \\ u^*(0) \end{pmatrix}$ is the identity of $z \in L^2(0,T;V(\Omega)'') \times (L^2(\Omega))'$ by the reflexive property of Hilbert space.
Then we have

$$\langle J_u(u_0, u), \frac{\partial u}{\partial u_0} h\rangle = \langle u^*(0), h\rangle.$$

Recall (3.9), we eventually find the optimal condition

$$\langle \tfrac{\partial J(u_0,u)}{\partial u_0}, h\rangle = \langle u^*(0), h\rangle + \langle J_{u_0}(u_0, u), h\rangle = 0,$$

Now the only task left is to solve equation (3.11) for $z$,

$$\begin{aligned}
\langle M_u^* z, \frac{\partial u}{\partial u_0} h\rangle &= \langle z, M_u \frac{\partial u}{\partial u_0} h\rangle = \langle \begin{pmatrix} (\frac{\partial}{\partial t} + A)\frac{\partial u}{\partial u_0} h \\ \frac{\partial u(0)}{\partial u_0} h \end{pmatrix}, \begin{pmatrix} u^* \\ u^*(0) \end{pmatrix}\rangle \\
&= \langle (\frac{\partial}{\partial t} + A)\frac{\partial u}{\partial u_0} h, u^*\rangle + \langle \frac{\partial u(0)}{\partial u_0} h, u^*(0)\rangle \\
&= \langle (-\frac{\partial}{\partial t} + A^*) u^*, \frac{\partial u}{\partial u_0} h\rangle + \langle \frac{\partial u(0)}{\partial u_0} h, u^*(0)\rangle + (u^*(T), \frac{\partial u(T)}{\partial u_0} h) \\
&\quad - (\frac{\partial u(0)}{\partial u_0} h, u^*(0))
\end{aligned}$$

$$= \langle (-\frac{\partial}{\partial t} + A^*)u^*, \frac{\partial u}{\partial u_0}h\rangle + (\frac{\partial u(T)}{\partial u_0}h, u^*(T))$$
$$= \langle J_u(u_0, u), \frac{\partial u}{\partial u_0}h)\rangle,$$

where the third line is taking integration by part with respect to $t$.

Setting $u^*(T) = 0$ we then have

$$-\frac{\partial u^*}{\partial t} + A^*u^* = J_u(u_0, u),$$
$$u^*(T) = 0.$$

This finally concludes the optimality system

$$\begin{cases} \frac{\partial u}{\partial t} + Au = f, \\ u(0) = u_0, \\ -\frac{\partial u^*}{\partial t} + A^*u^* = J_u(u_0, u), \\ u^*(T) = 0, \\ u^*(0) + J_{u_0}(u_0, u) = 0. \end{cases} \tag{3.13}$$

**Remark 4** *In the above derivation, (3.13) used the assumption* $H(\Omega) = L^2(\Omega)$*. If* $H(\Omega) \neq L^2(\Omega)$*, the optimal condition will be given implicitly by equation* $(u^*(0), h) + \langle J_{u_0}(u_0, u), h\rangle = 0$*. In the following parts of this Section, we will keep this assumption* $H(\Omega) = L^2(\Omega)$*.*

**3.2.2. Lagrange Multiplier Rule.** Another way, that is usually more convenient to derive the optimality system, is the Lagrange multiplier rule. Before using such method, one thing we should verify is the surjective of the derivative of the constraint operator $M$, which is essentially required by the implicit function theorem [51]. However, this is usually not an issue when operator $A$ in (3.2) is linear. In particular, we should be very careful for problems with nonlinear constraint.

To find the solution of the problem (3.1)-(3.2), we form a Lagrange functional:

$$\mathcal{L}(\lambda, u, u_0) = J(u_0, u) + \langle \lambda, M(u_0, u) \rangle, \tag{3.14}$$

where $\lambda \in L^2(0,T;V''(\Omega)) \times L^2(\Omega)'$ is the Lagrange multiplier.
By the isometrics property of Hilbert space, we identify $\lambda$ as $\begin{pmatrix} u^* \\ u(0)^* \end{pmatrix} \in L^2(0,T;V(\Omega)) \times L^2(\Omega)$ and rewrite (3.14) as

$$\begin{aligned}\mathcal{L}(u^*, u^*(0), u, u_0) &= J(u_0, u) + \langle M(u, u_0), \begin{pmatrix} u^* \\ u(0)^* \end{pmatrix} \rangle \\ &= J(u_0, u) + \langle \frac{\partial u}{\partial t} + Au - f, u^* \rangle + \langle u^*(0), u(0) - u_0 \rangle.\end{aligned}$$

Doing calculus of variation with respect to $u^*, u^*(0)$, we recover the constraint equation

$$\frac{\partial u}{\partial t} + Au = f, \quad u(0) = u_0. \tag{3.15}$$

Doing calculus of variation with respect to $u, u_0$, we have

$$\langle \frac{\partial v}{\partial t} + Av - f, u^* \rangle + \langle u^*(0), v(0) \rangle + \langle J_u(u_0, u), v \rangle = 0, \tag{3.16}$$

$$\langle u^*(0), -h \rangle + \langle J_{u_0}(u_0, u), h \rangle = 0. \tag{3.17}$$

Taking integration by part with respect to $t$ on (3.16), we find out

$$\langle -\frac{\partial u^*}{\partial t} + A^* u^*, v \rangle + \langle J_u(u_0, u), v \rangle + \langle u^*(T), v(T) \rangle = 0. \tag{3.18}$$

Setting $u(T) = 0$ we have

$$-\frac{\partial u^*}{\partial t} + A^* u^* = -J_u(u_0, u), \quad u^*(T) = 0. \tag{3.19}$$

Summarizing (3.15)-(3.19), we obtain the optimality system

$$\begin{cases} \frac{\partial u}{\partial t} + Au = f, \\ u(0) = u_0, \\ -\frac{\partial u^*}{\partial t} + A^* u^* = -J_u(u_0, u), \\ u^*(T) = 0, \\ u^*(0) - J_{u_0}(u_0, u) = 0, \end{cases} \tag{3.20}$$

which is the same as (3.13) derived in section 3.2.1.

The dual method and the Lagrange multiplier rule are both important techniques in optimization. However, the Lagrange multiplier rule is preferred in lots of constrained optimizations because it is more straightforward to be implemented and allows the optimization problems to be solved without explicitly parameterizing the constraints. In next subsection, we will employ the Lagrange multiplier rule to explore the discrete variational data assimilation (or discrete PDE-constrained optimization).

## 3.3. DISCRETE VARIATIONAL DATA ASSIMILATION

Usually, for computational purpose, we have multiple ways to deal with a variational data assimilation or a PDE-constrained optimization problem. One is to first optimize it and then discretize it, such as a direct numerical solving of the optimality system (3.20) in section 3.1. The second is to first discretize it and then optimize it, which is going to be discussed in this section.

We use a finite element method as a spatial discretization and the backward Euler scheme as an example to discretize the time. Let $\mathcal{T}_h$ denote a family of triangulation of $\Omega_h$ that is an approximation of $\Omega$. Assume the triangulation $\mathcal{T}_h$ satisfies the usual sort of quasi-uniformity condition. Associated with $\mathcal{T}_h$ is the finite element space $V_h$ =span$\{u_i\}_{i=1}^{i=N_b}$, where $u_i$ are piecewise polynomials and $N_b$ is the number of finite element nodes. The admissible set of discrete optimal variable is then considered as $U_h = V_h \cap U_{ad}$.

For the time discretization we uniformly construct a temporal grid $0 = t_0 < t_1 < t_2 < t_3 ... < t_n ... < t_N = T$ with time step $\tau = \frac{T}{N}$. Let $I_n = (t_{n-1}, t_n]$ denote the $n^{th}$ sub-interval. We will use the finite-dimensional space

$$V_{\tau,h} = \{v : [0,T] \to V_h : v|_{I_n} \in V_h \quad \text{is constant in time}\}.$$

Let $v^n$ be the value of $v \in V_{\tau,h}$ at $t_n$ and $V^n_{\tau,h}$ be the restriction to $I_n$ of the functions in $V_{\tau,h}$.

Given specific $h, \tau$ and $\gamma > 0$, a fully discretization of the problem (3.1)-(3.2) can be formulated as:

$$\min_{u_{0,h} \in U_h} J_h(u_{0,h}) \tag{3.21}$$

subject to

$$\begin{cases} \langle \dfrac{u_h^{n+1} - u_h^n}{\tau}, v_h \rangle + \langle A u_h^{n+1}, v_h \rangle = \langle f_{n+1}, v_h \rangle, \quad \forall v_h \in V_h, \\ u_h^0 = u_{0,h} \quad u_{0,h} \in L^2(\Omega), \end{cases} \tag{3.22}$$

where

$$F_h(u_{0,h}) = \frac{1}{2}\tau \sum_{n=1}^{N} \|\hat{u}^n - u_h^n\|^2_{\mathcal{H}} + \frac{\gamma}{2}\|u_{0,h}\|^2_H. \tag{3.23}$$

Using the similar argument in continuous formulation (3.1)-(3.2), it is not difficult to prove the wellposedness of problem (3.21)-(3.23).

In order to solve for $u_{0,h}$, we use Lagrange multiplier rule and form the discrete Lagrange functional:

$$\begin{aligned}\mathcal{L}(\bar{u}_h, u_{0,h}, \bar{u}_h^*) &= \frac{1}{2}\tau\sum_{n=1}^{N}\|\hat{u}^n - u_h^n\|_{\mathcal{H}}^2 + \frac{\gamma}{2}\|u_{0,h}\|_H^2 \\ &+ \tau\sum_{n=0}^{N-1}\langle\frac{u_h^{n+1} - u_h^n}{\tau} + Au_h^{n+1} - f_{n+1}, u_h^{*n}\rangle + (u_h^0 - u_{0,h}, u_h^{*0}),\end{aligned} \tag{3.24}$$

where $\bar{u}_h = (u_h^1, u_h^2, u_h^3, \ldots.., u_h^N)$ and $\bar{u}_h^* = (u_h^{*0}, u_h^{*1}, u_h^{*2}, u_h^{*3}, \ldots.., u_h^{*N-1})$. Based on a adjoint notation in the sense of $\langle Au, v\rangle = \langle A^*v, u\rangle$, we rewrite (3.24) as

$$\begin{aligned}\mathcal{L}(\bar{u}_h, u_{0,h}, \bar{u}_h^*) &= \frac{1}{2}\tau\sum_{n=1}^{N}\|\hat{u}^n - u_h^n\|_{\mathcal{H}}^2 + \frac{\gamma}{2}\|u_{0,h}\|_H^2 + \tau\sum_{n=0}^{N-1}(\langle\frac{u_h^{n+1} - u_h^n}{\tau}, u_h^{*n}\rangle \\ &+ \langle Au_h^{n+1}, u_h^{*n}\rangle - \langle f_{n+1}, u_h^{*n}\rangle) + (u_h^N, u_h^{*N}) - (u_h^N, u_h^{*N}) + (u_h^0 - u_{0,h}, u_h^{*0}) \\ &= \frac{1}{2}\tau\sum_{n=1}^{N}\|\hat{u}^n - u_h^n\|_{\mathcal{H}}^2 + \frac{\gamma}{2}\|u_{0,h}\|_H^2 + \tau\sum_{n=1}^{N}\langle\frac{u_h^{*n-1} - u_h^{*n}}{\tau}, u_h^n\rangle \\ &+ \tau\sum_{n=1}^{N}\langle A^*u_h^{*n-1}, u_h^n\rangle - \tau\sum_{n=1}^{N}\langle f_n, u_h^{*n-1}\rangle + (u_h^N, u_h^{*N}) - (u_{0,h}, u_h^{*0}).\end{aligned} \tag{3.25}$$

By using standard techniques from the calculus of variations, we derive equations that correspond to rendering (3.25) stationary. Variations in the Lagrange multiplier $\bar{u}_h^*$ recover the constraint equation (3.22). Variations with respect to $u_{0,h}$ and $u_h^n$, for $n = 1, 2, 3....N-1$ yield

$$\frac{\partial\mathcal{L}(\bar{u}_h, u_{0,h}, \bar{u}_h^*)}{\partial u_{0,h}}v_h = (\gamma u_{0,h}, v_h) - (u_h^{*0}, v_h) = 0, \tag{3.26}$$

$$\frac{\partial\mathcal{L}(\bar{u}_h, u_{0,h}, \bar{u}_h^*)}{\partial u_h^n}w_h = \tau\langle\frac{u_h^{*n-1} - u_h^{*n}}{\tau}, w_h\rangle + \tau\langle A^*u_h^{*n-1}, w_h\rangle - \tau\langle\hat{u}^n - u_h^n, w_h\rangle = 0. \tag{3.27}$$

Imposing $u_h^{*N} = 0$ when calculating the variation with respect to $u_h^N$ finally gives the discrete optimality system,

$$\begin{cases} \dfrac{u_h^{n+1} - u_h^n}{\tau} + Au_h^{n+1} = f_{n+1}, \\ u_h^0 = u_{0,h}, \\ -\dfrac{u_h^{*n+1} - u_h^{*n}}{\tau} + A^* u_h^{*n} = \hat{u}^{n+1} - u_h^{n+1}, \\ u_h^{*N} = 0, \\ u_{0,h} = \dfrac{1}{\gamma} u_h^{*0} \end{cases} \tag{3.28}$$

for $n = 0, 1, 2, 3.....N - 1$.

**Remark 5** *One may observe that the discrete optimality system* (3.28) *is the same as the direct full discretization of* (3.20). *This is because of the special symmetric property of Euler's scheme, including the explicit Euler scheme. However, such coincidence may not happen for other temporal discretization schemes, such as the Crank-Nicolson and most of the Runge-Kutta methods.*

A direct solving of the optimality systems (3.20) and (3.28) will couple space and time together and result in an extreme large linear system, which is challenging to the computational resource. In next section, we present several iterative algorithms to overcome this difficulty.

## 3.4. GRADIENT DESCENT METHOD

When talking about descent method to find optimizer for a smooth cost functional, there are two basics that matter: the descent direction and the descent step size along that direction. In this section, we recall two methods with descent directions that rely on the information from the first order derivative (or the gradient), thereby being named as gradient descent method. Based on different descent directions, we further classify them as the conjugate gradient method and the steepest descent method. We also will discuss practical techniques to determine a desirable choice of the descent step size.

To begin with, we need to calculate the first order derivative of the cost functional (3.21) and find out its dual element (or gradient) in the admissible set:

$$\langle J_h'(u_{0,h}), v_h\rangle = \tau\sum_{n=1}^{N}\langle \hat{u}^n - u_h^n, (u_h^n)'v_h\rangle + (\gamma u_{0,h}, v_h), \quad \forall v_h \in U_h. \tag{3.29}$$

Note that $(u_h^n)'v_h$ is essentially the solution by solving the following discretized equation

$$\begin{cases} \langle \dfrac{\mathcal{U}_h^{n+1} - \mathcal{U}_h^n}{\tau}, w_h\rangle + \langle A\mathcal{U}_h^{n+1}, w_h\rangle = 0, \\ \mathcal{U}_h^0 = v_h, \end{cases} \tag{3.30}$$

for $n = 0, 1, 2, 3, \cdot, N-1$.

We now introduce a set of adjoint variables $\{u_h^{*n}\}_{n=0}^{N-1}$ to replace each $w_h$, for $n = 0, 1, 2, 3, \cdot, N-1$, in (3.30):

$$\begin{cases} \langle \dfrac{\mathcal{U}_h^{n+1} - \mathcal{U}_h^n}{\tau}, u_h^{*n}\rangle + \langle A\mathcal{U}_h^{n+1}, u_h^{*n}\rangle = 0, \\ (\mathcal{U}_h^0 - v_h, u_h^{*0}) = 0. \end{cases} \tag{3.31}$$

Multiplying the first equation with $\tau$ in (3.31) and adding all equations together, for $n = 0, 1, 2, 3, \cdot, N-1$, we do the following manipulations

$$\begin{aligned} 0 &= \tau\sum_{n=0}^{N-1}(\langle \frac{\mathcal{U}_h^{n+1} - \mathcal{U}_h^n}{\tau}, u_h^{*n}\rangle + \langle A\mathcal{U}_h^{n+1}, u_h^{*n}\rangle) + (\mathcal{U}_h^0 - v_h, u_h^{*0}) \\ &= \tau\sum_{n=0}^{N-1}(\langle \frac{\mathcal{U}_h^{n+1} - \mathcal{U}_h^n}{\tau}, u_h^{*n}\rangle + \langle A\mathcal{U}_h^{n+1}, u_h^{*n}\rangle) \end{aligned} \tag{3.32}$$

$$
\begin{aligned}
&+ (\mathcal{U}_h^N, u_h^{*N}) - (\mathcal{U}_h^N, u_h^{*N}) + (\mathcal{U}_h^0 - v_h, u_h^{*0}) \\
&= \tau \sum_{n=1}^{N} \langle \frac{u_h^{*n-1} - u_h^{*n}}{\tau}, \mathcal{U}_h^n \rangle + \tau \sum_{n=1}^{N} \langle A u_h^{*n-1}, \mathcal{U}_h^n \rangle \\
&+ (\mathcal{U}_h^N, u_h^{*N}) - (v_h, u_h^{*0}).
\end{aligned} \tag{3.33}
$$

We can connect (3.29) and (3.33) by setting the following equations:

$$
\begin{aligned}
&\langle \frac{u_h^{*n-1} - u_h^{*n}}{\tau}, \mathcal{U}_h^n \rangle + \langle A u_h^{*n-1}, \mathcal{U}_h^n \rangle = \langle \hat{u} - u, \mathcal{U}_h^n \rangle, \\
&u_h^{*N} = 0,
\end{aligned} \tag{3.34}
$$

which leads to

$$
\begin{aligned}
\langle J_h'(u_{0,h}), v_h \rangle &= \tau \sum_{n=1}^{N} \langle \hat{u}^n - u_h^n, (u_h^n)' v_h \rangle + (\gamma u_{0,h}, v_h) \\
&= -(v_h, u_h^{*0}) + (\gamma u_{0,h}, v_h) \\
&= (\gamma u_{0,h} - u_h^{*0}, v_h).
\end{aligned} \tag{3.35}
$$

Finally, $\gamma u_{0,h} - u_h^{*0}$ gives the gradient of (3.21) at current initial condition $u_{0,h}$. Basically, the above is a similar argument compared with (3.24) − (3.25), and $\gamma u_{0,h} - u_h^{*0}$ is the representation of the linear functional $J_h'(u_{0,h})$ in the admissible set $U_h$.

**3.4.1. Steepest Descent Method.** With the result in (3.35), we can discuss the steepest descent method [52, 53] to solve the discrete optimization problem (3.21): given $\vec{u}_{0,h}^{(0)}$ and a tolerance $\epsilon$, solve the following equations sequentially until the stop criteria $\|\vec{u}_{0,h}^{(i+1)} - \vec{u}_{0,h}^{(i)}\|_H \leq \epsilon$ (or $\|\gamma u_{0,h}^{(i+1)} - u_h^{*0(i+1)}\|_H \leq \epsilon$) is satisfied:

$$
\begin{cases}
\dfrac{u_h^{n+1(i)} - u_h^{n(i)}}{\tau} + A u_h^{n+1(i)} = f^{n+1}, \\
u_h^{0(i)} = u_{0,h}^{(i)},
\end{cases} \tag{3.36}
$$

$$\begin{cases} -\dfrac{u_h^{*n+1(i)} - u_h^{*n(i)}}{\tau} + A^* u_h^{*n(i)} = \hat{u}^{n+1} - u^{n+1(i)}, \\ u_h^{*N(i)} = 0, \end{cases} \tag{3.37}$$

$$u_{0,h}^{(i+1)} = u_{0,h}^{(i)} + \eta^{i+1}(u_h^{*0(i)} - \gamma u_{0,h}^{(i)}), \tag{3.38}$$

where $n = 0, 1, 2, 3....N$ is time evolution step, $i = 0, 1, 2, 3....$ represents the iteration step, $\eta^{i+1}$ is called the learning rate at each iteration, and $u_{0,h}^{(i)}$, $u_h^{n(i)}$, $u_h^{*n(i)}$ are iterative sequences. Since gradient is a local information for a given functional, a choice of $\eta^{i+1}$ usually happens between $(0, 1]$.

We illustrate the steepest descent algorithm as follows:

**Algorithm 1** *Step 0 (Initialization): Specify a convergence tolerance $\epsilon$, guess initial function $\vec{u}_{0,h}^{(0)}$, and start the iteration step $i = 1$.*

*Step 1 (Forward phase): Use $u_{0,h}^{(i)}$ as initial condition to solve equation (3.36) for $u_h^{(i)}$.*

*Step 2 (Backward phase): Pass $u_h^{(i)}$ to equation (3.37) and solve equation (3.37) backward for $u_h^{*0(i)}$.*

*Step 3 (Update phase): Use $\eta^{i+1}$ from $(0, 1)$ and then update*

$$u_{0,h}^{(i+1)} = u_{0,h}^{(i)} + \eta^{i+1}(u_h^{*0(i)} - \gamma u_{0,h}^{(i)}).$$

*Step 4 (Criteria for stopping the iteration): Compute $\|u_h^{*0(i)} - \gamma u_{0,h}^{(i)}\|$, if $\|u_h^{*0(i)} - \gamma u_{0,h}^{(i)}\| \leq \epsilon$ then stop; otherwise, increase i by 1 and go back to Step 1.*

**3.4.2. Inexact Line Search Steepest Descent Method.** To reduce the iterations and improve computational efficiency in the steepest descent method, we can optimize the choice of the learning rate $\eta^{i+1}$ instead of randomly picking $\eta^{i+1}$ between $(0, 1]$ at each iteration. This is equivalent to deal with another minimization problem:

$$\min_{\eta^{i+1}\in\mathbb{R}} u(\eta^{i+1}) = J_h(u_{0,h}^{(i)} + \eta^{i+1}(u_h^{*0(i)} - \gamma u_{0,h}^{(i)})). \tag{3.39}$$

Normally, solving problem (3.39) is also computationally expensive and tractable, which is what we do not want to handle. An alternative way is to determine a relatively optimal $\eta^{i+1}$ by Armijo backtracking rule. The idea is given as follows: we start a guess of relatively large $\eta^{i+1}$, such as a real number 1 or 2, then shrink $\eta^{i+1}$ proportional to a constant $\rho$ between $(0, 1)$ if $u_{0,h}^{(i)} + \eta^{i+1}(u_h^{*0(i)} - \gamma u_{0,h}^{(i)})$ does not provide a noticeable value decreasing for the cost functional (3.21).

Such idea is also called the inexact line search method, which is mathematically described as: find $\eta^{i+1}$ via repeatedly solving (3.36) with initial value

$$u_{0,h}^{(i+1)} = u_{0,h}^{(i)} + \eta^{i+1}(\vec{u}_h^{*0(i)} - \gamma u_{0,h}^{(i)}) \quad \text{by updating} \quad \eta^{i+1} = \rho\eta^{i+1}, \tag{3.40}$$

until the following inequality is satisfied

$$J_h(u_{0,h}^{(i+1)}) \le J_h(u_{0,h}^{(i)}) + \delta\eta^{i+1}\langle J_h'(u_{0,h}^{(i)}), u_h^{*0(i)} - \gamma u_{0,h}^{(i)}\rangle, \tag{3.41}$$

where $\eta^{i+1}$ is typically initialized as a constant equal to or greater than 1, and $\delta$ and $\rho$ are chosen between $(0, 1)$.

The straight line

$$y(\eta^{i+1}) = J_h(u_{0,h}^{(i)}) + \delta\eta^{i+1}\langle J_h'(u_{0,h}^{(i)}), u_h^{*0(i)} - \gamma u_{0,h}^{(i)}\rangle \tag{3.42}$$

is a search line, which measures the value decreasing of the cost functional (3.21). If $J_h(u_{0,h}^{(i)} + \eta^{i+1}(u_h^{*0(i)} - \gamma u_{0,h}^{(i)}))$ is underneath the line (3.42), $\eta^{i+1}$ is a good candidate of descent step size, otherwise, we need to go through step (3.40) again until the inequality (3.41) is satisfied.

After the modification of the choice on $\eta^{i+1}$, we summarize the algorithm as follows:

**Algorithm 2** *Step 0 (Initialization): Specify a convergence tolerance $\epsilon$, guess initial function $\vec{u}_{0,h}^{(0)}$, and start the iteration step $i = 1$.*

*Step 1 (Forward phase): Use $u_{0,h}^{(i)}$ as initial condition to solve equation (3.36) for $u_h^{(i)}$.*

*Step 2 (Backward phase): Pass $u_h^{(i)}$ to equation (3.37) and solve equation (3.37) backward for $u_h^{*0(i)}$.*

*Step 3 (Inexact line search for $\eta^{i+1}$):*

*(1) Initialize a constant $\eta^{i+1} \geq 1$, set $0 < \rho < 1$ and $0 < \delta < 1$;*

*(2) use $u_{0,h}^{(i+1)} = u_{0,h}^{(i)} + \eta^{i+1}(u_h^{*0(i)} - \gamma u_{0,h}^{(i)})$ as initial value to solve equation (3.36) forward to obtain $u_h^n$ for computing $J_h(\vec{u}_{0,h}^{(i+1)})$;*

*(3) Update $\eta^{i+1} = \rho\eta^{i+1}$ until inequality (3.41) is attained.*

*(4) Output $\eta^{i+1}$.*

*Step 4 (Update phase): Use $\eta^{i+1}$ from step 3 and then update*

$$u_{0,h}^{(i+1)} = \vec{u}_{0,h}^{(i)} + \eta^{i+1}(u_h^{*0(i)} - \gamma u_{0,h}^{(i)}).$$

*Step 5 (Criteria for stopping the iteration): Compute $\|u_h^{*0(i)} - \gamma u_{0,h}^{(i)}\|$, if $\|u_h^{*0(i)} - \gamma u_{0,h}^{(i)}\| \leq \epsilon$ then stop; otherwise, increase $i$ by 1 and go back to Step 1.*

**Remark 6** *The steepest descent algorithm or descent with inexact line search is very stable and easy to be implemented by only calculating the first order derivative. However, we know the information contained in the first order derivative is a consequence of locally linear approximation of the cost functional. Therefore, slow convergence behavior, such as linear or sublinear rate, is often observed.*

**3.4.3. Conjugate Gradient Method.** To possess a faster convergence speed, besides using the gradient information, we also can try to incorporate the information from previous steps, that is interpreted as a momentum or inertial term, to accelerate the algorithm. Mathematically, this change essentially allows us to vary the steepest descent direction such that the new one could be $\mathcal{A}$-conjugate orthogonal to all previous descent directions, where $\mathcal{A}$ is connected to the property of the cost functional. This is then called the conjugate gradient method.

For detail, the conjugate gradient method is illustrated as: given $u_{0,h}^{(0)}$, $u_{0,h}^{(1)}$ and a tolerance $\epsilon$, solve the following equations sequentially until the stop criteria $\|\vec{u}_{0,h}^{(i+1)} - \vec{u}_{0,h}^{(i)}\|_H \leq \epsilon$ (or $\|\gamma u_{0,h}^{(i+1)} - u_h^{*0(i+1)}\|_H \leq \epsilon$) is satisfied:

$$\begin{cases} \dfrac{u_h^{n+1(i)} - u_h^{n(i)}}{\tau} + Au_h^{n+1(i)} = f^{n+1}, \\ u_h^{0(i)} = u_{0,h}^{(i)}, \end{cases} \tag{3.43}$$

$$\begin{cases} -\dfrac{u_h^{*n+1(i)} - u_h^{*n(i)}}{\tau} + A^* u_h^{*n(i)} = \hat{u}^{n+1} - u^{n+1(i)}, \\ u_h^{*N(i)} = 0, \end{cases} \tag{3.44}$$

$$u_{0,h}^{(i+1)} = u_{0,h}^{(i)} + \zeta^{i+1} B^i (u_h^{*0(i)} - \gamma u_{0,h}^{(i)}) + \eta^{i+1} C^i (u_{0,h}^{(i)} - u_{0,h}^{(i-1)}), \tag{3.45}$$

where $n = 0, 1, 2, 3, ..., N$ is the time evolution step, $i = 0, 1, 2, 3, ...$ represents the iteration step, $\zeta^{i+1}$ and $\eta^{i+1}$ are iterative parameters, $u_{0,h}^{(i)}$, $u_h^{n(i)}$, and $u_h^{*n(i)}$ are iterative sequences, and $B^i$ and $C^i$ are two symmetric positive definite matrices.

Following the ideas in [54, 55] we adopt $B^i$ and $C^i$ as identity matrices, $\zeta^{i+1}$ and $\eta^{i+1}$ are updated using

$$\zeta^{i+1} = \frac{1}{q^{i+1}}, \quad \eta^{i+1} = \frac{e^i}{q^{i+1}}, \tag{3.46}$$

where

$$e^i = \begin{cases} 0 & i = 0, \\ q^i \frac{\|\lambda^i\|_H^2}{\|\lambda^{i-1}\|_H^2} & i > 0, \end{cases}$$

$$q^{i+1} = \frac{\|\lambda^i\|_L^2}{\|\lambda^i\|_H^2} - e^i, \quad i = 0, 1, 2, 3, \ldots.$$

Here $\lambda^i = \gamma u_{0,h}^{(i)} - u_h^{*0(i)}$ and $\|\lambda^i\|_L = (L\lambda^i, \lambda^i)^{\frac{1}{2}}$. The operator $L$ acting on $\lambda^i$ is defined as follows

$$\begin{cases} \dfrac{\vec{\vartheta}_h^{n+1} - \vartheta_h^n}{\tau} + A\vartheta_h^{n+1} = 0, \\ \vartheta_h^0 = \lambda^i, \end{cases} \tag{3.47}$$

$$\begin{cases} -\dfrac{\vec{\vartheta}_h^{*n+1} - \vartheta_h^{*n}}{\tau} + A^*\vartheta_h^{*n} = -{\vartheta_h}^{n+1}, \\ \vartheta_h^{*N} = \vec{0}, \end{cases} \tag{3.48}$$

$$L\lambda^i = \gamma\lambda^i - \vartheta_h^{*0}. \tag{3.49}$$

Now the conjugate gradient algorithm can be summarized as follows:

**Algorithm 3** *Step 0 (Initialization): Specify a convergence tolerance $\epsilon$, guess two initial functions $u_{0,h}^{(0)}$ and $u_{0,h}^{(1)}$, and then start the iteration at step $i = 1$.*

*Step 1 (Forward phase): Use $u_{0,h}^{(i)}$ as the initial condition to solve (3.43) for $u_h^{(i)}$.*

*Step 2 (Backward phase): Pass $u_h^{(i)}$ to (3.44) and solve (3.44) backwards for $u_h^{*0(i)}$.*

*Step 3 (Computing for operator L):*

*(1) Set* $\lambda^i = \gamma u_{0,h}^{(i)} - u_h^{*0(i)}$ *and use it as initial value to solve equation (3.47) forward to obtain* $\vartheta_h$*;*

*(2) Pass* $\vartheta_h$ *to (3.48) and solve equation (3.48) backward for attaining* $\vartheta_h^{*0}$*;*

*(3) Compute* $L\lambda^i = \gamma\lambda^i - \vartheta_h^{*0}$.

*Step 4 (Update phase): Calculate* $\zeta^{i+1}, \eta^{i+1}$ *by using* (3.46) *and then update*

$$u_{0,h}^{(i+1)} = \vec{u}_{0,h}^{(i)} + \zeta^{i+1}(u_h^{*0(i)} - \gamma u_{0,h}^{(i)}) + \eta^{i+1}(u^{(i)}{}_{0,h} - u^{(i-1)}{}_{0,h}).$$

*Step 5 (Criteria for stopping the iteration): Compute* $\|u_{0,h}^{(i+1)} - u_{0,h}^{(i)}\|_H$. *If* $\|u_{0,h}^{(i+1)} - u_{0,h}^{(i)}\|_H \leq \epsilon$ *then stop. Otherwise increase i by* 1 *and go back to Step 1.*

**Remark 7** *The conjugate gradient method serves a linear or super linear convergence rate, and solves the discrete optimality system* (3.28) *effectively in most of cases. However, it is relatively less stable and hence causes the algorithm itself to diverge for some of the data assimilation scenarios that have low stability, e.g., small regularization parameter* $\gamma$ *in the cost functional* (3.21). *To tackle this numerical problem, we usually turn back to the steepest descent method which gains more stability at the cost of a lower convergence rate.*

## 3.5. PROPER ORTHOGONAL DECOMPOSITION IN OPTIMIZATION

Proper orthogonal decomposition (POD) is a data approximation method that aims at obtaining a low-dimensional representation of the high-dimensional processes. It does so by creating an optimal lower order basis, called POD modes, to minimize the information loss as less as possible. POD has numerous applications in fluid dynamics, image process, signal analysis, and data compression.

In this section, we employ an incremental POD method [56, 57] to optimize the computational resource in the variational data assimilation or PDE constrained optimization problem. Recall in section 3.4 the steepest descent method:

$$\begin{cases} \dfrac{u_h^{n+1(i)} - u_h^{n(i)}}{\tau} + Au_h^{n+1(i)} = f^{n+1}, \\ u_h^{0(i)} = u_{0,h}^{(i)}, \end{cases} \tag{3.50}$$

$$\begin{cases} -\dfrac{u_h^{*n+1(i)} - u_h^{*n(i)}}{\tau} + A^* u_h^{*n(i)} = \hat{u}^{n+1} - u^{n+1(i)}, \\ u_h^{*N(i)} = 0, \end{cases} \tag{3.51}$$

$$u_{0,h}^{(i+1)} = u_{0,h}^{(i)} + \eta^{i+1}(u_h^{*0(i)} - \gamma u_{0,h}^{(i)}). \tag{3.52}$$

To find the gradient at current point $u_{0,h}^{(i)}$, we need to solve equation (3.50) for obtaining $\{u_h^n\}_{n=1}^N$, then plug $\{u_h^n\}_{n=1}^N$ into the right side of equation (3.51) for a backward solving to find $u_h^{*0(i)}$. During this procedure, if the space dimension is considerable and the time evolution $T$ is long as well, we may run into trouble storing the data $\{u_h^n\}_{n=1}^N$ when solving (3.50). To address this issue, an incremental POD method is introduced as a data compression technique such that the storage of $\{u_h^n\}_{n=1}^N$ is not challenging to the computer memory anymore.

**3.5.1. Proper Orthogonal Decomposition/Singular Value Decomposition.** We first review the basic ideas of the data compression using proper orthogonal decomposition. Given a set of data $\{\gamma_k\}_{k=1}^m \in X$, find a set of proper orthonormal basis $\{\gamma_k\}_{k=1}^r$, $r << n$ and the corresponding coefficients such that the following objective functional is minimized:

$$\min_{\gamma^i \in X, i=1,2\ldots r} J_r = \sum_{k=1}^m \|\gamma_k - \gamma_k^r\|_X^2, \tag{3.53}$$

where $\|\cdot\|_X$ is a norm measuring distance we are interested in, and $\gamma_k^r = \sum_{j=1}^{r} \beta_{jk}\gamma_j$. To solve this minimization problem, we use the best operator approximation theory. To begin with, we define the Hilbert-Schmidt (HS) norm for a bounded linear operator.

**Definition 15** *Let $X$ and $Y$ be Hilbert spaces, an operator $T \in \mathscr{L}(X,Y)$ is Hilbert-Schmidt if $\sum_{n\geq 1} \|Tx_n\|^2 < \infty$ for some total orthonormal sequence $\{x_n\}_{n\geq 1} \in X$. The Hilbert-Schmidt norm of $T$ is defined as*

$$\|T\|_{HS} = (\sum_{n\geq 1} \|Tx_n\|^2)^{\frac{1}{2}}. \tag{3.54}$$

Define a linear bounded operator $L : \mathbb{K}^m \to X$ by $La = \sum_{k=1}^{m} a_k\gamma_k$, $a = \{a_k\}_{k=1}^{m} \in \mathbb{K}^m$.

**Lemma 1** *$L$ is Hilbert-Schmidt (HS) and $\|L\|_{HS}^2 = \sum_{k=1}^{m} \|\gamma_k\|_X^2$.*

**Lemma 2** *If $T \in \mathscr{L}(X,Y)$ is HS, then the best r-rank HS operator approximation of $T$ is given as:*

$$\min_{K_r \in B(X,Y)} \|T - K_r\|_{HS} = (\sum_{j\geq r+1} \sigma_j^2)^{\frac{1}{2}}, \tag{3.55}$$

*in which the minimum is achieved by the $r^{th}$ truncated singular value decomposition (SVD) $T_r$ of $T$, and $\sigma_j$ represents the $j^{th}$ singular value of $T$.*

Based on the definition of operator $L$, Lemma 1, and Lemma 2, we have

$$\min_{\gamma^i \in X, i=1,2\ldots r} J_r = \sum_{k=1}^{m} \|\gamma_k - \gamma_k^r\|_X^2 = \min_{K_r \in B(\mathbb{K}^m, X)} \|L - K_r\|_{HS}^2, \tag{3.56}$$

which gives the following Theorem:

**Theorem 10** *Let* $\{\gamma_k\}_{k=1}^n \in X$ *and* $L \in \mathscr{L}(\mathbb{K}^m, X)$ *as defined above, if the SVD of L is given as*

$$La = \sum_{k=1}^{m} \sigma_k (a, \psi_k)_{\mathbb{K}^m} \theta_k, \tag{3.57}$$

*where* $\sigma_k$ *are ordered singular values,* $\psi_k, \theta_k$ *are the corresponding singular vectors of* $T^*T$ *and* $TT^*$*. Then* $\gamma_k^r = \sum_{j=1}^r (\gamma_k, \theta_j)_X \theta_j$ *solves the minimization problem*

$$\min_{\gamma^i \in X, i=1,2\ldots r} J_r = \sum_{k=1}^{n} \|\gamma_k - \gamma_k^r\|_X^2. \tag{3.58}$$

**3.5.2. Incremental SVD in Standard and Weighted Euclidean Space.** In standard Euclidean space, a given linear operator $L : \mathbb{R}^m \to \mathbb{R}^n$ has a matrix representation, saying $U_{m\times n}$. The SVD of $L$ is equivalent to find a matrix decomposition of $U_{m\times n}$ such that

$$U_{m\times n} = V\Sigma W^T \tag{3.59}$$

where $V, W$ are left and right singular vector matrix that are orthonormal with respect to standard Euclidean inner product, and $\Sigma$ is an ordered diagonal matrix. In the following, we use $\|\cdot\|$ to stand for the norm of standard Euclidean space. If we add one more column $c$ onto $U$, i.e., $\begin{bmatrix} U & c \end{bmatrix}$, which becomes a matrix representation of a new operator $L^a : \mathbb{R}^{m+1} \to \mathbb{R}^n$. To find out the SVD of $L^a$ or $\begin{bmatrix} U & c \end{bmatrix}$, we first use a projection and normalization or QR decomposition to produce a left orthonormal matrix:

$$VV^Tc \Longrightarrow c - VV^Tc \Longrightarrow \frac{c - VV^Tc}{\|c - VV^Tc\|} \Longrightarrow \begin{bmatrix} V & \frac{(c-VV^Tc)}{\|c-VV^Tc\|} \end{bmatrix} \Longrightarrow \text{QR decomposition of } \begin{bmatrix} U & c \end{bmatrix}$$

$$\Longrightarrow \begin{bmatrix} V\Sigma W^T & c \end{bmatrix} = \begin{bmatrix} V & \frac{(c-VV^Tc)}{\|c-VV^Tc\|} \end{bmatrix} \begin{bmatrix} \Sigma W^T & V^Tc \\ 0 & \|c - VV^Tc\| \end{bmatrix}$$

We also ask for a right orthonormal matrix, hence a nature candidate of matrix decomposition for $\begin{bmatrix} U & c \end{bmatrix}$ is, [57]

$$\begin{bmatrix} U & c \end{bmatrix} = \begin{bmatrix} V\Sigma W^T & c \end{bmatrix} = \begin{bmatrix} V & \frac{(c-VV^Tc)}{\|c-VV^Tc\|} \end{bmatrix} \begin{bmatrix} \Sigma W^T & V^Tc \\ 0 & \|c-VV^Tc\| \end{bmatrix}$$
$$= \begin{bmatrix} V & \frac{(c-VV^Tc)}{\|c-VV^Tc\|} \end{bmatrix} \begin{bmatrix} \Sigma & V^Tc \\ 0 & \|c-VV^Tc\| \end{bmatrix} \begin{bmatrix} W^T & 0 \\ 0 & 1 \end{bmatrix}.$$

Let $Q = \begin{bmatrix} \Sigma & V^Tc \\ 0 & \|c-VV^Tc\| \end{bmatrix}$ and its standard SVD as $Q = V_Q \Sigma_Q W_Q^T$, this leads to

$$\begin{bmatrix} U & c \end{bmatrix} = \begin{bmatrix} V & \frac{(c-VV^Tc)}{\|c-VV^Tc\|} \end{bmatrix} V_Q \Sigma_Q W_Q^T \begin{bmatrix} W^T & 0 \\ 0 & 1 \end{bmatrix}$$
$$= \left( \begin{bmatrix} V & \frac{(c-VV^Tc)}{\|c-VV^Tc\|} \end{bmatrix} V_Q \right) \Sigma_Q \left( \begin{bmatrix} W & 0 \\ 0 & 1 \end{bmatrix} W_Q \right)^T.$$

We can easily verify that the matrix $\begin{bmatrix} V & \frac{(c-VV^Tc)}{\|c-VV^Tc\|} \end{bmatrix} V_Q$ and $\begin{bmatrix} W & 0 \\ 0 & 1 \end{bmatrix} W_Q$ are still orthonormal matrix based on a simple geometric fact, and $\Sigma_Q$ is an ordered diagonal matrix automatically since it is the diagonal matrix from the SVD of $Q$. So far, using the above matrix decomposition, we are very close to state an iterative process to find the SVD for a large-scale data matrix incrementally. Considering the singularity and instability in this scheme, we also need to truncate the negligible singular values and their corresponding information in singular vector matrix, and give additional study to the following cases as well.

If $\|c - VV^T c\|$ is small enough, the column vectors of the matrix $V$ and $c$ can be viewed as being linear dependent, i.e., $c - VV^T c = 0$. In this point, we have the matrix $Q$ as $\begin{bmatrix} \Sigma & V^T c \\ 0 & 0 \end{bmatrix}$ and do the following matrix decomposition:

$$\begin{aligned}
\begin{bmatrix} U & c \end{bmatrix} = \begin{bmatrix} V\Sigma W^T & c \end{bmatrix} &= \begin{bmatrix} V & 0 \end{bmatrix} \begin{bmatrix} \Sigma W^T & V^T c \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} V & 0 \end{bmatrix} \begin{bmatrix} \Sigma & V^T c \\ 0 & 0 \end{bmatrix} \begin{bmatrix} W^T & 0 \\ 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} V & 0 \end{bmatrix} V_Q \Sigma_Q W_Q^T \begin{bmatrix} W^T & 0 \\ 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} V & 0 \end{bmatrix} V_Q \begin{bmatrix} \Sigma_{Q(1:k,1:k)} & 0 \\ 0 & 0 \end{bmatrix} W_Q^T \begin{bmatrix} W^T & 0 \\ 0 & 1 \end{bmatrix} \\
&= VV_{Q(1:k;1:k)} \Sigma_{Q(1:k;1:k)} (W_{Q(:,1:k)})^T \begin{bmatrix} W^T & 0 \\ 0 & 1 \end{bmatrix} \\
&== VV_{Q(1:k;1:k)} \Sigma_{Q(1:k;1:k)} \left( \begin{bmatrix} W & 0 \\ 0 & 1 \end{bmatrix} W_{Q(:,1:k)} \right)^T .
\end{aligned}$$

We can observe that $V_{Q(1:k;1:k)} \Sigma_{Q(1:k;1:k)} (W_{Q(:,1:k)})^T$ is exactly the standard SVD of the matrix $R = \begin{bmatrix} \Sigma & V^T c \end{bmatrix}$.

We now summarize the above incremental process as follows:

**Algorithm 4** *Step 0: Specified tolerances $\epsilon_1$, $\epsilon_2$, and $\epsilon_3$, initialize the SVD for the first data vector.*

*Step 1: Given $U = V\Sigma W^T$ and new data vector $c$, compute $\frac{c - VV^T c}{\|c - VV^T c\|}$ and assemble* $Q = \begin{bmatrix} \Sigma & V^T c \\ 0 & \|c - VV^T c\| \end{bmatrix}$.

*Step 2:*

*(1) If* $\|c - VV^T c\| \leq \epsilon_3$*, compute the SVD of* $R = \begin{bmatrix} \Sigma & V^T c \end{bmatrix}$ *as* $R = V_R \Sigma_R W_R^T$*, then update* $V = VV_R, \Sigma = \Sigma_R$*, and* $W^T = (WW_R)^T$*.*

*(2) If* $\|c - VV^T c\| \geq \epsilon_3$*, compute the SVD of* $Q = \begin{bmatrix} \Sigma & V^T c \\ 0 & \|c - VV^T c\| \end{bmatrix}$ *as* $Q = V_Q \Sigma_Q W_Q^T$*, then update* $V = \begin{bmatrix} V & \frac{(c - VV^T c)}{\|c - VV^T c\|} \end{bmatrix} V_Q$*,* $\Sigma = \Sigma_Q$ *and* $W^T = \left( \begin{bmatrix} W & 0 \\ 0 & 1 \end{bmatrix} W_Q \right)^T$*.*

*Step 3: If* $(V(:, end), V(:, 1)) > \epsilon_2$*, do the Gram-Schmidt orthogonalization.*

*Step 4: Do the truncation for* $V, \Sigma$ *and* $W$ *if the singular value is less than the specified tolerance* $\epsilon_1$*.*

*Step 5: Do Step 1-5 until the last data vector is introduced.*

**Remark 8** *Step3 is important to be checked to eliminate the non-orthogonality of left singular vectors caused by roundoff error.*

Besides the incremental POD for a standard Euclidean space, we are also interested in the best data approximation quantified by other different metrics. In the following, we will show the incremental POD in a general finite dimension Hilbert space or, equivalently, a $M$ weighted Euclidean space.

Assume $M$ is a positive definite matrix, define the Hilbert space $\mathbb{R}^m_M$ induced by the weighted inner product:

$$(x, y)_M = y' M x \quad \forall x, y \in \mathbb{R}^n_M. \tag{3.60}$$

We now consider a slightly different minimization problem:

$$\min_{\gamma^i \in \mathbb{R}^n_M, i=1,2\ldots r} J_r = \sum_{k=1}^{m} \|\gamma_k - \gamma_k^r\|^2_{\mathbb{R}^n_M}, \tag{3.61}$$

where $\{\gamma_k\}_{k=1}^m$ is a set of data vectors and $\gamma_k^r = \sum_{j=1}^r \beta_{jk}\gamma_j$. Denote the data matrix $Z = \{\gamma_1, \gamma_2, \gamma_3.....\gamma_m\}$ and define the operator $L : \mathbb{R}^m \to \mathbb{R}_M^n$:

$$La = \sum_{k=1}^{m} a_k \gamma_k = Za \quad \forall a = \{a_i\}_{i=1}^m \in \mathbb{R}^m. \tag{3.62}$$

We can find out the adjoint operator of $L$ by

$$(Lx, y)_{\mathbb{R}_M^n} = y' MZx = (x, L^* y)_{\mathbb{R}^m} = (L^* y)' x \quad x \in \mathbb{R}^m, \;\; y \in \mathbb{R}_M^n, \tag{3.63}$$

which leads to $L^* y = Z' M y$.

Based on the Theorem 10, we know the problem (3.61) eventually boils down to a SVD of the compact operator $L$. Before looking into the incremental decomposition of $L$, we clarify the following notations for understanding:

- $(x, y)_X = x^T M y$, $X$ is $M$-weighted $\mathbb{R}^n$ space.
- $Z^* = Z^T M$.
- $L : \mathbb{R}^m \to X, \quad L^* : X \to \mathbb{R}^m$.
- $L^* L : X \to X, \quad LL^* : \mathbb{R}^m \to \mathbb{R}^m$.

If we have a SVD of $L$: $Z = (\gamma_1, \gamma_2, ...\gamma_n) = V_M \Sigma_M W_M^T$, $Z$ is the matrix representation of $L$, then the above clarification implies that $V_M$ should be a $M$-orthonormal matrix and $W_M$ is a orthonormal matrix in standard sense.

A simple geometric fact is that the multiplication of a matrix with a standard orthonormal matrix is only a rotation without changing its shape, hence a $M$-orthonormal matrix are still $M$-orthonormal after transformed by a standard orthonormal matrix. This hints us only to modify the incremental SVD in the standard inner product space a little to obtain the incremental SVD algorithm in a weighted inner product space.

Given a matrix $U$ and its M-weighted SVD $U = V_M \Sigma_M W_M^*$, if we add one more column $c$ onto $U$, i.e., $\begin{bmatrix} U & c \end{bmatrix}$. We have the following matrix decomposition:

$$[U \quad c] = \begin{bmatrix} V_M \Sigma_M W_M^* & c \end{bmatrix} = \begin{bmatrix} V_M & \frac{c - V_M V_M^* c}{\|c - V_M V_M^* c\|_M} \end{bmatrix} \begin{bmatrix} \Sigma_M & V_M^* c \\ 0 & \|c - V_M V_M^* c\|_M \end{bmatrix} \begin{bmatrix} W_M & 0 \\ 0 & 1 \end{bmatrix}^T.$$

Let $Q = \begin{bmatrix} \Sigma_M & V_M^* c \\ 0 & \|c - V_M V_M^* c\|_M \end{bmatrix}$ and its standard SVD $Q = V_Q \Sigma_Q W_Q^T$, then we have

$$\begin{bmatrix} U & c \end{bmatrix} = \begin{bmatrix} V_M & \frac{c - V_M V_M^* c}{\|c - V_M V_M^* c\|_M} \end{bmatrix} V_Q \Sigma_Q W_Q^T \begin{bmatrix} W_M & 0 \\ 0 & 1 \end{bmatrix}^T \tag{3.64}$$

$$= \left( \begin{bmatrix} V_M & \frac{c - V_M V_M^* c}{\|c - V_M V_M^* c\|_M} \end{bmatrix} V_Q \right) \Sigma_Q \left( \begin{bmatrix} W_M & 0 \\ 0 & 1 \end{bmatrix} W_Q \right)^T. \tag{3.65}$$

We can formally verify that $\begin{bmatrix} V_M & \frac{c - V_M V_M^* c}{\|c - V_M V_M^* c\|_M} \end{bmatrix} V_Q$ is a M-orthonormal matrix and $\begin{bmatrix} W_M & 0 \\ 0 & 1 \end{bmatrix} W_Q$ is orthonormal with respect to standard inner product. Therefore, (3.65) is a M-weighted SVD of $\begin{bmatrix} U & c \end{bmatrix}$. Again, if $\|c - V_M V_M^* c\|_M$ is small enough, we do the decomposition similar to what we have done in section 3.5.2.

We then summarize the incremental M-weighted SVD algorithm [56] as follows:

**Algorithm 5** *Step 0: Specified a tolerance $\epsilon_1$, $\epsilon_2$ and initialize the M-weighted SVD for the first data vector.*

*Step 1: Given $U = V_M \Sigma_M W_M$ and new data vector $c$, compute $\frac{c - V_M V_M^* c}{\|c - V_M V_M^* c\|}$ and assemble $Q = \begin{bmatrix} \Sigma_M & V_M^* c \\ 0 & \|c - V_M V_M^* c\|_M \end{bmatrix}$.*

*Step 2:*

*(1) If* $\|c - V_M V_M^* c\|_M \leq \epsilon_3$*, compute the SVD of* $R = \begin{bmatrix} \Sigma & V_M^* c \end{bmatrix}$ *as* $R = V_R \Sigma_R W_R^T$*, then update* $V = V V_R, \Sigma = \Sigma_R$*, and* $W_M^T = (W_M W_R)^T$.

*(2) If* $\|c - V_M V_M^* c\|_M \geq \epsilon_3$*, compute the SVD of* $Q = \begin{bmatrix} \Sigma & V_M^* c \\ 0 & \|c - V_M V_M^* c\|_M \end{bmatrix}$ *as* $Q = V_Q \Sigma_Q W_Q^T$*, then update* $V_M = \begin{bmatrix} V_M & \frac{c - V_M V_M^* c}{\|c - V_M V_M^* c\|_M} \end{bmatrix} V_Q$*,* $\Sigma_M = \Sigma_Q$ *and* $W_M^T = \left( \begin{bmatrix} W_M & 0 \\ 0 & 1 \end{bmatrix} W_Q \right)^T$.

*Step 3: If* $V_M(:, end)' M V_M(:, 1) > \epsilon_2$*, do the Gram-Schmidt orthogonalization.*

*Step 4: Do the truncation for* $V_M, \Sigma_M$ *and* $W_M$ *if the singular value is less than the specified tolerance* $\epsilon_1$.

*Step 5: Do Step 1-5 until the last data vector is introduced.*

**3.5.3. Error Analysis.** For the proposed incremental algorithm, we are also interested in the information loss during such process . In [58], authors there developed a prior error estimate of information loss in sense of an infinity matrix norm, which can be improved by Hilbert-Schmidt norm since the data information is measured by the Hilbert-Schmidt norm (see Lemma 1). In this section, an more accurate error estimation will be presented to quantify the information loss. For this purpose, we first recall some properties of the Hilbert-Schmidt norm in finite dimension space. Note that the matrix representation of the POD operator is exactly the data matrix, in this sense, we abuse the Hilbert-Schmidt norm of the data matrix for presentation.

**Property 1**: The Hilbert-Schmidt norm is independent of the choice on the total orthonormal sequence, and $T$ is Hilbert-Schmidt iff $T^*$ is Hilbert-Schmidt.

**Property 2**: If $T$ is compact and $\sum_{n\geq 1} \sigma_n^2 < \infty$, then $T$ is Hilbert-Schmidt and $\|T\|_{HS}^2 = \sum_{n\geq 1} \sigma_n^2$.

Denote $C^j$ as a data matrix with $j$ column vectors, $c^{j+1}$ as a new data column vector. In finite dimension space, we verify the following facts about Hilbert-Schmidt norm.

**Fact 1**: $\|C^{k+1}\|_{HS} = \|C^k\|_{HS} + \|c^{k+1}\|_{\mathbb{R}^n_M}$, where $C^{k+1} = [C^k \quad c^{k+1}]$.

**Proof:** We choose $\hat{e}_1, \hat{e}_2, \hat{e}_3, ..., \hat{e}_m, \in \mathbb{R}^m$ and $e_1, e_2, e_3, ..., e_m, e_{m+1} \in \mathbb{R}^{m+1}$ as two sets orthonormal sequences, then based on the definition of $HS$ norm, we calculate

$$\|C^{k+1}\|^2_{HS} = \sum_{i=1}^{m+1} \|C^{k+1} e_i\|^2_{\mathbb{R}^n_M}$$
$$= \sum_{i=1}^{m} \|C^k \hat{e}_i\|^2_{\mathbb{R}^n_M} + \|C^{k+1} e_i\|^2_{\mathbb{R}^n_M}$$
$$= \|C^k\|^2_{HS} + \|c^{k+1}\|^2_{\mathbb{R}^n_M}.$$

or

$$\|C^{k+1}\|^2_{HS} = \sum_{i=1}^{k+1} \|c^i\|^2_{\mathbb{R}^n_M} = \sum_{i=1}^{k} \|c^i\|^2_{\mathbb{R}^n_M} + \|c^{k+1}\|^2_{\mathbb{R}^n_M}$$
$$= \|C^k\|^2_{HS} + \|c^{k+1}\|^2_{\mathbb{R}^n_M}.$$

**Fact 2**: If the $p$ truncation is only applied in the algorithm, we have $\|C^{k+1} - \hat{C}^{k+1}\|_{HS} = \|c^{k+1} - VV^* c\|_{\mathbb{R}^n_M}$.

**Proof:**

$$\|C^{k+1} - \widehat{C}^{k+1}\|^2_{HS} = \|[C^k \quad c^{k+1}] - [C^k \quad VV^* c^{k+1}]\|^2_{HS}$$
$$= \|[0 \quad c^{k+1} - VV^* c^{k+1}]\|^2_{HS}$$
$$= \|c^{k+1} - VV^* c^{k+1}\|^2_{\mathbb{R}^n_M},$$

where the last equality is a consequence of **Fact 1**.

**Fact 3**: If we only apply SVD truncation on $C^{k+1}$, we have $\|C^{k+1} - \bar{C}^{k+1}\|_{HS}^2 = \sum_{i>l} \sigma_i^2$, $l$ is where the singular value truncation starts. This fact is a consequence of **Property 2**.

We denote $U^k$, $\widehat{U}^k$, and $\bar{U}^k$ as the full data matrix, data matrix with $p$ truncation, and the final data matrix at the $k$ incremental step. Using all three facts above, we quantify the information loss in the incremental POD algorithm:

$$
\begin{aligned}
&\|U^{k+1} - \bar{U}^{k+1}\|_{HS} = \|U^{k+1} - [\bar{U}^k \quad c^{k+1}] + [\bar{U}^k \quad c^{k+1}] - \widehat{U}^{k+1} + \widehat{U}^{k+1} - \bar{U}^{k+1}\|_{HS} \\
&\leq \|U^{k+1} - [\bar{U}^k \quad c^{k+1}]\|_{HS} + \|[\bar{U}^k \quad c^{k+1}] - \widehat{U}^{k+1}\|_{HS} + \|\widehat{U}^{k+1} - \bar{U}^{k+1}\|_{HS} \\
&\leq \|[U^k - \bar{U}^k \quad 0]\|_{HS} + \|[0 \quad c^{k+1} - V^k V^{k*} c^{k+1}]\|_{HS} + \|\widehat{U}^{k+1} - \bar{U}^{k+1}\|_{HS} \\
&\leq \|U^k - \bar{U}^k\|_{HS} + \|c^{k+1} - V^k V^{k*} c^{k+1}\|_{\mathbb{R}^n_M} + (\sum_{i>l} (\sigma_i^{k+1}))^{\frac{1}{2}},
\end{aligned}
$$

where $V^k$ is the left singular vector matrix of $\bar{U}^k$ and $\sigma_i^{k+1}$ is the $i$th singular value of $\widehat{U}^{k+1}$.

We then estimate the error accumulation along the incremental process as follows:

$$
e^{k+1} \leq \begin{cases}
e^k & \text{if no truncation is applied,} \\
e^k + \|c^{k+1} - V^k V^{k*} c^{k+1}\|_{\mathbb{R}^n_M} & \text{if only p truncation is applied,} \\
e^k + (\sum_{i>l} (\sigma_i^{k+1})^2)^{\frac{1}{2}} & \text{if only singular value truncation is applied,} \\
e^k + \|c^{k+1} - V^k V^{k*} c^{k+1}\|_{\mathbb{R}^n_M} + (\sum_{i>l} (\sigma_i^{k+1})^2)^{\frac{1}{2}} & \text{if all truncations are applied.}
\end{cases}
$$

This error estimation provides guideline to set up the truncation thresholds for the incremental POD algorithm.

In addition, if one wants to monitor the ratio between the captured information and the full information along the incremental algorithm, we have the following estimation:

$$(r^{k+1})^2 = \frac{\|\bar{U}^{k+1}\|_{HS}^2}{\|U^{k+1}\|_{HS}^2} = \frac{\|\bar{U}^{k+1}\|_{HS}^2}{\sum_{i=1}^{k+1} \|c^k\|_{\mathbb{R}_M^n}^2} = \frac{\sum_{i=1}^{l} (\sigma_i^{k+1})^2}{\sum_{i=1}^{k+1} \|c^k\|_{\mathbb{R}_M^n}^2}.$$

For this consideration, we need to compute the $M$ weighted norm of the new introduced data $c^i$ at each step and add it to the next step for calculating the full information $\sum_{i=1}^{k+1} \|c^k\|_{\mathbb{R}_M^n}^2$.

**3.5.4. Data Compression in PDE Simulation.** Consider a time dependent partial differential equation:

$$\begin{aligned} &\frac{\partial u}{\partial t} - Au = f \quad \text{in } \Omega \times (0, T], \\ &u(\cdot, 0) = u_0 \quad \text{in } \Omega, \end{aligned} \tag{3.66}$$

where $A$ is a generic linear (or nonlinear) operator. In discrete level, we can discretize (3.66) as

$$\begin{aligned} &\frac{u_{n+1} - u_n}{\tau_n} - Au_{n+1} = f_{n+1} \quad \text{in } \Omega \times (0, T], \\ &u^0 = u_0 \quad \text{in } \Omega, \end{aligned} \tag{3.67}$$

where $\tau$ is the temporal step size.

We intend to compress data $Z = \{u_i\}_{i=1}^m \in X$ into a smaller size, this problem can be formulated as:

$$\min_{\gamma^i \in X, i=1,2\ldots r} J_r = \sum_{k=1}^{m} \tau_i \|u_k - u_k^r\|_X^2. \tag{3.68}$$

In physics, it means that we are trying to find a optimal $r$ dimension basis to capture certain energy defined by $\|\cdot\|_X$ along time as much as possible. Or (3.68) can be interpreted as the discrete Riemann sum of a spatial and temporary integral mathematically.

If we consider $X$ is a finite dimension Hilbert space in domain $\Omega$, we observe

$$\|x\|_X = (\sum_{i=1}^{n} x_i\psi_i, \sum_{i=1}^{n} x_i\psi_i)_X = x'Mx = (x,x)_M, \tag{3.69}$$

where $\psi_i$ could be finite element basis or other basis, and $M = [\int_\Omega \psi_i\psi_j dx]$ is a positive definite matrix. Based on the previous arguments, problem (3.68) is essentially a $M$-weighted SVD of the operator $L : \mathbb{R}^m \to \mathbb{R}^n_M$:

$$Lg = \sum_{k=1}^{m} g_k \tau_i^{\frac{1}{2}} u_k = Z\Delta^{\frac{1}{2}} g \quad \forall g = \{g_i\}_{i=1}^m \in \mathbb{R}^m, \tag{3.70}$$

where $\Delta = diag[\tau_1 \quad \tau_2 \quad \tau_3 \quad ...... \quad \tau_m]$.

We can easily prove that the SVD of the operator: $Z\Delta^{\frac{1}{2}} : \mathbb{R}^m \to \mathbb{R}^n_M$ is equivalent to the SVD of the operator: $Z\Delta : \mathbb{R}^m_\Delta \to \mathbb{R}^n_M$, [59]. This is mainly because that problem (3.68) is also equivalent to find the best HS operator approximation of the discrete linear operator $K$, which is defined by

$$Kg = \sum_{i=1}^{m} \tau_i u_i g_i. \tag{3.71}$$

Operator $L$ is essentially an approximation of the continuous $POD$ operator:

$$K_c g = \int_T u(x,t) g(t) dt. \tag{3.72}$$

In discrete level, we assume that $u(x,t)$ and $g(t)$ are piecewise constant function in term of time, i.e.,

$$u(x,t) = \sum_{i=1}^{m} u_i \chi_i(t), \quad g(t) = \sum_{i=1}^{m} g_i \chi_i(t), \tag{3.73}$$

where $\chi_i(t)$ are characteristic functions defined for time interval $(t_i, t_{i+1}]$.

From this point view, operator $K : L^2(0,T) \to X$ has the matrix representation $Z\Delta$, i.e., $Z\Delta : \mathbb{R}^m_\Delta \to \mathbb{R}^n_M$ in finite dimension Hilbert spaces. This is a more general perspective to have insight of POD operators, which also gives us an alternative to do the matrix decomposition: let $U\Delta_i := V\Sigma W^*$ be the SVD of $U\Delta : \mathbb{R}^m_\Delta \to \mathbb{R}^n_M$, then we have

$$\begin{aligned}
\begin{bmatrix} U & c \end{bmatrix} \Delta &= \begin{bmatrix} V\Sigma W^T \Delta_i & \tau_{i+1} c \end{bmatrix} \\
&= \begin{bmatrix} V & \frac{c-VV^*c}{\|c-VV^*c\|_M} \end{bmatrix} \begin{bmatrix} \Sigma W^T \Delta_i & \tau_{i+1} V^* c \\ 0 & \tau_{i+1} \|c - VV^*c\|_M \end{bmatrix} \\
&= \begin{bmatrix} V & \frac{c-VV^*c}{\|c-VV^*c\|_M} \end{bmatrix} \begin{bmatrix} \Sigma & \tau_{i+1}^{\frac{1}{2}} V^* c \\ 0 & \tau_{i+1}^{\frac{1}{2}} \|c - VV^*c\|_M \end{bmatrix} \begin{bmatrix} W & 0 \\ 0 & \tau_{i+1}^{-\frac{1}{2}} \end{bmatrix}^T \Delta.
\end{aligned}$$

Let $Q = \begin{bmatrix} \Sigma & \tau_{i+1}^{\frac{1}{2}} V^* c \\ 0 & \tau_{i+1}^{\frac{1}{2}} \|c - VV^*c\|_M \end{bmatrix}$ and $Q = V_Q \Sigma_Q W_Q^T$, then

$$\begin{bmatrix} U & c \end{bmatrix} \Delta = \left( \begin{bmatrix} V & \frac{c-VV^*c}{\|c-VV^*c\|_M} \end{bmatrix} V_Q \right) \Sigma_Q \left( \begin{bmatrix} W & 0 \\ 0 & \tau_{i+1}^{-\frac{1}{2}} \end{bmatrix} W_Q \right)^T \Delta.$$

It is not hard to verify that $\begin{bmatrix} V & \frac{c-VV^*c}{\|c-VV^*c\|_M} \end{bmatrix} V_Q$ is M-orthonormal and $\begin{bmatrix} W & 0 \\ 0 & \tau_{i+1}^{-\frac{1}{2}} \end{bmatrix} W_Q$ is $\Delta$-orthonormal.

Based on the above decomposition argument, we provide the following algorithm [59]:

**Algorithm 6** *Step 0: Specified tolerance $\epsilon_1$, $\epsilon_2$ and initialize $U$, $\Delta$ and $U\Delta = V\Sigma W^*$.*

*Step 1: Solve the PDE one more step and use the output as new vector data $c$, generate new left orthonormal matrix $\begin{bmatrix} V & \frac{c-VV^*c}{\|c-VV^*c\|_M} \end{bmatrix}$ and matrix $Q$ as $\begin{bmatrix} \Sigma & \tau_{i+1}^{\frac{1}{2}} V^* c \\ 0 & \tau_{i+1}^{\frac{1}{2}} \|c - VV^*c\|_M \end{bmatrix}$.*

*Step 2:*

*(1) If* $\|c - VV^*c\|_M \le \epsilon_3$*, compute the SVD of* $R = \begin{bmatrix} \Sigma & V^*c \end{bmatrix}$ *as* $R = V_R \Sigma_R W_R^T$*, then update* $V = VV_R, \Sigma = \Sigma_R$*, and* $W^T = (WW_R)^T$.

*(2) If* $\|c - VV^*c\|_M \ge \epsilon_3$*, compute the SVD of* $Q = \begin{bmatrix} \Sigma & \tau_{i+1}^{\frac{1}{2}} V^*c \\ 0 & \tau_{i+1}^{\frac{1}{2}} \|c - VV^*c\|_M \end{bmatrix}$ *as* $Q = V_Q \Sigma_Q W_Q^T$*, then update* $V_M = \begin{bmatrix} V_M & \frac{c - VV^*c}{\|c - VV^*c\|_M} \end{bmatrix} V_Q$*,* $\Sigma = \Sigma_Q$*,* $W^T = \left( \begin{bmatrix} W & 0 \\ 0 & \tau_{i+1}^{-\frac{1}{2}} \end{bmatrix} W_Q \right)^T$. *and* $\Delta = diag \begin{bmatrix} \Delta & \tau_{i+1} \end{bmatrix}$.

*Step 3: If* $(V(:, end), V(:, 1))_M > \epsilon_2$*, do the Gram-Schmidt orthogonalization.*

*Step 4: Do the truncation for* $V, \Sigma$ *and* $W$ *if the singular value is less than tolerance* $\epsilon_1$.

*Step 6: Do Step 1-4 until the PDE runs out of the final time moment.*

**Remark 9** *Note that in the algorithm 6, we do not need to store the original data after using them, hence computer memory would be saved when truncations are applied.*

**3.5.5. Data Compression in Variational Data Assimilation.** Consider the variational data assimilation problem that identifies an initial condition for a dynamics system:

$$\min_{u_0 \in U_{ad}} J(u_0) = \frac{1}{2} \int_0^T \|\hat{u} - u(u_0)\|_{\mathcal{H}}^2 dt + \frac{\gamma}{2} \|u_0\|_H^2, \tag{3.74}$$

subject to

$$\begin{aligned} \frac{\partial u}{\partial t} - Au &= f \quad \text{in } \Omega \times (0, T], \\ u(\cdot, 0) &= u_0 \quad \text{in } \Omega. \end{aligned} \tag{3.75}$$

In discrete level, this optimization problem can be stated as

$$\min_{u_{0,h} \in U_h} J_h(u_{0,h}) = \frac{1}{2} \tau \sum_{n=1}^{N} \|\hat{u}^n - u_h^n\|_{\mathcal{H}}^2 + \frac{\gamma}{2} \|u_{0,h}\|_H^2. \tag{3.76}$$

subject to

$$\begin{cases} \dfrac{u_{n+1} - u_n}{\tau} - Au_{n+1} = f_{n+1} \quad \text{in } \Omega \times (0, T], \\ u^0 = u_0 \quad \text{in } \Omega. \end{cases} \tag{3.77}$$

A standard calculus variation of the cost functional (3.76) gives the following optimality system:

$$\begin{cases} \dfrac{u_{n+1} - u_n}{\tau} - Au_{n+1} = f_{n+1} \quad \text{in } \Omega \times (0, T], \\ u^0 = u_0 \quad \text{in } \Omega, \\ -\dfrac{u_{n+1} - u_n}{\tau} - A^* u_{n+1} = \hat{u}^{n+1} - u_h^{n+1} \quad \text{in } \Omega \times (0, T], \\ u^N = 0 \quad \text{in } \Omega, \\ u_{0,h} = \dfrac{1}{\gamma} u_h^{*0} \end{cases} \tag{3.78}$$

for $n = 0, 1, 2, 3.....N - 1$.

As mentioned at the beginning of section 3.5, the use of gradient method to solve (3.78) will encounter storage difficulties for $\{u^n\}_{n=1}^N$. To overcome this difficulty, we apply the incremental POD algorithm developed in above sections to the gradient descent methods.

**Algorithm 7** *Step 0 (Initialization): Specify a convergence tolerance $\epsilon$, guess initial function $u_{0,h}$, and start the iteration step $i = 1$.*

*Step 1 (Forward phase): Use $u_{0,h}$ as initial condition to solve first equation in (3.78) forward to obtain $u^n$, at the same time implement Algorithm 6 to compress data matrix $\{u^n\}_{n=1}^N$ as $V_M \Sigma_M W_M^*$ or $V_M \Sigma_M W^T \Delta$ .*

*Step 2 (Backward phase): At each time moment $t_n$, reconstruct the data $u^n = V_M \Sigma_M W^T(:, n)$ from the compressed data in Step 1 and plug the reconstructed data to the second equation of (3.78) to solve backward for $u^{*0}$. Note that the the time related information is indicated in matrix $W$.*

*Step 3 (Inexact line search for $\eta^{i+1}$):*

*(1) Initialize a constant $\eta^{i+1} \geq 1$, set $0 < \rho < 1$ and $0 < \delta < 1$;*

*(2) Use $u_{0,h}^{(i+1)} = u_{0,h}^{(i)} + \eta^{i+1}(u_h^{*0(i)} - \gamma u_{0,h}^{(i)})$ as initial value to solve equation (3.78) forward to obtain $u_h^n$ for computing $J_h(u_{0,h}^{(i+1)})$;*

*(3) Update $\eta^{i+1} = \rho\eta^{i+1}$ until inequality $J_h(\vec{u}_{0,h}^{(i+1)}) \leq J_h(\vec{u}_{0,h}^{(i)}) + \delta\eta^{i+1}\langle J_h'(\vec{u}_{0,h}^{(i)}), \vec{u}_h^{*0(i)} - \gamma\vec{u}_{0,h}^{(i)}\rangle_{U_h^*\times U_h}$ is attained.*

*(4) Output $\eta^{i+1}$.*

*Step 4 (Update phase): Use a proper learning rate $\eta^{i+1}$ from step 3 and then update*

$$u_{0,h}^{(i+1)} = u_{0,h}^{(i)} + \eta^{i+1}(u_h^{*0(i)} - \gamma u_{0,h}^{(i)}).$$

*Step 5 (Criteria for stopping the iteration): Compute $\|u_h^{*0(i)} - \gamma u_{0,h}^{(i)}\|_H$, if $\|u_h^{*0(i)} - \gamma u_{0,h}^{(i)}\|_H \leq \epsilon$ then stop; otherwise, increase i by 1 and go back to Step 1.*

### 3.6. PARALLEL ALGORITHM

In this section, we start from a new perspective to optimize the computational resource for the variational data assimilation problem, i.e., parallelization. Recall solving the variational data assimilation problem (3.1)-(3.2) ends up with dealing with the optimality system:

$$\begin{cases} \frac{\partial u}{\partial t} + Au = f, \\ u(0) = u_0, \\ -\frac{\partial u^*}{\partial t} + A^*u^* = \hat{u} - u, \\ u^*(T) = 0, \\ u^*(0) - \gamma u(0) = 0. \end{cases} \tag{3.79}$$

One main challenge to solve (3.79) is induced by its coupled temporal and space nature. We hereby propose a time parallel algorithm to decouple the system along time and mitigate the computational burden.

The time parallel algorithm was first developed by Lions, Maday and Turinici for addressing the computational cost in ODE and PDE [60, 61], it is then widely extended to solve problems with extreme computational cost, such as optimal control [62–64]. Motivated by their philosophy, we parallelize the variational data assimilation problem to improve the computation efficiency. We first partition the time domain of the VDA optimality system into smaller sub-intervals, and do computation in a parallel manner for each sub-interval. The communication among each sub-interval is carried on by a multiple shooting method. By such design, the original problem (3.79) is finally transformed to solving a nonlinear equation, which can be handled using a root-finding method, such as Newton's method.

For notation convenience, we change the adjoint variable $u^*$ in (3.79) to $p$ for the presentation only in Section 3.6. The parallel algorithm starts by guessing the solution of $u$ and $p$ in (3.79) at time $t_k$, $k = 0, 1, 2, 3....N$, where $t_k$ is the grid point partitioned from interval $(0, T]$. Denote these guesses with $U^k$ and $P^k$, let $u(t_{k+1}; P_{k+1}, U^k)$ and $p(t_k; P_{k+1}, U^k)$ be solutions of the following two-point boundary problem:

$$\begin{cases} u_t + Au = f, \quad \text{in } \Omega \times (t_k, t_{k+1}], \\ u(\cdot, t_k) = U^k, \quad \text{in } \Omega, \\ -p_t + A^* p = \hat{u} - u, \quad \text{in } \Omega \times [t_{k+1}, t_k), \\ p(\cdot, t_{k+1}) = P^{k+1}, \quad \text{in } \Omega. \end{cases} \tag{3.80}$$

where $U^k$ and $P^k$ are also called shooting variables. If $(U^k,\ P^k)$ and $(u(t_k;P_k,U^{k-1}),$ $p(t_k;P_{k+1},U^k))$ are all solutions of (3.79) at time $\{t_k\}_{k=1}^N$, they should satisfy the following matching conditions:

$$\begin{cases} U^0 - \dfrac{1}{\gamma}P^0 = 0, \\ U^1 - u(t_1;U^0,P^1) = 0, \\ U^2 - u(t_2;U^1,P^2) = 0, \\ U^3 - u(t_3;U^2,P^3) = 0, \\ \ldots\ldots\ldots. \\ U^N - u(t_N;U^{N-1},P^N) = 0, \\ P^1 - u(t_1;U^1,P^2) = 0, \\ P^2 - u(t_2;U^2,P^3) = 0, \\ P^3 - u(t_3;U^3,P^4) = 0, \\ \ldots\ldots\ldots. \\ P^{N-1} - p(t_{N-1};U^{N-1},P^N) = 0, \\ P^N = 0. \end{cases} \tag{3.81}$$

For convenience, we abuse notations for a while, that abbreviates $u(t_k;P_k,U^{k-1})$ and $p(t_k;P_{k+1},U^k)$ as $u(t_k)$ and $p(t_k)$, respectively. They will be equivalent notation in the following presentation. We use (3.81) to define a nonlinear equation $F(\vec{X}) = 0$, where $\vec{X} = (U^0,U^1,U^2,...,U^{N-1},U^N,P^0,P^1,P^2,P^3,...,P^{N-1},P^N)^T$. To solve for $\vec{X}$, we use the Newton's method, i.e.,

$$\vec{X}_{k+1} = \vec{X}_k - J^{-1}(\vec{X}_k)F(\vec{X}_k), \quad J^{-1} \text{ is the inverse of Jacobian matrix,}$$

which is also equivalent to

$$J(\vec{X}_k)(\vec{X}_{k+1} - \vec{X}_k) = -F(\vec{X}_k). \tag{3.82}$$

The next step is to calculate the Jacobian matrix:

$$\begin{pmatrix} 1 & 0 & \cdots & 0 & -\frac{1}{\gamma} & 0 & 0 & \cdots & 0 & 0 \\ -\frac{\partial u(t_1)}{\partial U^0} & 1 & \cdots & 0 & 0 & \frac{\partial u(t_1)}{\partial P^1} & 0 & \cdots & 0 & 0 \\ 0 & -\frac{\partial u(t_2)}{\partial U^1} & \cdots & 0 & 0 & 0 & -\frac{\partial u(t_2)}{\partial P^2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & 0 & 0 & 0 & -\frac{\partial u(t_N)}{\partial P^N} \\ -\frac{\partial p(t_0)}{\partial U^0} & 0 & \cdots & 0 & 1 & -\frac{\partial p(t_0)}{\partial P^1} & 0 & \cdots & 0 & 0 \\ 0 & -\frac{\partial p(t_1)}{\partial U^1} & \cdots & 0 & 0 & 1 & -\frac{\partial p(t_1)}{\partial P^2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -\frac{\partial p(t_{N-1})}{\partial U^{N-1}} & 0 & 1 & 0 & \cdots & 1 & -\frac{\partial p(t_{N-1})}{\partial P^N} \\ 0 & 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Based on the Jacobian matrix, we can explicitly write (3.82) as:

$$\begin{cases} U_{i+1}^0 - U_i^0 - \dfrac{1}{\gamma}(P_{i+1}^0 - P_i^0) = \dfrac{1}{\gamma}P_i^0 - U_i^0 \\ -\dfrac{\partial u(t_1)}{\partial U_i^0}(U_{i+1}^0 - U_i^0) + U_{i+1}^1 - Ui^1 - \dfrac{u(t_1)}{\partial P_i^1}(P_{i+1}^1 - P_i^1) = u(t_1) - U_i^1, \\ \cdots\cdots\cdots\cdots \\ -\dfrac{\partial u(t_N)}{\partial U_i^{N-1}}(U_{i+1}^{N-1} - U_i^{N-1}) + U_{i+1}^N - Ui^N - \dfrac{u(t_N)}{\partial P_i^N}(P_{i+1}^N - P_i^N) = u(t_N) - U_i^N, \\ -\dfrac{\partial p(t_0)}{\partial U_i^0}(U_{i+1}^0 - U_i^0) + P_{i+1}^0 - P_i^0 - \dfrac{p(t_0)}{\partial P_i^1}(P_{i+1}^1 - P_i^1) = p(t_0) - P_i^0, \\ \cdots\cdots\cdots\cdots\cdots \\ -\dfrac{\partial p(t_{N-1})}{\partial U_i^{N-1}}(U_{i+1}^{N-1} - U_i^{N-1}) + P_{i+1}^{N-1} - Pi^{N-1} - \dfrac{p(t_{N-1})}{\partial P_i^N}(P_{i+1}^N - P_i^N) \\ = p(t_{N-1}) - P_i^{N-1}, \\ P_{i+1}^N - P_i^N = -P_i^N. \end{cases} \tag{3.83}$$

Rearranging (3.83) gives us the iterative update of $\vec{X}$:

$$U_{i+1}^0 - U_i^0 - \frac{1}{\gamma}(P_{i+1}^0 - P_i^0) = \frac{1}{\gamma}P_i^0 - U_i^0 \tag{3.84}$$

$$U_{i+1}^1 = u(t_1) + \frac{u(t_1)}{\partial P_i^1}(P_{i+1}^1 - P_i^1) + \frac{\partial u(t_1)}{\partial U_i^0}(U_{i+1}^0 - U_i^0), \tag{3.85}$$

$$U_{i+1}^2 = u(t_2) + \frac{u(t_2)}{\partial P_i^2}(P_{i+1}^2 - P_i^2) + \frac{\partial u(t_2)}{\partial U_i^1}(U_{i+1}^1 - U_i^1), \tag{3.86}$$

$$\cdots\cdots\cdots\cdots \tag{3.87}$$

$$U_{i+1}^{N-1} = u(t_{N-1}) + \frac{\partial u(t_{N-1})}{\partial U_i^{N-2}}(U_{i+1}^{N-2} - U_i^{N-2}) + \frac{u(t_{N-1})}{\partial P_i^{N-1}}(P_{i+1}^{N-1} - P_i^{N-1}), \tag{3.88}$$

$$U_{i+1}^N = u(t_N) + \frac{\partial u(t_N)}{\partial U_i^{N-1}}(U_{i+1}^{N-1} - U_i^{N-1}) + \frac{u(t_N)}{\partial P_i^N}(P_{i+1}^N - P_i^N), \tag{3.89}$$

$$P_{i+1}^0 = p(t_0) + \frac{\partial p(t_0)}{\partial U_i^0}(U_{i+1}^0 - U_i^0) + \frac{p(t_0)}{\partial P_i^1}(P_{i+1}^1 - P_i^1), \tag{3.90}$$

$$P_{i+1}^1 = p(t_1) + \frac{\partial p(t_1)}{\partial U_i^1}(U_{i+1}^1 - U_i^1) + \frac{p(t_1)}{\partial P_i^2}(P_{i+1}^2 - P_i^1), \tag{3.91}$$

$$\cdots\cdots\cdots\cdots\cdots \tag{3.92}$$

$$P_{i+1}^{N-1} = p(t_{N-1}) + \frac{\partial p(t_{N-1})}{\partial U_i^{N-1}}(U_{i+1}^{N-1} - U_i^{N-1}) + \frac{p(t_{N-1})}{\partial P_i^N}(P_{i+1}^N - P_i^N), \tag{3.93}$$

$$P_{i+1}^N = 0. \tag{3.94}$$

Define $\delta^p = \frac{\partial p(t_k)}{\partial U_i^k}(U_{i+1}^k - U_i^k) + \frac{p(t_k)}{\partial P_i^{k+1}}(P_{i+1}^{k+1} - P_i^{k+1})$, $\delta^u = \frac{\partial u(t_{k+1})}{\partial U_i^k}(U_{i+1}^k - U_i^k) + \frac{u(t_{k+1})}{\partial P_i^{k+1}}(P_{i+1}^{k+1} - P_i^{k+1})$. We observe that $\delta^p$ and $\delta^u$ are the total variation of $u, p$ in (3.80) with respect to variables $U^k$ and $P^{k+1}$ at interval $(t_k, t_{k+1}]$. Therefore, the variation can be obtained by solving the following equation, for $k = 0, 1, 2, \cdot, N-1$,:

$$\begin{cases} \delta_t^u + A\delta^u = 0, \\ \delta^u(t_k) = U_{i+1}^k - U_i^k, \\ -\delta_t^p + A^*\delta^p = -\delta^u, \\ \delta^p(t_{k+1}) = P_{i+1}^{k+1} - P_i^{k+1}. \end{cases} \tag{3.95}$$

Note that

$$\delta^u(t_k) = U_{i+1}^k - U_i^k = U_{i+1}^k - u(t_k) + u(t_k) - U_i^k = \sigma_u^k + S_k^u, \tag{3.96}$$

$$\delta^p(t_{k+1}) = P_{i+1}^{k+1} - P_i^{k+1} = P_{i+1}^{k+1} - p(t_k) + p(t_k) - P_i^{k+1} = \sigma_p^{k+1} + S_{k+1}^p, \tag{3.97}$$

where $S_k^u = u(t_k) - U_i^k$ and $S_{k+1}^p = p(t_k) - P_i^{k+1}$. Recall (3.84) that $\sigma_u^k = U_{i+1}^k - u(t_k)$ and $\sigma_p^{k+1} = P_{i+1}^{k+1} - p(t_k)$. Using a backward finite difference scheme to discretize (3.95), for $k = 0, 2, 3, ..., N-1$, we have

$$\begin{cases} \dfrac{\delta_u^{k+1} - \delta^u(t_k)}{\tau} + A\delta_u^{k+1} = 0, \\ \delta^u(t_k) = U_{i+1}^k - U_i^k, \\ \dfrac{\delta_p^k - \delta^p(t_{k+1})}{\tau} + A^*\delta_p^k = -\delta_u^{k+1}, \\ \delta^p(t_{k+1}) = P_{i+1}^{k+1} - P_i^{k+1}. \end{cases} \tag{3.98}$$

By a simply algebraic work on (3.98) with (3.96) - (3.97), the locally solving (3.95) is then approximately equivalent to solving the following equation globally:

$$\begin{cases} \dfrac{\sigma_u^{k+1} - \sigma_u^k}{\tau} + A\sigma_u^{k+1} = \dfrac{S_u^k}{\tau}, \\ \sigma_u(\cdot, 0) = \dfrac{1}{\gamma}\sigma_p^0, \\ \dfrac{\sigma_p^k - \sigma_p^{k+1}}{\tau} + A^*\sigma_p^k = -\sigma_u^{k+1} + \dfrac{S_p^{k+1}}{\tau}, \\ \sigma_p(\cdot, T) = 0. \end{cases} \tag{3.99}$$

We finally have an explicit and simplified update of $\vec{X}$:

$$U_{i+1}^0 = \frac{1}{\gamma}P_{i+1}^0, \;\; U_{i+1}^1 = u(t_1) + \sigma_u^1, \;\; U_{i+1}^2 = u(t_2) + \sigma_u^2, \;\; \cdots \;\; U_{i+1}^N = u(t_N) + \sigma_u^N,$$

$$P_{i+1}^0 = p(t_0) + \sigma_p^0, \;\; P_{i+1}^1 = p(t_1) + \sigma_p^1, \;\; \cdots \;\; P_{i+1}^{N-1} = p(t_{N-1}) + \sigma_p^{N-1}, \;\; P_{i+1}^N = 0.$$

Now we can summarize the parallel algorithm as:

**Algorithm 8** *Step 0 (Initialization):Solving*

$$\begin{cases} u_t + Au = f, \quad in\ \Omega \times (0,T], \\ u(\cdot,0) = u_0, \quad in\ \Omega, \\ -p_t + A^*p = \hat{u} - u, \quad in\ \Omega \times [0,T), \\ p(\cdot,T) = 0, \quad in\ \Omega, \\ u_0 = \frac{1}{\gamma}p(0). \end{cases} \tag{3.100}$$

*in a coarse time grid discretization as an initial guesses of $U^k$ and $P^k$, which is cheap.*

*Step 1 (Parallel running): Solving*

$$\begin{cases} u_t + Au = f, \quad in\ \Omega \times (t_k, t_{k+1}], \\ u(\cdot,t_k) = U^k, \quad in\ \Omega, \\ -p_t + A^*p = \hat{u} - u, \quad in\ \Omega \times [t_{k+1}, t_k), \\ p(\cdot,t_{k+1}) = P^{k+1}, \quad in\ \Omega. \end{cases} \tag{3.101}$$

*independently with a refined discretization at each sub-interval in an parallel manner when $k \geq 1$. For $k = 0$, we need to solve the following equation:*

$$\begin{cases} u_t + Au = f, \quad in\ \Omega \times (t_0, t_1], \\ u(\cdot,0) = \frac{1}{\gamma}p(0), \quad in\ \Omega, \\ -p_t + A^*p = \hat{u} - u, \quad in\ \Omega \times [t_0, t_1), \\ p(\cdot,T) = P^1, \quad in\ \Omega. \end{cases} \tag{3.102}$$

*All purposes here are to obtain $u(t_k; P^k, U^{k-1})$ and $p(t_k; P^{k+1}, U^k)$.*

*Step 2 (Correction): Computing the defections* $S_u^k = u(t_k; P^k, U^{k-1}) - U^k$ *and* $S_p^k = p(t_k; P^{k+1}, U^k) - P^k$*, then using them to find the correction terms. This can be obtained by solving the equation:*

$$\begin{cases} \sigma_t^u + A\sigma_u = 0, \quad in\ \Omega \times (t_k, t_{k+1}], \\ \sigma^u(\cdot, t_k) = U_{i+1}^k - U_i^k, \quad in\ \Omega, \\ -\sigma_t^p + A^*\sigma_p = -\sigma_u, \quad in\ \Omega \times [t_k, t_{k+1}), \\ \sigma^p(\cdot, t_{k+1}) = P_{i+1}^{k+1} - P_i^{k+1}, \quad in\ \Omega, \end{cases} \tag{3.103}$$

*where i is the number of iteration from Newton's method. Note, this step is a local problem in continuous level. However, after discretization, (3.103) can be replaced by solving equation (3.99) globally.*

*Step 3 (Update):*

$$U^k = \sigma_u^k + u(t_k; P^k, U^{k-1}),$$
$$P^k = \sigma_p^k + p(t_k; P^{k+1}, U^k).$$

*Step 4 (Stop criteria): Running Step 1 and Step 3 until a pre-defined stop tolerance* $\epsilon$ *is greater than the quantity* $max\{\|\sigma_u^k\|_{\mathcal{H}}, \|\sigma_p^k\|_{\mathcal{H}}\}$ *.*

Due to the decoupling of time in the optimality system, the linear system from the discretization of each subproblem in the parallel scheme is not formidable anymore, so it is possible to discretize and solve them directly. An alternative to deal with the subproblems is using optimization technique, such as gradient descent methods, which is necessary to be mentioned to handle the optimality system with even larger spatial and temporal scale. Then an equivalent discrete illustration of Algorithm 8 is presented as follows:

**Algorithm 9** *Step 0 (Initialization): Solving*

$$\begin{cases} \dfrac{u^{k+1}-u^k}{\tau} + Au^{k+1} = f_{k+1}, \\ u^0 = \dfrac{1}{\gamma} p^0, \\ \dfrac{p^k - p^{k+1}}{\tau} + A^* p^k = \hat{u}^{k+1} - u^{k+1}, \\ p^N = 0, \end{cases} \tag{3.104}$$

*in a coarse grid discretization as an initial guesses of* $U^k$ *and* $P^k$*. This is equivalent to solve the minimization problem:*

$$F(u_0) = \frac{1}{2}\tau \sum_{k=1}^{N} \|\hat{u}^k - u^k\|_{\mathcal{H}}^2 + \frac{\gamma}{2}\|u_0\|_H^2 \tag{3.105}$$

*subject to*

$$\begin{cases} \dfrac{u^{k+1}-u^k}{\tau} + Au^{k+1} = f_{k+1}, \\ u^0 = u_0. \end{cases} \tag{3.106}$$

*Step 1 (Parallel running): Solving*

$$\begin{cases} \dfrac{u^{n+1}-u^n}{\tau} + Au^{n+1} = f_{n+1}, \\ u(t_k) = U^k, \\ \dfrac{p^n - p^{n+1}}{\tau} + A^* p^n = \hat{u}^{n+1} - u^{n+1}, \\ p(t_{k+1}) = P^{k+1}, \end{cases} \tag{3.107}$$

*independently with a refined discretization at each subinterval in a parallel manner when* $k \geq 1$*. Note that, in our presentation in this section,* $k$ *is a global time step index,* $n$ *is a local time step index. An important advantage in this step is the computation in* (3.107) *can be finished in one step without any iterations, i.e., no optimization is involved.*

*For $k = 0$, we are supposed to solve*

$$\begin{cases} \dfrac{u^{n+1} - u^n}{\tau} + Au^{n+1} = f_{n+1}, \\ u(0) = U^0, \\ \dfrac{p^n - p^{n+1}}{\tau} + A^* p^n = \hat{u}^{n+1} - u^{n+1}, \\ p(t_1) = P^1, \\ U^0 = \dfrac{1}{\gamma} P^0. \end{cases} \tag{3.108}$$

*For this purpose, we firstly solve an auxiliary equation*

$$\begin{cases} \dfrac{w^n - w^{n+1}}{\tau} + A^* w^n = 0, \\ w(t_1) = P_1, \end{cases} \tag{3.109}$$

*to attain $w(0)$. Then $u(0)$, $p(0; P^1, U^0)$ and $u(t_1; P^1, U^0)$ are given by solving the following minimizing problem:*

$$F(u_0) = \frac{1}{2}\tau \sum_{n=1}^{N} \|\hat{u}^n - u^n\|_{\mathcal{H}}^2 + \frac{\gamma}{2}\|u_0 - \frac{1}{\gamma} w(0)\|_H^2 \tag{3.110}$$

*subject to*

$$\begin{cases} \dfrac{u^{n+1} - u^n}{\tau} + Au^{n+1} = f_{n+1}, \\ u^0 = u_0. \end{cases} \tag{3.111}$$

*It is not difficult to see the solving of the (3.108) is equivalent to the solving of (3.109)-(3.111), then $p(0; P^1, U^0)$ is given by $p(0; P^1, U^0) = \gamma u(0)$.*

*Step 2 (Correction): Computing the defections $S_u^k = u(t_k; P^k, U^{k-1}) - U^k$ and $S_p^k = p(t_k; P^{k+1}, U^k) - P^k$, then using them to obtain the correction terms, which can be* *provided by solving*

$$\begin{cases} \dfrac{\sigma_u^{k+1} - \sigma_u^k}{\tau} + A\sigma_{k+1}^u = \dfrac{S_u^k}{\tau}, \\ \sigma_u(\cdot, 0) = \dfrac{1}{\gamma}\sigma_p^0, \\ \dfrac{\sigma_p^k - \sigma_p^{k+1}}{\tau} + A^*\sigma_p^k = -\sigma_u^{k+1} + \dfrac{S_p^{k+1}}{\tau}, \\ \sigma_p(\cdot, T) = 0. \end{cases} \tag{3.112}$$

*Similarly, to solve (*3.112*), we do the following manipulations: solving an auxiliary equation*

$$\begin{cases} \dfrac{z^n - z^{n+1}}{\tau} + A^* z^n = \dfrac{S_p^{n+1}}{\tau}, \\ z(T) = 0, \end{cases} \tag{3.113}$$

*to first find out* $z(0)$*, then we solve the minimization problem*

$$F(u_0) = \frac{1}{2}\tau \sum_{k=1}^{N} \|\sigma_u^k\|_{\mathcal{H}}^2 + \frac{\gamma}{2}\|\sigma_{u,0} - \frac{1}{\gamma}z(0)\|_H^2 \tag{3.114}$$

*subject to*

$$\begin{cases} \dfrac{\sigma_u^{k+1} - \sigma_u^k}{\tau} + A\sigma_u^{k+1} = \dfrac{S_u^k}{\tau}, \\ \sigma_u(0) = \sigma_u^0. \end{cases} \tag{3.115}$$

*Again, problem (3.113)-(3.115) is equivalent to (*3.112*).*

*Step 3 (Update):*

$$U^k = \sigma_u^k + u(t_k; P^k, U^{k-1}),$$

$$P^k = \sigma_p^k + p(t_k; P^{k+1}, U^k).$$

*Step 4 (Stop Criteria): Running Step 1 and Step 3 until a pre-defined stop tolerance* $\epsilon$ *is greater than the quantity max*$\{\|\sigma_u^k\|_{\mathcal{H}}, \|\sigma_p^k\|_{\mathcal{H}}\}$ *.*

# 4. DATA ASSIMILATION FOR PARABOLIC INTERFACE EQUATION

## 4.1. BACKGROUND FOR SECOND ORDER PARABOLIC INTERFACE EQUATION

Parabolic interface equations model physical or engineering problems when two or more distinct materials or fluids with different conductivities or diffusions are involved. Unlike a normal parabolic equation, many important features, such as the lower global regularity, interface jump conditions, and discontinuous coefficients, need to be addressed with more considerations both theoretically and numerically, see, e.g., [38, 39, 65–67].

Over the past few decades, a vast amount of literature employing variational methods has been contributed to investigate the data assimilation problem for parabolic equations. In [33], J. L. Lions provided an elementary introduction of the adjoint method and dual method to recover parameters for parabolic partial differential equations. Motivated by this approach, researchers afterwards employed similar thoughts on the initial recovery of parabolic types of dynamics systems. In [34], efficient numerical methods were developed to attain the optimal initial condition of the heat equation. In [37], a forward approach to reconstruct the initial state was presented for the convection-diffusion equation and a practical algorithm is established. Moreover, the nonlinear parabolic equations, such as in water movement and in radiative heat transfer problems, were studied in [35, 36] by reducing nonlinearity. However, to our current knowledge, few studies have investigated data assimilation for parabolic interface equations. The main interests of this section is to investigate the VDA problem for the parabolic interface equation.

We consider the following second order parabolic interface equation:

$$\begin{cases} u_t - \nabla \cdot \beta(x, y)\nabla u = f, \quad \text{in } \Omega \times (0, T], \\ u(\cdot, 0) = u_0, \quad \text{in } \Omega, \\ u = 0, \quad \text{on } \partial\Omega \times (0, T], \end{cases} \tag{4.1}$$

together with the jump interface condition,

$$[u]|_\Gamma = 0, \quad [\beta(x,y)\frac{\partial u}{\partial \vec{n}}]|_\Gamma = 0. \tag{4.2}$$

Here $\Omega \subset \mathbb{R}^2$ is an open bounded domain, the curve $\Gamma$ is a smooth interface that separates $\Omega$ into two subdomains $\Omega^+$ and $\Omega^-$ such that $\Omega = \Omega^+ \cup \Omega^- \cup \Gamma$, $[u]|_\Gamma = u^+|_\Gamma - u^-|_\Gamma$ is the jump of function $u$ across the interface $\Gamma$, where $u^+ = u|_{\Omega^+}$ and $u^- = u|_{\Omega^-}$, $\vec{n}$ is the unit normal vector along interface $\Gamma$ pointing to $\Omega^-$, $\frac{\partial u}{\partial \vec{n}}$ is the normal derivative of $u$, and $\beta(x,y)$ is assumed to be a positive piecewise constant function

$$\beta(x,y) = \begin{cases} \beta^+ & \text{if } (x,y) \in \Omega^+, \\ \beta^- & \text{if } (x,y) \in \Omega^-, \end{cases}$$

and the source term $f$ is given discontinuously as

$$f(x,y,t) = \begin{cases} f^+ & \text{if } (x,y) \in \Omega^+, \\ f^- & \text{if } (x,y) \in \Omega^-. \end{cases}$$

We now introduce necessary preliminaries for the discussion of the data assimilation problem concerning equations (4.1)-(4.2). Let $\|\cdot\|_0$ denote the $L^2$-norm with the usual $L^2$ inner product $(\cdot,\cdot)$, $\|\cdot\|_\infty$ denote the $L^\infty$-norm, and $\|\cdot\|_m$ denote the standard norm in the Sobolev space $W^{m,2}(\Omega)$, which is also written as $H^m(\Omega)$. For the temporal-spatial function spaces over $(0,T)\times\Omega$, we define, for $1 \le p < \infty$,

$$W^{m,p}(0,T;\mathcal{B}) = \left\{u(t) \in \mathcal{B} \text{ for a.e } t \in (0,T) \text{ and } \sum_{j=0}^{m}\int_0^T \|u^{(j)}(t)\|_{\mathcal{B}}^p dt < \infty\right\}$$

and for $p = \infty$

$$W^{m,\infty}(0,T;\mathcal{B}) = \Big\{u(t) \in \mathcal{B} \text{ for a.e } t \in (0,T) \text{ and } \max_{0\le j\le m}(ess \sup_{0\le t\le T} \|u^{(j)}(t)\|_{\mathcal{B}}) < \infty\Big\}$$

which are equipped with corresponding norms

$$\|u\|_{W^{m,p}(0,T;\mathcal{B})} = \Big(\sum_{j=0}^{m}\int_0^T \|u^{(j)}(t)\|_{\mathcal{B}}^p dt\Big)^{\frac{1}{p}},$$

$$\|u\|_{W^{m,\infty}(0,T;\mathcal{B})} = \max_{0\le j\le m}(ess \sup_{0\le t\le T} \|u^{(j)}(t)\|_{\mathcal{B}}),$$

where $\mathcal{B}$ is a general Banach space. As usual, we let $L^p(0,T;\mathcal{B}) = W^{0,p}(0,T;\mathcal{B})$ and $H^m(0,T;\mathcal{B}) = W^{m,2}(0,T;\mathcal{B})$.

We shall also need the following spaces:

$$X = H^1(\Omega) \cap H^2(\Omega^+) \cap H^2(\Omega^-),$$

$$Y = L^2(\Omega) \cap H^1(\Omega^+) \cap H^1(\Omega^-),$$

equipped with norms

$$\|u\|_X = \|u\|_{H^1(\Omega)} + \|u\|_{H^2(\Omega^+)} + \|u\|_{H^2(\Omega^-)},$$

$$\|u\|_Y = \|u\|_{L^2(\Omega)} + \|u\|_{H^1(\Omega^+)} + \|u\|_{H^1(\Omega^-)}.$$

We write $Y(0,T) = L^2(0,T;X) \cap H^1(0,T;Y)$.

To introduce a weak form of the interface problem (4.1)-(4.2), we define the continuous bilinear form $a(\cdot,\cdot)$: $H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ and the associated operator $A$ : $H_0^1(\Omega) \to H^{-1}(\Omega)$ as follows:

$$
\begin{aligned}
a(u,v) &= \int_{\Omega} \beta(x,y)\nabla u \cdot \nabla v dxdy \\
&= \int_{\Omega^+} \beta^+ \nabla u \cdot \nabla v dxdy + \int_{\Omega^-} \beta^- \nabla u \cdot \nabla v dxdy, \\
a(u,v) &= \langle Au, v\rangle,
\end{aligned}
$$

where $\langle\cdot,\cdot\rangle$ again defines the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

Setting $W(0,T) = L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega))$, the weak formulation is derived as follows:

$$
\begin{aligned}
&\text{Given } f \in L^2(0,T;H^{-1}(\Omega)), \text{ find } u \in W(0,T) \text{ satisfying} \\
&\langle \frac{\partial u}{\partial t}, v\rangle + a(u,v) = \langle f, v\rangle, \quad \forall v \in H_0^1(\Omega), \qquad (4.3) \\
&u(\cdot,0) = u_0 \quad \text{at } t = 0.
\end{aligned}
$$

Note that (4.3) can be expressed in the form:

$$
\frac{\partial u}{\partial t} + Au - f = 0.
$$

Throughout this Section, $C$ is a generic positive constant that is independent of the mesh parameter $h$ and the time step $\tau$ and is not necessarily the same at each occurrence.

## 4.2. MODELING THE VARIATIONAL DATA ASSIMILATION

**4.2.1. Mathematics Formulation and Well-posedness.** Let $U_{ad}$ denote the admissible solutions set that could be either $L^2(\Omega)$ or a closed convex subset of $L^2(\Omega)$. Given $T > 0$, $\gamma > 0$, and a distributed observation $\hat{u}$, the variational data assimilation for the second order parabolic interface equation is given by

$$\min_{u_0 \in U_{ad}} J(u_0) = \frac{1}{2}\int_0^T \|\hat{u} - u(u_0)\|_0^2 dt + \frac{\gamma}{2}\|u_0\|_0^2, \tag{4.4}$$

subject to

$$\begin{cases} u_t - \nabla \cdot \beta(x,y)\nabla u = f, \quad \text{in } \Omega \times (0,T], \\ u(\cdot,0) = u_0, \quad \text{in } \Omega, \\ u = 0, \quad \text{on } \partial\Omega \times [0,T], \\ [u]|_\Gamma = 0, \quad \text{on } \Gamma \times (0,T], \\ [\beta(x,y)\frac{\partial u}{\partial \vec{n}}]|_\Gamma = 0, \quad \text{on } \Gamma \times (0,T], \end{cases} \tag{4.5}$$

where the mapping $u(u_0) : L^2(\Omega) \to W(0,T)$ is defined as the solution of (4.5) with the initial value $u_0$. The mapping $u(u_0)$ is continuous and uniquely defined [39].

The minimization of $\int_0^T \frac{1}{2}\|\hat{u} - u(u_0)\|_0^2 dt$ is the primary goal, which is to drive the state variable $u(u_0)$ close to the distributed observations $\hat{u}$ over $(0,T) \times \Omega$ via adjusting the initial condition $u_0$. The second term $\frac{\gamma}{2}\|u_0\|_0^2$ is a Tikhonov regularization and plays a key role in guaranteeing the existence of the minimizer as well as the stability of the data assimilation problem. In particular, $\gamma$ is a significantly characterized parameter. According to the reliability of observations, it can be used to measure the cost acted on the initial condition and balance the minimizing distribution in the cost functional.

For the minimization problem (4.4)-(4.5), provided that $\partial\Omega$ and $\Gamma$ are smooth enough and $f, \hat{u} \in L^2(0,T; L^2(\Omega))$, we have the following existence and uniqueness result.

**Theorem 11** *There exists a unique solution $u_0 \in U_{ad}$ for the data assimilation problem (4.4)-(4.5). Furthermore, the solution $u_0$ can be characterized by*

$$\begin{aligned} \langle J'(u_0), v - u_0 \rangle = &\int_0^T \int_\Omega (u(u_0) - \hat{u})(u(v) - u(u_0)) dxdydt \\ &+ \gamma \int_\Omega u_0(v - u_0) dxdy \geq 0 \quad \forall v \in U_{ad}. \end{aligned} \tag{4.6}$$

**Theorem 12** *The solution of problem* (4.4)*-*(4.5) *continuously depends on the observational data* $\hat{u}$ *and the parameter* $\gamma$.

Moreover, small $\gamma$ will degrade the stability of the data assimilation problem.

**Remark 10** *Theorem 11 and Theorem 12 can be verified immediately by using Theorem 8 in the Section 2 and Theorem 9 in the Section 3, since the problem* (4.4)*-*(4.5) *is a special case of the optimization problem considered there.*

**4.2.2. Derivation of the Optimality System.** Based on the wellposedness results, we can now determine the optimal initial condition. Using the gradient information from the Gâteaux derivative and the adjoint method, we derive the optimality system characterized by the first order necessary condition.

Computing the first order Gâteaux derivative of the cost functional (4.4) at any given direction $h \in L^2(\Omega)$ gives us

$$\frac{DJ(u_0)}{Du_0} \cdot h = -\int_0^T \int_\Omega (\hat{u} - u)(\frac{Du(u_0)}{Du_0} h) dxdydt + \int_\Omega \gamma u_0 h dxdy. \tag{4.7}$$

Equation (4.7) should vanish if $u_0$ is the minimizer in $U_{ad} = L^2(\Omega)$, thereby allowing us to solve for $u_0$. However, $\int_0^T \int_\Omega (\hat{u} - u(u_0))(\frac{Du(u_0)}{Du_0} h) dxdydt$ is an intractable term to be evaluated. Hence, expressing $u_0$ explicitly is difficult. To address this difficulty, the adjoint method is used.

**Lemma 3** *The mapping* $u(u_0) : L^2(\Omega) \to W(0,T)$ *defined as the solution of* (4.5) *with initial condition* $u_0$ *has a Gâteaux derivative* $\frac{Du(u_0)}{Du_0} h$ *in every direction* $h \in L^2(\Omega)$. *Moreover,* $\frac{Du(u_0)}{Du_0} h$ *solves the second order parabolic interface equation with zero force and initial condition h.*

**Remark 11** *Lemma* 3 *holds because of the linearity of the constraint equation.*

Oriented by the Gâteaux derivative, Lemma 3 gives us the insight to represent

$$\int_0^T \int_\Omega (\hat{u} - u(u_0))(\frac{Du(u_0)}{Du_0} h) dxdydt$$

based on the adjoint method, which can be seen in the following theorem.

**Theorem 13** *Given an observational function $\hat{u} \in L^2(0, T; L^2(\Omega))$ and if $u_0 \in L^2(\Omega)$ is the optimal solution for (4.4)-(4.5), then $u_0$ is obtained as*

$$u_0 = \frac{1}{\gamma} u^*(\cdot, 0) \tag{4.8}$$

*where $u^*$ is the solution to the associated adjoint equation,*

$$\begin{cases} -u_t^* - \nabla \cdot \beta(x, y) \nabla u^* = \hat{u} - u, \quad \text{in } \Omega \times [0, T), \\ u^*(\cdot, T) = 0, \quad \text{in } \Omega, \\ u^* = 0, \quad \text{on } \partial\Omega \times [0, T), \\ [u^*]|_\Gamma = 0, \quad \text{on } \Gamma \times [0, T), \\ [\beta(x, y) \frac{\partial u^*}{\partial \vec{n}}]|_\Gamma = 0, \quad \text{on } \Gamma \times [0, T). \end{cases} \tag{4.9}$$

**Proof:** Considering the formula (4.7), the main purpose of this proof is to find an easily evaluated adjoint representation for

$$\int_0^T \int_\Omega (\hat{u} - u)(\frac{Du(u_0)}{Du_0} h) dxdydt. \tag{4.10}$$

Lemma 3 reminds us to make use of the equation,

$$\begin{cases} u_t - \nabla \cdot \beta(x,y)\nabla u = 0, \quad \text{in } \Omega \times (0,T], \\ u(\cdot,0) = h, \quad \text{in } \Omega, \\ u = 0, \quad \text{on } \partial\Omega \times (0,T], \\ [u]|_\Gamma = 0, \quad \text{on } \Gamma \times (0,T], \\ [\beta(x,y)\dfrac{\partial u}{\partial \vec{n}}]|_\Gamma = 0, \quad \text{on } \Gamma \times (0,T]. \end{cases} \tag{4.11}$$

By taking the $L^2(0,T;L^2(\Omega))$ inner product on the first equation in (4.11) with $u^* \in W(0,T)$, we obtain

$$\int_0^T \int_\Omega \frac{\partial u}{\partial t} u^* dxdydt + \int_0^T \int_{\Omega_+} \beta^+ \nabla u \nabla u^* dxdydt + \int_0^T \int_{\Omega^-} \beta^- \nabla u \nabla u^* dxdydt = 0.$$

Integrating by parts in time for the first term leads to

$$\begin{aligned} &\int_\Omega uu^*|_0^T dxdy - \int_0^T \int_\Omega \frac{\partial u^*}{\partial t} udxdydt + \int_0^T \int_{\Omega_+} \beta^+ \nabla u \nabla u^* dxdydt \\ &+ \int_0^T \int_{\Omega^-} \beta^- \nabla u \nabla u^* dxdydt = 0. \end{aligned} \tag{4.12}$$

By imposing

$$\begin{aligned} &- \int_0^T \int_\Omega \frac{\partial u^*}{\partial t} udxdydt + \int_0^T \int_{\Omega_+} \beta^+ \nabla u \nabla u^* dxdydt \\ &+ \int_0^T \int_{\Omega^-} \beta^- \nabla u \nabla u^* dxdydt = \int_0^T \int_\Omega (\hat{u} - u) udxdydt, \\ &u^*(\cdot,T) = 0, \end{aligned} \tag{4.13}$$

(4.12) can be simplified as

$$- \int_\Omega uu^*(\cdot,0)dxdy = \int_0^T \int_\Omega (\hat{u} - u) udxdydt. \tag{4.14}$$

Taking integration by parts in space $\Omega = \Omega^+ \cup \Omega^-$ on (4.13) gives the adjoint equation (4.9), together with $u^*(\cdot, T) = 0$. We also know that $u = \frac{Du(u_0)}{Du_0} h$ and $u(\cdot, 0) = h$. The proof is completed by substituting (4.14) into (4.7) and letting $\frac{DJ(u_0)}{Du_0} \cdot h$ vanish, i.e.,

$$\frac{DJ(u_0)}{Du_0} \cdot h = -\int_\Omega u^*(\cdot, 0) h dx dy + \int_\Omega \gamma u_0 h dx dy = 0,$$
$$\int_\Omega (\gamma u_0 - u^*(\cdot, 0)) h dx dy = 0 \quad \forall h \in L^2(\Omega).$$

By the completeness of the $L^2(\Omega)$ space, we have $u_0 = \frac{1}{\gamma} u^*(\cdot, 0)$.

Since the minimization problem (4.4)-(4.5) is strictly convex, the first order necessary condition in Theorem 3.4 is also sufficient. In order to attain the optimal solution, we need to solve the following equation systems (the optimality system):

*the forward state equation*

$$\begin{cases} u_t - \nabla \cdot \beta(x, y) \nabla u = f, \quad \text{in } \Omega \times (0, T], \\ u(\cdot, 0) = u_0, \quad \text{in } \Omega, \\ u = 0, \quad \text{on } \partial\Omega \times (0, T], \\ [u]|_\Gamma = 0, \quad \text{on } \Gamma \times (0, T], \\ [\beta(x, y) \dfrac{\partial u}{\partial \vec{n}}]|_\Gamma = 0, \quad \text{on } \Gamma \times (0, T], \end{cases} \tag{4.15}$$

*the backward adjoint equation*

$$\begin{cases} -u_t^* - \nabla \cdot \beta(x, y) \nabla u^* = \hat{u} - u, \quad \text{in } \Omega \times [0, T), \\ u^*(\cdot, T) = 0, \quad \text{in } \Omega, \\ u^* = 0, \quad \text{on } \partial\Omega \times [0, T), \\ [u^*]|_\Gamma = 0, \quad \text{on } \Gamma \times [0, T), \\ [\beta(x, y) \dfrac{\partial u^*}{\partial \vec{n}}]|_\Gamma = 0, \quad \text{on } \Gamma \times [0, T), \end{cases} \tag{4.16}$$

and

$$u_0 = \frac{1}{\gamma} u^*(\cdot, 0), \tag{4.17}$$

where $u_0 \in L^2(\Omega)$, $u \in W(0,T)$, $u^* \in W(0,T)$, $\hat{u} \in L^2(0,T; L^2(\Omega))$, $f \in L^2(0,T; L^2(\Omega))$.

**Remark 12** *If the admissible set is considered as $U_{ad} = \{u_0 \in L^2(\Omega) : a \leq u_0 \leq b\}$, then the optimal solution is an orthogonal projection of $\frac{1}{\gamma}u^*(\cdot, 0)$ onto $U_{ad}$, i.e., $u_0 = \max\{a, \min\{b, \frac{1}{\gamma}u^*(\cdot, 0)\}\}$.*

## 4.3. A FINITE ELEMENT APPROXIMATION AND CONVERGENCE ANALYSIS

**4.3.1. Numerical Approximation.** To numerically compute the solution discussed in section 4.2, in this section we present a fully discrete approximation to the data assimilation problem (4.4)-(4.5) that uses a piecewise linear finite element method in space and the backward Euler scheme in time.

For the spatial discretization, we first approximate the smooth interface $\Gamma$ and boundary $\partial\Omega$ with line segments, the union of such line segments forms an approximate interface $\Gamma_h$ and boundary $\partial\Omega_h$. The domain circumscribed by $\partial\Omega_h$ is denoted with $\Omega_h$, which is an approximation of $\Omega$. $\Gamma_h$ divides $\Omega_h$ into two subdomains $\Omega_h^+$ and $\Omega_h^-$, which form an approximation of $\Omega^+$ and $\Omega^-$ respectively.

Let $\mathcal{T}_h^+$ denote a family of triangulation of $\Omega_h^+$ and $\mathcal{T}_h^-$ denote a family of triangulation of $\Omega_h^-$ such that

$$\mathcal{T}_h = \mathcal{T}_h^+ \cup \mathcal{T}_h^-.$$

We need the vertices on $\partial\Omega_h$ or $\Gamma_h$ of a triangle $\tau_h \in \mathcal{T}_h$ to be vertices of $\partial\Omega_h$ or $\Gamma_h$ respectively. We also assume the triangulation $\mathcal{T}_h$ satisfies the usual sort of quasi-uniformity condition.

Associated with $\mathcal{T}_h$ is the finite element space $V_h$ =span$\{\phi_i\}_{i=1}^{i=N_b}$, where $\phi_i$ is piecewise linear polynomials and $N_b$ is the number of finite element nodes. The admissable set of discrete solutions is denoted by $U_h = V_h \cap U_{ad}$.

For the time discretization we uniformly construct a time grid $0 = t_0 < t_1 < t_2 < t_3 ... < t_n ... < t_N = T$ with time step $\tau = \frac{T}{N}$. Let $I_n = (t_{n-1}, t_n]$ denote the $n$th sub-interval. We use the finite-dimensional space

$$V_{\tau,h} = \{v : [0,T] \to V_h : v|_{I_n} \in V_h \quad \text{is constant in time}\}.$$

Let $v^n$ be the value of $v \in V_{\tau,h}$ at $t_n$ and $V_{\tau,h}^n$ be the restriction to $I_n$ of the functions in $V_{\tau,h}$.

Given specific $h, \tau$ and $\gamma > 0$, the fully discrete approximation of problem (4.4)-(4.5) is stated as

$$\min_{u_{0,h} \in U_h} J_h(u_{0,h}) \tag{4.18}$$

subject to

$$\begin{cases} (\dfrac{u_h^{n+1} - u_h^n}{\tau}, v_h) + a(u_h^{n+1}, v_h) = (f_{n+1}, v_h), \quad \forall v_h \in V_h, \\ u_h^0 = u_{0,h}, \end{cases} \tag{4.19}$$

where

$$J_h(u_{0,h}) = \frac{1}{2}\tau \sum_{n=1}^{N} \|\hat{u}^n - u_h^n\|_0^2 + \frac{\gamma}{2}\|u_{0,h}\|_0^2. \tag{4.20}$$

Similar to the proof for the wellposedness of the continuous data assimilation problem, one can prove the wellposedness of the fully discrete data assimilation problem (4.18)-(4.20).

**Theorem 14** *Given $\tau = \frac{T}{N}$ and mesh size $h$, for each fixed regularization parameter $\gamma$, there exists a unique optimal solution $u_{0,h} \in U_h$ such that the cost functional* (4.20) *is minimized.*

**Theorem 15** *The solution of problem* (4.18)*-*(4.20) *continuously depends on the observational data* $\hat{u}$ *and the parameter* $\gamma$.

Note that small $\gamma$ will also reduce the stability of the discrete data assimilation problem.

In order to derive the discrete optimality system and solve for $u_{0,h}$, we apply an approach different from that of the continuous problem, i.e., a Lagrange multiplier rule. We form the Lagrange functional as

$$\begin{aligned}\mathcal{L}(\bar{u}_h, u_{0,h}, \bar{u}_h^*) &= \frac{1}{2}\tau \sum_{n=1}^{N} \|\hat{u}^n - u_h^n\|_0^2 + \frac{\gamma}{2}\|u_{0,h}\|_0^2 \\ &+ \tau \sum_{n=0}^{N-1} \langle \frac{u_h^{n+1} - u_h^n}{\tau} + Au_h^{n+1} - f_{n+1}, u_h^{*n}\rangle + (u_h^0 - u_{0,h}, u_h^{*0}),\end{aligned} \tag{4.21}$$

where $\bar{u}_h = (u_h^1, u_h^2, u_h^3, \ldots.., u_h^N)$ and $\bar{u}_h^* = (u_h^{*0}, u_h^{*1}, u_h^{*2}, u_h^{*3}, \ldots.., u_h^{*N-1})$. Going through the almost identical steps (3.24)-(3.28) in section 3.3 and using the self-adjointness of the operator $A$ in sense of $\langle Au, v\rangle_{H^{-1}\times H_0^1} = \langle Av, u\rangle_{H^{-1}\times H_0^1}$, we end up with the discrete optimality system,

$$\begin{cases} \dfrac{u_h^{n+1} - u_h^n}{\tau} + Au_h^{n+1} = f_{n+1}, \\ u_h^0 = u_{0,h}, \\ -\dfrac{u_h^{*n+1} - u_h^{*n}}{\tau} + Au_h^{*n} = \hat{u}^{n+1} - u_h^{n+1}, \\ u_h^{*N} = 0, \\ u_{0,h} = \dfrac{1}{\gamma}u_h^{*0} \end{cases} \tag{4.22}$$

for $n = 0, 1, 2, 3\ldots..N-1$.

**4.3.2. Finite Element Convergence Analysis.** We shall expect the discrete solution in (4.22) to converge to the solution of (4.15)-(4.17). That is, given fixed $\gamma$, $u_{0,h} \to u_0$, $u_h \to u$ and $u_h^* \to u^*$ should be attained while the time step $\tau$ and finite element mesh size $h$ diminish.

**Theorem 16** *For each fixed regularization parameter $\gamma$, let $\{u_{0,h}\}_{h>0}$ be the corresponding sequence of minimizer of the discrete data assimilation problem (4.18)-(4.20). Then $\{u_{0,h}\}_{h>0}$ converges to the optimal solution $u_0$ of the continuous problem (4.4)-(4.5).*

**Proof:** It is not difficult to see $J_h(u_{0,h}) \le C$ for some constant $C$ independent of $h$ and $\tau$. Then the coercevity of $J_h(u_{0,h})$ implies the boundedness of $\{u_{0,h}\}$ in $L^2(\Omega)$. Hence we can extract a subsequence $\{u_{0,h'}\}$ from $\{u_{0,h}\}$ such that $\{u_{0,h'}\}$ weakly converges to $\mu^*$ in $L^2(\Omega)$. We conclude furthermore

$$\lim_{h',\tau\to 0} \frac{1}{2}\tau \sum_{n=1}^{N} \|\hat{u}^n - u_h^n(u_{0,h'})\|_0^2 \to \frac{1}{2}\int_0^T \|\hat{u} - u(\mu^*)\|_0^2 dt.$$

Thus, for $\forall v \in U_{ad}$, by the weakly lower semicontinuity we deduce

$$\begin{aligned} J(\mu^*) &\le \liminf_{h',\tau\to 0} \frac{1}{2}\tau \sum_{n=1}^{N} \|\hat{u}^n - u_{h'}^n(u_{0,h'})\|_0^2 + \frac{\gamma}{2} \liminf_{h',\tau\to 0} \|u_{0,h'}\|_0^2 \\ &\le \liminf_{h',\tau\to 0} J_{h'}(u_{0,h'}) \le \liminf_{h',\tau\to 0} J_{h'}(\pi_{h'}(v)) \\ &= \frac{1}{2}\int_0^T \|\hat{u} - u(v)\|_0^2 dt + \frac{\gamma}{2}\|v\|_0^2 \\ &= J(v) \end{aligned} \tag{4.23}$$

where $\pi_h$ is the $L^2$ projection operator from $U_{ad}$ to $U_h$.

Then (4.23) and the uniqueness result in Theorem 18 imply $\mu^*$ is the optimal solution of the problem (4.4)-(4.5) and thus the theorem is proved.

Besides a general convergence result in Theorem 21, under appropriate assumptions, we can obtain the optimal finite element convergence rate for $u_0 - u_{0,h}$, $u - u_h$, $u^* - u_h^*$.

Compared with the classical FEM approximation analysis, the difficulties in our case lie in the undetermined initial condition from the forward state equation and the Gelarkin orthogonality we miss on the backward adjoint equations, both of which would lead to the invalidity of the classical analysis framework. In order to overcome these difficulties, we introduce the following auxiliary equations to bridge the analysis in the data assimilation problem and the classical FEM approximation results:

$$\begin{cases} \dfrac{\partial u(u_{0,h})}{\partial t} - \nabla \cdot (\beta(x,y)\nabla u) = f, \quad \text{in } \Omega \times (0,T], \\ u(u_{0,h})(\cdot,0) = u_{0,h}, \quad \text{in } \Omega, \\ u(u_{0,h}) = 0, \quad \text{on } \partial\Omega \times (0,T], \\ [u(u_{0,h})]|_\Gamma = 0, \quad \text{on } \Gamma \times (0,T], \\ [\beta(x,y)\dfrac{\partial u(u_{0,h})}{\partial \vec{n}}]|_\Gamma = 0, \quad \text{on } \Gamma \times (0,T], \end{cases} \tag{4.24}$$

$$\begin{cases} -\dfrac{\partial u^*(u_h)}{\partial t} - \nabla \cdot (\beta(x,y)\nabla u^*(u_h)) = \hat{u} - u_h, \quad \text{in } \Omega \times [0,T), \\ u^*(u_h)(\cdot,T) = 0, \quad \text{in } \Omega, \\ u^*(u_h) = 0, \quad \text{on } \partial\Omega \times [0,T) \\ [u^*(u_h)]|_\Gamma = 0, \quad \text{on } \Gamma \times [0,T), \\ [\beta(x,y)\dfrac{\partial u^*(u_h)}{\partial \vec{n}}]|_\Gamma = 0, \quad \text{on } \Gamma \times [0,T). \end{cases} \tag{4.25}$$

$$\begin{cases} -\dfrac{\partial u^*(u_{0,h})}{\partial t} - \nabla \cdot (\beta(x,y)\nabla u^*(u_{0,h})) = \hat{u} - u(u_{0,h}), \ \text{in } \Omega \times [0,T), \\ u^*(u_{0,h})(\cdot,T) = 0, \quad \text{in } \Omega, \\ u^*(u_{0,h}) = 0, \quad \text{on } \partial\Omega \times [0,T), \\ [u^*(u_{0,h})]|_\Gamma = 0, \quad \text{on } \Gamma \times [0,T), \\ [\beta(x,y)\dfrac{\partial u^*(u_{0,h})}{\partial \vec{n}}]|_\Gamma = 0, \quad \text{on } \Gamma \times [0,T), \end{cases} \tag{4.26}$$

The motivation of the constructions for (4.24) and (4.25) is to remove the uncertainties on the initial condition and source term. We then convert the target error estimate into an intermediate error that can be bounded finally by using (4.26) and the additional equalities $u_0 = \frac{1}{\gamma}u^*(\cdot,0)$ and $u_{0,h} = \frac{1}{\gamma}u_h^{*0}$ in the optimality systems. The details are demonstrated in the following theorem and lemmas.

**Theorem 17** *Let* $(u, u^*, u_0) \in W(0,T)\times W(0,T)\times U_{ad}$ *and* $(u_h, u_h^*, u_{0,h}) \in V_{\tau,h}\times V_{\tau,h}\times U_h$ *be solutions of the continuous optimality system* (4.15) − (4.17) *and discrete optimality system* (4.22) *respectively. Assuming the solutions are smooth enough, then the following error estimate holds*

$$\begin{aligned} & \|u_0 - u_{0,h}\|_0 + \|u - u_h\|_{L^2(0,T;L^2(\Omega))} + \|u^* - u_h^*\|_{L^2(0,T;L^2(\Omega))} \\ \leq \; & C(\gamma)(h^2|\log h| + \tau). \end{aligned} \tag{4.27}$$

This is the major theorem we are going to show in this section. To prove it, some useful inequalities need to be derived from the auxiliary equations first.

**Lemma 4** *Let* $(u(u_{0,h}), u^*(u_{0,h}), u^*(u_h)) \in W(0,T)\times W(0,T)\times W(0,T)$ *be solutions for equations* (4.24)*,* (4.26)*, and* (4.25) *respectively. Let* $(u, u^*, u_0) \in W(0,T)\times W(0,T)\times U_{ad}$ *be the solution of* (4.15)*-*(4.17) *and let* $(u_h, u_h^*, u_{0,h}) \in V_{\tau,h}\times V_{\tau,h}\times U_h$ *be the solution of* (4.22)*. Then we have the following inequalities*

$$\|u - u(u_{0,h})\|_{L^2(0,T;L^2(\Omega))} \leq C\|u_0 - u_{0,h}\|_0, \tag{4.28}$$

$$\|u^*(u_h) - u^*(u_{0,h})\|_{L^2(0,T;L^2(\Omega))} \leq C\|u(u_{0,h}) - u_h\|_{L^2(0,T;L^2(\Omega))}, \tag{4.29}$$

$$\sup_{0\leq t\leq T} \|u^*(u_h) - u^*(u_{0,h})\| \leq C\|u(u_{0,h}) - u_h\|_{L^2(0,T;L^2(\Omega))}, \tag{4.30}$$

$$\|u^* - u^*(u_{0,h})\|_{L^2(0,T;L^2(\Omega))} \leq C\|u_0 - u_{0,h}\|_0. \tag{4.31}$$

**Proof:** By subtractions between the equations (4.15) and (4.24), (4.16) and (4.26), and (4.25) and (4.26), respectively, we obtain the following equations

$$\begin{cases} \dfrac{\partial(u-u(u_{0,h}))}{\partial t} - \nabla\cdot(\beta(x,y)\nabla(u-u(u_{0,h}))) = 0, \quad \text{in } \Omega\times(0,T], \\ (u-u(u_{0,h}))(\cdot,0) = u_0 - u_{0,h}, \quad \text{in } \Omega, \\ u-u(u_{0,h}) = 0, \quad \text{on } \partial\Omega\times(0,T], \\ [u-u(u_{0,h})]|_\Gamma = 0, \quad \text{on } \Gamma\times(0,T], \\ [\beta(x,y)\dfrac{\partial(u-u(u_{0,h}))}{\partial\vec{n}}]|_\Gamma = 0, \quad \text{on } \Gamma\times(0,T], \end{cases} \tag{4.32}$$

$$\begin{cases} -\dfrac{\partial(u^*-u^*(u_{0,h}))}{\partial t} - \nabla\cdot(\beta(x,y)\nabla(u^*-u^*(u_{0,h}))) = \\ u(u_{0,h}) - u, \quad \text{in } \Omega\times[0,T), \\ (u^*-u^*(u_{0,h}))(\cdot,T) = 0, \quad \text{in } \Omega, \\ (u^*-u^*(u_{0,h})) = 0, \quad \text{on } \partial\Omega\times[0,T), \\ [(u^*-u^*(u_{0,h}))]|_\Gamma = 0, \quad \text{on } \Gamma\times[0,T), \\ [\beta(x,y)\dfrac{\partial(u^*-u^*(u_{0,h}))}{\partial\vec{n}}]|_\Gamma = 0, \quad \text{on } \Gamma\times[0,T), \end{cases} \tag{4.33}$$

$$\begin{cases} -\dfrac{\partial(u^*(u_h)-u^*(u_{0,h}))}{\partial t} - \nabla\cdot(\beta(x,y)\nabla(u^*(u_h)-u^*(u_{0,h}))) = \\ u(u_{0,h}) - u_h, \quad \text{in } \Omega\times[0,T), \\ (u^*(u_h)-u^*(u_{0,h}))(\cdot,T) = 0, \quad \text{in } \Omega, \\ (u^*(u_h)-u^*(u_{0,h})) = 0, \quad \text{on } \partial\Omega\times[0,T), \\ [(u^*(u_h)-u^*(u_{0,h}))]|_\Gamma = 0, \quad \text{on } \Gamma\times[0,T), \\ [\beta(x,y)\dfrac{\partial(u^*(u_h)-u^*(u_{0,h}))}{\partial\vec{n}}]|_\Gamma = 0, \quad \text{on } \Gamma\times[0,T). \end{cases} \tag{4.34}$$

By taking the $L^2(0,t;L^2(\Omega))$ inner product with $u-u(u_{0,h})$ for the first equation in (4.32), we obtain

$$\int_0^t \frac{1}{2}\frac{d\|u-u(u_{0,h})\|_0^2}{ds}ds+\int_0^t(\beta(x,y)\nabla(u-u(u_{0,h})),\nabla(u-u(u_{0,h})))ds=0. \tag{4.35}$$

Equation (4.35) infers

$$\begin{aligned}&\|(u-u(u_{0,h}))(\cdot,t)\|_0^2+2\int_0^t(\beta(x,y)\nabla(u-u(u_{0,h})),\nabla(u-u(u_{0,h})))ds\\&=\|u_0-u_{0,h}\|_0^2.\end{aligned}$$

Thus, we have the inequality

$$\|(u-u(u_{0,h}))(\cdot,t)\|_0\le\|u_0-u_{0,h}\|_0 \quad \text{for } 0\le t\le T. \tag{4.36}$$

Integrating with respect to $t$ on both sides from 0 to $T$ for (4.36) gives us

$$\|u-u(u_{0,h})\|_{L^2(0,T;L^2(\Omega))}\le C\|u_0-u_{0,h}\|. \tag{4.37}$$

Taking the suprenum with respect to $t$ on (4.36), we obtain

$$\sup_{0\le t\le T}\|u-u(u_{0,h})\|\le\|u_0-u_{0,h}\|_0. \tag{4.38}$$

Taking the $L^2(\Omega)$ inner product on the first equation of (4.33) with $u^*-u^*(u_{0,h})$, we have

$$\begin{aligned}&-\frac{1}{2}\frac{d\|u^*-u^*(u_{0,h})\|_0^2}{dt}+(\beta(x,y)\nabla(u^*-u^*(u_{0,h})),\nabla(u^*-u^*(u_{0,h})))\\&=(u(u_{0,h})-u,u^*-u^*(u_{0,h})).\end{aligned} \tag{4.39}$$

Applying the Cauchy-Schwarz inequality and Young's inequality on the right hand side of (4.39), we have

$$
\begin{aligned}
&-\frac{1}{2}\frac{d\|u^* - u^*(u_{0,h})\|_0^2}{dt} + (\beta(x,y)\nabla(u^* - u^*(u_{0,h})), \nabla(u^* - u^*(u_{0,h}))) \\
&\leq \frac{1}{2}\|u(u_{0,h}) - u\|_0^2 + \frac{1}{2}\|u^* - u^*(u_{0,h})\|_0^2.
\end{aligned} \tag{4.40}
$$

Using the Grönwall inequality on (4.40) leads to

$$
\begin{aligned}
&\|(u^* - u^*(u_{0,h}))(\cdot,t)\|_0^2 + C\int_t^T \|\nabla(u^* - u^*(u_{0,h}))\|_0^2 ds \\
&\leq C\int_t^T \frac{1}{2}\|u(u_{0,h}) - u\|_0^2 ds + \|(u^* - u^*(u_{0,h}))(\cdot,T)\|_0^2 \\
&= C\int_t^T \frac{1}{2}\|u(u_{0,h}) - u\|_0^2 ds.
\end{aligned} \tag{4.41}
$$

Thus, we have

$$
\|(u^* - u^*(u_{0,h}))(\cdot,t)\|_0^2 \leq C\int_0^T \frac{1}{2}\|u(u_{0,h}) - u\|_0^2 ds,
$$

which implies another two inequalities:

$$
\|u^* - u^*(u_{0,h})\|_{L^2(0,T;L^2(\Omega))} \leq C\|u(u_{0,h}) - u\|_{L^2(0,T;L^2(\Omega))}, \tag{4.42}
$$

$$
\sup_{0\leq t\leq T} \|u^* - u^*(u_{0,h})\|_0 \leq C\|u(u_{0,h}) - u\|_{L^2(0,T;L^2(\Omega))}. \tag{4.43}
$$

In addition, taking the $L^2(\Omega)$ inner product on the first equation of (4.34) with $u^*(u_h) - u^*(u_{0,h})$ and similarly applying the Cauchy-Schwarz inequality, Young's inequality, and the Grönwall inequality step by step, we have the following inequalities:

$$
\|u^*(u_h) - u^*(u_{0,h})\|_{L^2(0,T;L^2(\Omega))} \leq C\|u(u_{0,h}) - u_h\|_{L^2(0,T;L^2(\Omega))}, \tag{4.44}
$$

$$
\sup_{0\leq t\leq T} \|u^*(u_h) - u^*(u_{0,h})\|_0 \leq C\|u(u_{0,h}) - u_h\|_{L^2(0,T;L^2(\Omega))}. \tag{4.45}
$$

Finally, combining (4.42) and (4.37) leads to

$$\|u^* - u^*(u_{0,h})\|_{L^2(0,T;L^2(\Omega))} \leq C\|u_0 - u_{0,h}\|_0. \tag{4.46}$$

Now we are in position to build up connections between the inequalities derived above and classical FEM convergence results. By using the triangle inequality and (4.28) we can bound $\|u - u_h\|_{L^2(0,T;L^2(\Omega))}$ as follows:

$$\begin{aligned} \|u - u_h\|_{L^2(0,T;L^2(\Omega))} &\leq \|u - u(u_{0,h})\|_{L^2(0,T;L^2(\Omega))} + \|u(u_{0,h}) - u_h\|_{L^2(0,T;L^2(\Omega))} \\ &\leq C\|u_0 - u_{0,h}\|_0 + \|u(u_{0,h}) - u_h\|_{L^2(0,T;L^2(\Omega))}. \end{aligned} \tag{4.47}$$

From inequalities (4.29) and (4.31), $\|u^* - u_h^*\|_{L^2(0,T;L^2(\Omega))}$ can be bounded similarly as

$$\begin{aligned} &\|u^* - u_h^*\|_{L^2(0,T;L^2(\Omega))} \\ &\leq \|u^* - u^*(u_{0,h})\|_{L^2(0,T;L^2(\Omega))} + \|u^*(u_{0,h}) - u^*(u_h)\|_{L^2(0,T;L^2(\Omega))} \\ &\quad + \|u^*(u_h) - u_h^*\|_{L^2(0,T;L^2(\Omega))} \\ &\leq C\|u_0 - u_{0,h}\|_0 + C\|u(u_{0,h}) - u_h\|_{L^2(0,T;L^2(\Omega))} \\ &\quad + \|u^*(u_h) - u_h^*\|_{L^2(0,T;L^2(\Omega))}. \end{aligned} \tag{4.48}$$

Note that $u_h$ and $u_h^*$ are the classical FEM approximations of $u(u_{0,h})$ and $u^*(u_h)$, respectively. Convergence and error estimates between them are obtained directly while traditional regularities are satisfied. From inequalities (4.47) and (4.48), we see that the convergence analysis now points to the only undetermined term $\|u_0 - u_{0,h}\|_0$. Another two conditions $u_0 = \frac{1}{\gamma}u^*(\cdot, 0)$ and $u_{0,h} = \frac{1}{\gamma}u_h^{*0}$ will be used to bound $\|u_0 - u_{0,h}\|_0$.

**Lemma 5** *Under the same conditions for* $u, u^*, u_0, u_{0,h}, u^*(u_{0,h})$*, and* $u(u_{0,h})$ *as in Lemma 4, we have the following error estimate:*

$$\|u_0 - u_{0,h}\|_0 \leq \frac{1}{\gamma}\|u^*(u_{0,h})(\cdot,0) - u_h^{*0}\|_0. \tag{4.49}$$

**Proof:** Taking the $L^2(\Omega)$ norm of $u_0 - u_{0,h}$ and applying equalities $u_0 = \frac{1}{\gamma}u^*(\cdot,0)$ and $u_{0,h} = \frac{1}{\gamma}u_h^{*0}$, we have

$$\begin{aligned}
&\|u_0 - u_{0,h}\|^2 = (u_0 - u_{0,h}, u_0 - u_{0,h}) \\
&= \frac{1}{\gamma}(u^*(\cdot,0) - u_h^{*0}, u_0 - u_{0,h}) \\
&= \frac{1}{\gamma}(u^*(\cdot,0) - u^*(u_{0,h})(\cdot,0), u_0 - u_{0,h}) + \frac{1}{\gamma}(u^*(u_{0,h})(\cdot,0) - u_h^{*0}, u_0 - u_{0,h}).
\end{aligned} \tag{4.50}$$

We now use (4.32) and (4.33) from the proof in Lemma 4 and take the $L^2(0,T;L^2(\Omega))$ inner product with $u^* - u^*(u_{0,h})$ on the first equation of (5.77). This leads to

$$\begin{aligned}
&\int_0^T (\frac{\partial(u - u(u_{0,h}))}{\partial t}, u^* - u^*(u_{0,h}))dt \\
&+ \int_0^T (\beta(x,y)\nabla(u - u(u_{0,h})), \nabla(u^* - u^*(u_{0,h})))dt = 0.
\end{aligned} \tag{4.51}$$

Taking integration by parts with respect to $t$ on the first term of (4.51) gives us

$$\begin{aligned}
&((u - u(u_{0,h}))(\cdot,T), (u^* - u^*(u_{0,h}))(\cdot,T)) - ((u - u(u_{0,h}))(\cdot,0), (u^* - u^*(u_{0,h}))(\cdot,0)) \\
&- \int_0^T (\frac{\partial(u^* - u^*(u_{0,h}))}{\partial t}, u - u(u_{0,h}))dt \\
&+ \int_0^T (\beta(x,y)\nabla(u^* - u^*(u_{0,h})), \nabla(u - u(u_{0,h})))dt = 0.
\end{aligned}$$

Using (4.33) we simplify the previous equality:

$$((u - u(u_{0,h}))(\cdot,0), (u^* - u^*(u_{0,h}))(\cdot,0)) = -\int_0^T (u - u(u_{0,h}), u - u(u_{0,h}))dt.$$

Since $\int_0^T (u - u(u_{0,h}), u - u(u_{0,h}))dt$ is nonnegative, we have

$$(u^*(\cdot, 0) - u^*(u_{0,h})(\cdot, 0), u_0 - u_{0,h}) \leq 0. \tag{4.52}$$

Combining (4.52) with equality (4.50) leads to

$$\begin{aligned}\|u_0 - u_{0,h}\|_0^2 &\leq \frac{1}{\gamma}(u^*(u_{0,h})(\cdot, 0) - u_h^{*0}, u_0 - u_{0,h}) \\ &\leq \frac{1}{\gamma}\|u^*(u_{0,h})(\cdot, 0) - u_h^{*0}\|_0\|u_0 - u_{0,h}\|_0.\end{aligned}$$

Hence,

$$\|u_0 - u_{0,h}\|_0 \leq \frac{1}{\gamma}\|u^*(u_{0,h})(\cdot, 0) - u_h^{*0}\|_0.$$

By using the triangle inequality and (4.30), the last step necessary for Theorem 17 is provided by

$$\begin{aligned}&\|u_h^{*0} - u^*(u_{0,h})(\cdot, 0)\|_0 \leq \|u_h^{*0} - u^*(u_h)(\cdot, 0)\|_0 + \|u^*(u_h)(\cdot, 0) - u^*(u_{0,h})(\cdot, 0)\|_0 \\ &\leq \max_{0\leq i\leq N-1} \|u_h^{*i} - u^*(u_h)(\cdot, t_i)\|_0 + \sup_{0\leq t<T} \|u^*(u_h) - u^*(u_{0,h})\|_0 \\ &\leq \max_{0\leq i\leq N-1} \|u_h^{*i} - u^*(u_h)(\cdot, t_i)\|_0 + C\|u_h - u(u_{0,h})\|_{L^2(0,T;L^2(\Omega))}.\end{aligned} \tag{4.53}$$

Rearranging inequalities (4.47), (4.48), (4.49), and (4.53), we conclude

$$\begin{aligned}&\|u_0 - u_{0,h}\|_0 + \|u - u_h\|_{L^2(0,T;L^2(\Omega))} + \|u^* - u_h^*\|_{L^2(0,T;L^2(\Omega))} \\ &\leq \frac{C}{\gamma} \max_{0\leq i\leq N-1} \|u_h^{*i} - u^*(u_h)(\cdot, t_i)\|_0 + \frac{C}{\gamma}\|u_h - u(u_{0,h})\|_{L^2(0,T;L^2(\Omega))} \\ &+ \|u^*(u_h) - u_h^*\|_{L^2(0,T;L^2(\Omega))}\end{aligned}$$

Using results in [39], the following classical error bounds holds:

$$\max_{0\le i\le N-1} \|u_h^{*i} - u^*(u_h)(\cdot, t_i)\|_0 \le C(h^2|_0 \log h| + \tau),$$
$$\|u^*(u_h) - u_h^*\|_{L^2(0,T;L^2(\Omega))} \le C(h^2|\log h| + \tau),$$
$$\|u(u_{0,h}) - u_h\|_{L^2(0,T;L^2(\Omega))} \le C(h^2|\log h| + \tau).$$

Finally, we have the convergence result,

$$\begin{aligned} & \|u_0 - u_{0,h}\|_0 + \|u - u_h\|_{L^2(0,T;L^2(\Omega))} + \|u^* - u_h^*\|_{L^2(0,T;L^2(\Omega))} \\ \le\ & C(\gamma)(h^2|\log h| + \tau), \end{aligned}$$

which completes the proof of Theorem 17.

Also, the dependence of the constant $C$ on $\gamma$ implies small regularization parameters may cause the numerical accuracy to degenerate. Hence, in practice one needs to use more refined mesh size $h$ and time step $\tau$ to reduce the finite element approximation error caused by small $\gamma$.

## 4.4. ITERATIVE METHODS SOLVING THE DISCRETE OPTIMALITY SYSTEM

Due to the forward in time nature in the state equation and backward in time nature of the adjoint equation, solving the discrete optimality system directly would generate a massive linear system and encounters computational difficulty. Considering the stability in data assimilation problem, in this section we develop two iterative algorithms, based on the conjugate gradient method and the steepest descent method, to decouple the discrete optimality system, which improve the computation efficiency significantly.

**4.4.1. Matrix Formulation.** We first derive the matrix formulation of the fully discrete optimality system (4.22). By definition of the operator $A$, the discrete optimality system (4.22) can be written as

$$\begin{cases} (\dfrac{u_h^{n+1}-u_h^n}{\tau}, v_h) + a(u_h^{n+1}, v_h) = (f_{n+1}, v_h), \\ u_h^0 = u_{0,h}, \\ -(\dfrac{u_h^{*n+1}-u_h^{*n}}{\tau}, v_h) + a(u_h^{*n}, v_h) = (\hat{u}^{n+1} - u_h^{n+1}, v_h), \\ u_h^{*N} = 0, \\ u_{0,h} = \dfrac{1}{\gamma} u_h^{*0} \end{cases} \tag{4.54}$$

for $n = 0, 1, 2, 3, ...., N-1$.

Considering the integral formula of the first equation in (4.54), we have

$$\begin{aligned} &\int_{\Omega} \frac{u_h^{n+1}-u_h^n}{\tau} v_h dxdy + \int_{\Omega^+} \beta^+ \nabla u_h^{n+1} \nabla v_h dxdy \\ &+ \int_{\Omega^-} \beta^- \nabla u_h^{n+1} \nabla v_h dxdy = \int_{\Omega^+} f_{n+1}^+ v_h dxdy + \int_{\Omega^-} f_{n+1}^- v_h dxdy. \end{aligned} \tag{4.55}$$

For each time moment $n$, $u_h^n = \sum_{j=1}^{N_b} u_j^n \phi_j$, plugging $u_h^n$ into (4.55) and using $v_h = \{\phi_i\}_{i=1}^{N_b}$ to test (4.55) respectively, we obtain

$$\begin{aligned} &\int_{\Omega} \frac{\sum_{j=1}^{N_b}(u_j^{n+1}-u_j^n)}{\tau} \phi_j \phi_i dxdy + \int_{\Omega^+} \sum_{j=1}^{N_b} u_j^{n+1} \beta^+ \frac{\partial \phi_j}{\partial x} \frac{\partial \phi_i}{\partial x} dxdy \\ &+ \int_{\Omega^+} \sum_{j=1}^{N_b} u_j^{n+1} \beta^+ \frac{\partial \phi_j}{\partial y} \frac{\partial \phi_i}{\partial y} dxdy + \int_{\Omega^-} \sum_{j=1}^{N_b} u_j^{n+1} \beta^- \frac{\partial \phi_j}{\partial x} \frac{\partial \phi_i}{\partial x} dxdy \\ &+ \int_{\Omega^-} \sum_{j=1}^{N_b} u_j^{n+1} \beta^- \frac{\partial \phi_j}{\partial y} \frac{\partial \phi_i}{\partial y} dxdy = \int_{\Omega^+} \sum_{j=1}^{N_b} f_{n+1}^+ \phi_i dxdy \\ &+ \int_{\Omega^-} \sum_{j=1}^{N_b} f_{n+1}^- \phi_i dxdy. \end{aligned}$$

Then the matrix formulation of the fully discrete forward state equation can be written as

$$\begin{cases} M \dfrac{\vec{u}_h^{n+1} - \vec{u}_h^n}{\tau} + Q \vec{u}_h^{n+1} = \vec{b}^{n+1}, \\ \vec{u}_h^0 = \vec{u}_{0,h}, \end{cases} \tag{4.56}$$

where

$$
\begin{aligned}
M &= \left[\int_{\Omega} \phi_j \phi_i dxdy\right]_{i,j=1}^{N_b}, \\
Q &= \left[\int_{\Omega^+} \beta^+ \frac{\partial \phi_j}{\partial x}\frac{\partial \phi_i}{\partial x} dxdy\right]_{i,j=1}^{N_b} + \left[\int_{\Omega^+} \beta^+ \frac{\partial \phi_j}{\partial y}\frac{\partial \phi_i}{\partial y} dxdy\right]_{i,j=1}^{N_b} \\
&+ \left[\int_{\Omega^-} \beta^- \frac{\partial \phi_j}{\partial x}\frac{\partial \phi_i}{\partial x} dxdy\right]_{i,j=1}^{N_b} + \left[\int_{\Omega^-} \beta^- \frac{\partial \phi_j}{\partial y}\frac{\partial \phi_i}{\partial y} dxdy\right]_{i,j=1}^{N_b}, \\
\vec{b}^{n+1} &= \left[\int_{\Omega^+} f_{n+1}^+ \phi_i dxdy\right]_{i=1}^{N_b} + \left[\int_{\Omega^-} f_{n+1}^- \phi_i dxdy\right]_{i=1}^{N_b}.
\end{aligned}
$$

Similarly, the matrix formulation of the fully discrete backward adjoint equation can be written as

$$
\begin{cases}
-M\dfrac{\vec{u}_h^{*n+1} - \vec{u}_h^{*n}}{\tau} + Q\vec{u}_h^{*n} = \vec{b}^{*n+1}, \\
\vec{u}_h^{*N} = \vec{0},
\end{cases}
\tag{4.57}
$$

where

$$
\begin{aligned}
\vec{b}^{*n+1} &= \vec{b}_{obs}^{*n+1} - \vec{b}_{u_h}^{n+1} = \left[\int_{\Omega} (\hat{u}^{n+1} - u_h^{n+1})\phi_i dxdy\right]_{i=1}^{N_b}, \\
\vec{b}_{obs}^{*n+1} &= \left[\int_{\Omega} \hat{u}^{n+1}\phi_i dxdy\right]_{i=1}^{N_b}, \\
\vec{b}_{u_h}^{n+1} &= \left[\int_{\Omega} u_h^{n+1}\phi_i dxdy\right]_{i=1}^{N_b}.
\end{aligned}
$$

Finally, the matrix formulation of the fully discrete optimality system is given by

$$
\begin{cases}
M\dfrac{\vec{u}_h^{n+1} - \vec{u}_h^{n}}{\tau} + Q\vec{u}_h^{n+1} = \vec{b}^{n+1}, \\
\vec{u}^0 = \vec{u}_{0,h}, \\
-M\dfrac{\vec{u}_h^{*n+1} - \vec{u}_h^{*n}}{\tau} + Q\vec{u}_h^{*n} = \vec{b}^{*n+1}, \\
\vec{u}_h^{*N} = \vec{0}, \\
\vec{u}_{0,h} = \dfrac{1}{\gamma}\vec{u}_h^{*0}.
\end{cases}
\tag{4.58}
$$

**4.4.2. The Conjugate Gradient Method.** Based on the conjugate gradient method and ideas in [54, 68], to efficiently solve (4.58) we propose the following iterative method to decouple the equation system (4.58): given $\vec{u}_{0,h}^{(0)}$, $\vec{u}_{0,h}^{(1)}$ and $\epsilon$, solve the following equations sequentially until the stop criteria $\|\vec{u}_{0,h}^{(i+1)} - \vec{u}_{0,h}^{(i)}\|_0 \leq \epsilon$ (or $\|\gamma\vec{u}_{0,h}^{(i+1)} - \vec{u}_h^{*0(i+1)}\|_0 \leq \epsilon$) is satisfied:

$$\begin{cases} M\dfrac{\vec{u}_h^{n+1(i)} - \vec{u}_h^{n(i)}}{\tau} + Q\vec{u}_h^{n+1(i)} = \vec{b}^{n+1}, \\ \vec{u}_h^{0(i)} = \vec{u}_{0,h}^{(i)}, \end{cases} \tag{4.59}$$

$$\begin{cases} -M\dfrac{\vec{u}_h^{*n+1(i)} - \vec{u}_h^{*n(i)}}{\tau} + Q\vec{u}_h^{*n(i)} = \vec{b}^{*n+1(i)}, \\ \vec{u}_h^{*N(i)} = \vec{0}, \end{cases} \tag{4.60}$$

$$\vec{u}_{0,h}^{(i+1)} = \vec{u}_{0,h}^{(i)} + \zeta^{i+1}B^i(\vec{u}_h^{*0(i)} - \gamma\vec{u}_{0,h}^{(i)}) + \eta^{i+1}C^i(\vec{u}_{0,h}^{(i)} - \vec{u}_{0,h}^{(i-1)}), \tag{4.61}$$

where $n = 0, 1, 2, 3, ..., N$ is the time evolution step, $i = 0, 1, 2, 3, ...$ represents the iteration step, $\zeta^{i+1}$ and $\eta^{i+1}$ are iterative parameters, $\vec{u}_{0,h}^{(i)}$, $\vec{u}_h^{n(i)}$, and $\vec{u}_h^{*n(i)}$ are iterative sequences, and $B^i$ and $C^i$ are two symmetric positive definite matrices.

Following the ideas in [54, 55] we adopt $B^i$ and $C^i$ as identity matrices. Then $\zeta^{i+1}$ and $\eta^{i+1}$ are updated using

$$\zeta^{i+1} = \frac{1}{q^{i+1}}, \quad \eta^{i+1} = \frac{e^i}{q^{i+1}}, \tag{4.62}$$

where

$$e^i = \begin{cases} 0 & i = 0, \\ q^i \frac{\|\vec{\lambda}^i\|_0^2}{\|\vec{\lambda}^{i-1}\|_0^2} & i > 0, \end{cases}$$

$$q^{i+1} = \frac{\|\vec{\lambda}^i\|_L^2}{\|\vec{\lambda}^i\|_0^2} - e^i, \quad i = 0, 1, 2, 3, ....$$

Here $\vec{\lambda}^i = \gamma\vec{u}_{0,h}^{(i)} - \vec{u}_h^{*0(i)}$ and $\|\vec{\lambda}^i\|_L = (L\vec{\lambda}^i, \vec{\lambda}^i)^{\frac{1}{2}}$. The operator $L$ acting on $\vec{\lambda}^i$ is defined as follows

$$\begin{cases} M\dfrac{\vec{\phi}_h^{n+1} - \vec{\phi}_h^n}{\tau} + Q\vec{\phi}_h^{n+1} = \vec{0}, \\ \vec{\phi}_h^0 = \vec{\lambda}^i, \end{cases} \tag{4.63}$$

$$\begin{cases} -M\dfrac{\vec{\phi}_h^{*n+1} - \vec{\phi}_h^{*n}}{\tau} + Q\vec{\phi}_h^{*n} = -\vec{b}_{\phi_h}^{*n+1}, \\ \vec{\phi}_h^{*N} = \vec{0}, \end{cases} \tag{4.64}$$

$$L\vec{\lambda}^i = \gamma\vec{\lambda}^i - \vec{\phi}_h^{*0}. \tag{4.65}$$

The above iterative scheme is carried out concisely using the following algorithm:

**Algorithm 10** *Step 0 (Initialization): Specify a convergence tolerance $\epsilon$, guess two initial functions $\vec{u}_{0,h}^{(0)}$ and $\vec{u}_{0,h}^{(1)}$, and then start the iteration at step $i = 1$.*

*Step 1 (Forward phase): Use $\vec{u}_{0,h}^{(i)}$ as the initial condition to solve (4.59) for $\vec{u}_h^{(i)}$.*

*Step 2 (Backward phase): Pass $\vec{u}_h^{(i)}$ to (4.60) and solve (4.60) backwards for $\vec{u}_h^{*0(i)}$.*

*Step 3 (Computing for operator L):*

*(1) Set $\vec{\lambda}^i = \gamma\vec{u}_{0,h}^{(i)} - \vec{u}_h^{*0(i)}$ and use it as initial value to solve equation (4.63) forward to obtain $\vec{\phi}_h$;*

*(2) Pass $\vec{\phi}_h$ to (5.89) and solve equation (4.64) backward for attaining $\vec{\phi}_h^{*0}$;*

*(3) Compute $L\vec{\lambda}^i = \gamma\vec{\lambda}^i - \vec{\phi}_h^{*0}$.*

*Step 4 (Update phase): Calculate $\zeta^{i+1}, \eta^{i+1}$ by using* (4.62) *and then update*

$$\vec{u}_{0,h}^{(i+1)} = \vec{u}_{0,h}^{(i)} + \zeta^{i+1}(\vec{u}_h^{*0(i)} - \gamma\vec{u}_{0,h}^{(i)}) + \eta^{i+1}(\vec{u}_{0,h}^{(i)} - \vec{u}_{0,h}^{(i-1)}).$$

*Step 5 (Criteria for stopping the iteration): Compute $\|\vec{u}_{0,h}^{(i+1)} - \vec{u}_{0,h}^{(i)}\|$. If $\|\vec{u}_{0,h}^{(i+1)} - \vec{u}_{0,h}^{(i)}\| \le \epsilon$ then stop. Otherwise increase i by* 1 *and go back to Step 1.*

**4.4.3. The Steepest Descent Method.** The conjugate gradient method serves an high convergence rate and solves the discrete optimality system (4.58) effectively for most of cases. However, it is relatively less stable and hence causes the algorithm to diverge for some of the data assimilation scenarios that have low stability, e.g., small regularization parameter $\gamma$ in the cost functional (4.18).

To tackle this numerical problem, we adopt the steepest descent method in [52, 53]which gains more stability at the cost of a lower convergence rate. For the purpose we need to calculate the gradient of the cost functional (4.4) and find out its representation in the admissable set.

$$J_h'(u_{0,h})v_h = \tau \sum_{n=1}^{N} (\hat{u}^n - u_h^n, (u_h^n)' v_h) + (\gamma u_{0,h}, v_h), \quad \forall v_h \in U_h. \tag{4.66}$$

By using the similar techniques as (3.24)-(3.28) in section 3.3, the gradient of the cost functional (4.4) is obtained as

$$J_h'(u_{0,h})v = (\gamma u_{0,h}, v_h) - (u_h^{*0}, v_h), \tag{4.67}$$

And $\gamma u_{0,h} - u_h^{*0}$ is the representation of the linear functionals $F_h'(u_{0,h})$ in the admissable set $U_h$.

With the result in (4.67), we now present the steepest descent method to solve the discrete data assimilation problem: given $\vec{u}_{0,h}^{(0)}$ and $\epsilon$, solve the following equations sequentially until the stop criteria $\|\vec{u}_{0,h}^{(i+1)} - \vec{u}_{0,h}^{(i)}\|_0 \leq \epsilon$ (or $\|\gamma \vec{u}_{0,h}^{(i+1)} - \vec{u}_h^{*0(i+1)}\|_0 \leq \epsilon$) is satisfied:

$$\begin{cases} M \dfrac{\vec{u}_h^{n+1(i)} - \vec{u}_h^{n(i)}}{\tau} + Q \vec{u}_h^{n+1(i)} = \vec{b}^{n+1}, \\ \vec{u}_h^{0(i)} = \vec{u}_{0,h}^{(i)}, \end{cases} \tag{4.68}$$

$$\begin{cases} -M\dfrac{\vec{u}_h^{*n+1(i)}-\vec{u}_h^{*n(i)}}{\tau}+Q\vec{u}_h^{*n(i)}=\vec{b}^{*n+1(i)}, \\ \vec{u}_h^{*N(i)}=\vec{0}, \end{cases} \tag{4.69}$$

$$\vec{u}_{0,h}^{(i+1)}=\vec{u}_{0,h}^{(i)}+\eta^{i+1}(\vec{u}_h^{*0(i)}-\gamma\vec{u}_{0,h}^{(i)}), \tag{4.70}$$

where $n=0,1,2,3....N$ is time evolution step, $i=0,1,2,3....$ represents the iteration step, $\eta^{i+1}$ is called the learning rate at each iteration, $\vec{u}_{0,h}^{(i)}$, $\vec{u}_h^{n(i)}$, $\vec{u}_h^{*n(i)}$ are iterative sequences.

To reduce the iterations and improve computational efficiency, the learning rate $\eta^{i+1}$ is determined by applying the inexact line search algorithm: find $\eta^{i+1}$ via repeatedly solving (4.68) with initial value

$$\vec{u}_{0,h}^{(i+1)}=\vec{u}_{0,h}^{(i)}+\eta^{i+1}(\vec{u}_h^{*0(i)}-\gamma\vec{u}_{0,h}^{(i)}) \quad \text{by updating} \quad \eta^{i+1}=\rho\eta^{i+1},$$

until the following inequality is satisfied

$$J_h(\vec{u}_{0,h}^{(i+1)})\leq J_h(\vec{u}_{0,h}^{(i)})+\delta\eta^{i+1}\langle J_h'(\vec{u}_{0,h}^{(i)}),\vec{u}_h^{*0(i)}-\gamma\vec{u}_{0,h}^{(i)}\rangle_{U_h^*\times U_h}, \tag{4.71}$$

where $\eta^{i+1}$ is typically initialized as a constant equal to or greater than 1, and $\delta$ and $\rho$ are chosen between $(0,1)$.

The above steepest descent method is implemented concisely using the following algorithm:

**Algorithm 11** *Step 0 (Initialization): Specify a convergence tolerance $\epsilon$, guess initial function $\vec{u}_{0,h}^{(0)}$, and start the iteration step $i=1$.*

*Step 1 (Forward phase): Use $\vec{u}_{0,h}^{(i)}$ as initial condition to solve equation (4.68) for $\vec{u}_h^{(i)}$.*

*Step 2 (Backward phase): Pass $\vec{u}_h^{(i)}$ to equation (4.69) and solve equation (4.69) backward for $\vec{u}_h^{*0(i)}$.*

*Step 3 (Inexact line search for $\eta^{i+1}$):*

*(1) Initialize a constant $\eta^{i+1} \geq 1$, set $0 < \rho < 1$ and $0 < \delta < 1$;*

*(2) use $\vec{u}_{0,h}^{(i+1)} = \vec{u}_{0,h}^{(i)} + \eta^{i+1}(\vec{u}_h^{*0(i)} - \gamma\vec{u}_{0,h}^{(i)})$ as initial value to solve equation (4.68) forward to obtain $\vec{u}_h^n$ for computing $F_h(\vec{u}_{0,h}^{(i+1)})$;*

*(3) Update $\eta^{i+1} = \rho\eta^{i+1}$ until inequality (4.71) is attained.*

*(4) Output $\eta^{i+1}$.*

*Step 4 (Update phase): Use $\eta^{i+1}$ from step 3 and then update*

$$\vec{u}_{0,h}^{(i+1)} = \vec{u}_{0,h}^{(i)} + \eta^{i+1}(\vec{u}_h^{*0(i)} - \gamma\vec{u}_{0,h}^{(i)}).$$

*Step 5 (Criteria for stopping the iteration): Compute $\|\vec{u}_h^{*0(i)} - \gamma\vec{u}_{0,h}^{(i)}\|$, if $\|\vec{u}_h^{*0(i)} - \gamma\vec{u}_{0,h}^{(i)}\| \leq \epsilon$ then stop; otherwise, increase i by 1 and go back to Step 1.*

**Remark 13 (Incremental POD for gradient methods)** *As mentioned in the Section 3.5.1, we may run into memory difficulties to store data $\{u_h^i\}_{i=1}^N$ in the Forward phase for both conjugate gradient method and steepest descent method. It is then time to use the incremental POD data compression technique. Define the needed thresholds for truncations and re-orthogonalization, recall the weighted matrix, i.e., mass matrix $M$, and time step size $\tau$ have already been provided in above, we implement the algorithm 6 on the step of Forward phase to compress $\{u_h^i\}_{i=1}^N$ into a smaller size matrix, then reconstruct $\{u_h^i\}$ at each time step when used for the Backward phase solving. By doing so, storing large-scale data is avoided.*

## 4.5. NUMERICAL EXPERIMENTS

Based on the iterative methods developed in previous section, this section will present numerical results to demonstrate the performance of the reconstructed initial condition. The finite element space is chosen on continuous piecewise linear polynomial with mesh size $h$, and backward Euler scheme is used with time step $\tau = h^2$ correspondingly.

$L^\infty, L^2$ error norm will be considered respectively. We especially focus on the $L^2$ error norm, since the way we measure the distance between observations and state variable in the cost functional is in an $L^2$ norm sense.

**4.5.1. Verification of the Finite Element Convergence Rate.** Before showing the data assimilation performance, we provide an example to verify the conclusions in Theorem 17. Given a set of smooth observations and for each fixed regularization parameter $\gamma$, we expect to observe that the finite element approximation converges in a second order regarding to $L^2$ norm. Mesh sizes of $1/4$, $1/8$, $1/16$, $1/32$ and time step sizes of $1/16$, $1/64$, $1/256$, $1/1024$ are used, respectively. For each fixed $\gamma$, the discrete solution with $h = 1/64$ and $\tau = 1/4096$ will be used to represent the analytical solution.

The distributed observations are given by

$$u^+ = \sin(\pi \cdot x)\sin(\pi \cdot y)\sin(t+1),\; u^- = 2\sin(\pi \cdot x)\sin(\pi \cdot y)\sin(t+1).$$

Other relevant parameters are set as: $\beta^+ = 1$, $\beta^- = \frac{1}{2}$, $\Omega^+ = (0,1)\times(0,1)$, $\Omega^- = (1,2)\times(0,1)$, $\Gamma : x = 1$, and $T = 1$. The boundary condition and jump interface condition satisfy $u = 0$ on $\partial\Omega$, $[u]|_\Gamma = 0$ on $\Gamma$, and $[\beta(x,y)\frac{\partial u}{\partial \vec{n}}]|_\Gamma = 0$. Both $f^+$ and $f^-$ can be computed by using $u^+$, $u^-$, $\beta^+$, and $\beta^-$.

Numerical results are displayed in Table 4.1, where the $L^2$ norm errors appear to satisfy the optimal convergence rate. In additional, at each column, the error between the analytical solution and numerical solution is tending to be larger when $\gamma$ decreases, which indicates the coefficient $C(\gamma)$ in inequality (4.27) is proportional to $\frac{1}{\gamma}$. And this negative behavior is reduced effectively while smaller mesh size $h$ and time step $\tau$ are used.

**4.5.2. Data Assimilation Performance without Observation Noise.** We now investigate the numerical performance for the data assimilation problem utilizing the iterative methods in section 4.4. In this example, we assume there is no noise in the distributed

Table 4.1. The finite element convergence rate for the recovered initial condition $u_0$ of the Second order Parabolic Interface equation. Here $u_0 \approx u_{0,\frac{1}{64}}$.

| Finite Element Convergence Rate | | | | | |
|---|---|---|---|---|---|
| $\gamma$ | $\|\|u_0 - u_{0,\frac{1}{8}}\|\|$ | $\|\|u_0 - u_{0,\frac{1}{16}}\|\|$ | rate | $\|\|u_0 - u_{0,\frac{1}{32}}\|\|$ | rate # |
| 1 | $1.0 \times 10^{-3}$ | $2.8 \times 10^{-4}$ | 1.85 | $5.7 \times 10^{-5}$ | 2.30 |
| $\frac{1}{10}$ | $3.9 \times 10^{-3}$ | $1.1 \times 10^{-3}$ | 1.84 | $2.7 \times 10^{-4}$ | 2.03 |
| $\frac{1}{50}$ | $1.1 \times 10^{-2}$ | $2.7 \times 10^{-3}$ | 2.03 | $5.1 \times 10^{-4}$ | 2.40 |
| $\frac{1}{200}$ | $2.7 \times 10^{-2}$ | $6.2 \times 10^{-3}$ | 2.12 | $1.5 \times 10^{-3}$ | 2.05 |
| $\frac{1}{500}$ | $3.6 \times 10^{-2}$ | $7.2 \times 10^{-3}$ | 2.32 | $1.4 \times 10^{-3}$ | 2.36 |
| $\frac{1}{1000}$ | $4.0 \times 10^{-2}$ | $7.9 \times 10^{-3}$ | 2.33 | $1.6 \times 10^{-3}$ | 2.31 |

observations. Hence, the sample is given by the exact solution. We test the expected numerical performance by adjusting the regularization parameter $\gamma$. The spatial and temporal step sizes are set to $1/50$ and $1/200$, respectively.

Distributed observations are generated by the exact solution

$$u^+ = \sin(\pi \cdot x)\sin(\pi \cdot y)\sin(t+1), \;\; u^- = 2\sin(\pi \cdot x)\sin(\pi \cdot y)\sin(t+1).$$

Other relevant parameters are set as: $\beta^+ = 1$, $\beta^- = \frac{1}{2}$, $\Omega^+ = (0,1) \times (0,1)$, $\Omega^- = (1,2) \times (0,1)$, $\Gamma : x = 1$, $T = 1$. The boundary condition and jump interface condition satisfy: $u = 0$ on $\partial\Omega$, $[u]|_\Gamma = 0$ on $\Gamma$, and $[\beta(x,y)\frac{\partial u}{\partial n}]|_\Gamma = 0$ on $\Gamma$. Both $f^+$ and $f^-$ can be computed by using $u^+, u^-, \beta^+, \beta^-$.

In Table 4.2 and 4.3, the error between the numerical results and the observations (or the exact solution) becomes smaller as $\gamma$ decreases from 1 to $\frac{1}{10000}$. The numerical results for the recovered initial condition becomes more accurate correspondingly. These observations match our practical expectation.

Furthermore, the convergence comparison of the two iterative methods indicates that, in practice, the conjugate gradient method is preferred for the moderate $\gamma$ because of its higher convergence rate, and the steepest gradient is a prior option for the small $\gamma$ due to its capability for maintaining the stability.

Table 4.2. Conjugate Gradient Method: data assimilation result without noise for the Second order Parabolic Interface equation. $u$ and $u_h$ are the exact solution and the numerical simulation performance, $||u-u_h||_{L^2} = \sum_{n=1}^{N} \tau \|u^n - u_h^n\|_0$, and $||u-u_h||_{L^\infty} = \sum_{n=1}^{N} \tau \|u^n - u_h^n\|_{L^\infty(\Omega)}$.

| $L^2, L^\infty$ Norm of $u-u_h$: Conjugate Gradient Method | | | |
|---|---|---|---|
| $\gamma$ | $\|\|u-u_h\|\|_{L^2}$ | $\|\|u-u_h\|\|_{L^\infty}$ | Iteration # |
| 1 | $9.26\times10^{-2}$ | $1.70\times10^{-1}$ | 3 |
| $\frac{1}{10}$ | $6.32\times10^{-2}$ | $1.18\times10^{-1}$ | 12 |
| $\frac{1}{200}$ | $8.50\times10^{-3}$ | $1.87\times10^{-2}$ | 14 |
| $\frac{1}{2000}$ | $2.42\times10^{-3}$ | $6.20\times10^{-3}$ | 21 |
| $\frac{1}{10000}$ | \ | \ | $\infty$ |

Table 4.3. Steepest Descent Method: data assimilation result without noise for the Second order Parabolic Interface equation. $u$ and $u_h$ are the exact solution and the numerical simulation performance, $||u-u_h||_{L^2} = \sum_{n=1}^{N} \tau \|u^n - u_h^n\|_0$, and $||u-u_h||_{L^\infty} = \sum_{n=1}^{N} \tau \|u^n - u_h^n\|_{L^\infty(\Omega)}$.

| $L^2, L^\infty$ Norm of $u-u_h$: Steepest Descent Method | | | |
|---|---|---|---|
| $\gamma$ | $\|\|u-u_h\|\|_{L^2}$ | $\|\|u-u_h\|\|_{L^\infty}$ | Iteration# |
| 1 | $9.26\times10^{-2}$ | $1.70\times10^{-1}$ | 4 |
| $\frac{1}{10}$ | $6.35\times10^{-2}$ | $1.19\times10^{-1}$ | 14 |
| $\frac{1}{200}$ | $1.08\times10^{-2}$ | $1.90\times10^{-2}$ | 50 |
| $\frac{1}{2000}$ | $3.42\times10^{-3}$ | $5.83\times10^{-3}$ | 59 |
| $\frac{1}{10000}$ | $2.08\times10^{-3}$ | $4.82\times10^{-3}$ | 60 |

**4.5.3. Data Assimilation Performance Test with Observation Noise.** We now consider a more realistic case, which introduces noise with a normal distribution $N(0, 1/100)$ into the observation sample. For the finite element approximation we still use a mesh size of $1/50$ and a time step of $1/200$. The numerical data assimilation performance is shown by altering $\gamma$.

The exact solution is given by

$$u^+ = \sin(\pi\cdot x)\sin(\pi\cdot y)\sin(t+1),\ u^- = 2\sin(\pi\cdot x)\sin(\pi\cdot y)\sin(t+1).$$

Other relevant parameters are set as $\beta^+ = 1, \beta^- = \frac{1}{2}, \Omega^+ = (0,1)\times(0,1), \Omega^- = (1,2)\times(0,1)$, $\Gamma : x = 1$, and $T = 1$. The boundary condition and jump interface condition satisfy $u = 0$ on $\partial\Omega$, $[u]|_\Gamma = 0$ on $\Gamma$, and $[\beta(x,y)\frac{\partial u}{\partial \vec{n}}]|_\Gamma = 0$ on $\Gamma$. Both $f^+$ and $f^-$ are computed by using $u^+$, $u^-$, $\beta^+$, and $\beta^-$.

Based on this setup, the distributed observations are provided by adding noise from the normal distribution $N(0, 1/100)$ into the exact solution. In this experiment, due to the differences between the exact solution and the observations, we investigate the distance from the numerical results to both the exact solution and the observations.

Table 4.4. Conjugate Gradient Method: data assimilation result with noise for the Second order Parabolic Interface equation. $u$ and $u_h$ are the exact solution and the numerical simulation performance, $||u - u_h||_{L^2} = \sum_{n=1}^{N} \tau \|u^n - u_h^n\|_0$, and $||u - u_h||_{L^\infty} = \sum_{n=1}^{N} \tau \|u^n - u_h^n\|_{L^\infty(\Omega)}$.

| $L^2, L^\infty$ Norm of $u - u_h$: Conjugate Gradient Method | | | |
|---|---|---|---|
| $\gamma$ | $\|\|u - u_h\|\|_{L^2}$ | $\|\|u - u_h\|\|_{L^\infty}$ | Iteration # |
| 1 | $9.03 \times 10^{-2}$ | $1.70 \times 10^{-1}$ | 3 |
| $\frac{1}{10}$ | $6.32 \times 10^{-2}$ | $1.18 \times 10^{-1}$ | 16 |
| $\frac{1}{200}$ | $1.01 \times 10^{-2}$ | $1.83 \times 10^{-2}$ | 15 |
| $\frac{1}{2000}$ | \ | \ | $\infty$ |
| $\frac{1}{10000}$ | \ | \ | $\infty$ |

Table 4.5. Steepest Descent Method: data assimilation result with noise for the Second order Parabolic Interface equation. $u$ and $u_h$ are the exact solution and the numerical simulation performance, $||u - u_h||_{L^2} = \sum_{n=1}^{N} \tau \|u^n - u_h^n\|_0$, and $||u - u_h||_{L^\infty} = \sum_{n=1}^{N} \tau \|u^n - u_h^n\|_{L^\infty(\Omega)}$.

| $L^2, L^\infty$ Norm of $u - u_h$: Steepest Descent Method | | | |
|---|---|---|---|
| $\gamma$ | $\|\|u - u_h\|\|_{L^2}$ | $\|\|u - u_h\|\|_{L^\infty}$ | Iteration # |
| 1 | $9.02 \times 10^{-2}$ | $1.70 \times 10^{-1}$ | 4 |
| $\frac{1}{10}$ | $6.35 \times 10^{-2}$ | $1.19 \times 10^{-1}$ | 14 |
| $\frac{1}{200}$ | $1.09 \times 10^{-2}$ | $1.92 \times 10^{-2}$ | 50 |
| $\frac{1}{2000}$ | $3.21 \times 10^{-3}$ | $6.01 \times 10^{-3}$ | 59 |
| $\frac{1}{10000}$ | $2.50 \times 10^{-3}$ | $4.90 \times 10^{-3}$ | 60 |

Table 4.6. Steepest Descent Method: data assimilation result with noise for the Second order Parabolic Interface equation. $\hat{u}$ and $u_h$ are the exact solution and the numerical simulation performance, $||\hat{u} - u_h||_{L^2} = \sum_{n=1}^{N} \tau\|\hat{u}^n - u_h^n\|_0$, and $||\hat{u} - u_h||_{L^\infty} = \sum_{n=1}^{N} \tau\|\hat{u}^n - u_h^n\|_{L^\infty(\Omega)}$.

| $L^2, L^\infty$ Norm of $\hat{u} - u_h$: Steepest Descent Method | | | |
|---|---|---|---|
| $\gamma$ | $\|\|\hat{u} - u_h\|\|_{L^2}$ | $\|\|\hat{u} - u_h\|\|_{L^\infty}$ | Iteration # |
| 1 | $9.38 \times 10^{-2}$ | $1.94 \times 10^{-1}$ | 4 |
| $\frac{1}{10}$ | $6.75 \times 10^{-2}$ | $1.43 \times 10^{-1}$ | 14 |
| $\frac{1}{200}$ | $1.70 \times 10^{-2}$ | $4.77 \times 10^{-2}$ | 50 |
| $\frac{1}{2000}$ | $1.07 \times 10^{-2}$ | $3.71 \times 10^{-2}$ | 59 |
| $\frac{1}{10000}$ | $1.03 \times 10^{-2}$ | $3.58 \times 10^{-2}$ | 60 |

In Table 4.4 and 4.5, the convergence comparison again confirms the advantages of a higher convergence rate from the conjugate gradient method and reliable stability of the steepest descent method. In Table 4.6, the distance measured with the $L^2$ and $L^\infty$ norms between the observations and the numerical results always becomes smaller as $\gamma$ decreases. In Table 4.4, a desired accuracy of the recovered initial condition, based on the noisy observations, can be attained by adjusting the characterization parameter $\gamma$, which validates the proposed methods to solve the data assimilation problem for the second order parabolic interface equation.

**4.5.4. Data Assimilation Results using Incremental POD.** In this section, we present the numerical results using incremental POD data compression. The observed data and parabolic interface model have the same set-up in Section 4.5.2. For the numerical discretization, we use $\tau = 1/100$ and $h = 1/32$.

In Table 4.7, we show that the use of incremental POD saves storage at least %70 – %80, which effectively solves the memory issues for gradient descent method in the data assimilation problems. The promising part from incremental POD is that we might not compromise any accuracy with the approximated data if the POD truncation is less than the gradient tolerance, which are displayed in Table 4.8. This is probably due to the correction of the iterative gradient method. It also might be the total information loss in our example is very minimal. These are interesting behaviors deserving more investigation for the POD data compression in optimization problem.

Table 4.7. Memory saved from the incremental POD for the data assimilation of the Second order Parabolic Interface equation. $1024 \times 100$ is the data matrix size we need to save in gradient methods, $1024 \times 9$ or $1024 \times 15$ is the data matrix size after POD compression. The POD truncation thresholds are all $10^{-10}$.

| Memory Saved During Iterations | | | |
|---|---|---|---|
| $\gamma$ | Original Data size | Compressed Data size | Storage Saved |
| 1 | $1024 \times 100$ | $1024 \times 9$ | 82% |
| $\frac{1}{10}$ | $1024 \times 100$ | $1024 \times 15$ | 70% |
| $\frac{1}{50}$ | $1024 \times 100$ | $1024 \times 15$ | 70% |

Table 4.8. Data assimilation comparison between the use and no use of the incremental POD of the Second order Parabolic Interface equation. $u$ is the exact solution, $u_h$ is the numerical result without compression, $u_h^C$ is the results applied data compression, $||u - u_h||_{L^2} = \sum_{n=1}^{N} \tau \|u^n - u_h^n\|_0$ and $||u - u_h^C||_{L^2} = \sum_{n=1}^{N} \tau \|u^n - u_h^{nC}\|_0$.

| Error comparison between compressed and no-compressed data | | |
|---|---|---|
| $\gamma$ | $\|\|u - u_h\|\|_{L^2}$ | $\|\|u - u_h^C\|\|_{L^2}$ |
| 1 | $4.69 \times 10^{-2}$ | $4.69 \times 10^{-2}$ |
| $\frac{1}{10}$ | $6.51 \times 10^{-3}$ | $6.51 \times 10^{-3}$ |
| $\frac{1}{50}$ | $3.05 \times 10^{-3}$ | $3.05 \times 10^{-3}$ |

**4.5.5. Data Assimilation Results using Parallel Algorithm.** This section presents the numerical results using the parallel algorithm developed in Section 3.6. We still use the same model parameter as Section 4.5.2 to generate observations and test numerical performance.

Table 4.9 shows that the computational cost is reduced when we simulate the VDA problem in a parallel manner. In Table 4.10, the numerical results show that the parallel algorithm does not affect the data assimilation accuracy. One disadvantage observed from our numerical experiments is that some of the decoupled problems are not well-conditioned. It means the preconditioning is necessary for the parallel algorithm, and it will be an interesting near future work.

Table 4.9. Computational cost saved from a sequential test of the Parallel algorithm

| Sequential test of Parallel algorithm $\gamma = 1$ no noise | | |
|---|---|---|
| # of partition | # of iteration | Time saved |
| 8 | 6 | $\approx 25\%$ |
| 16 | 8 | $\approx 50\%$ |
| 32 | 8 | $\approx 75\%$ |

Table 4.10. Data assimilation performance comparison between the parallel algorithm and the steepest gradient descent. $||u - u_h||^p_{L^2} = \sum_{n=1}^N \tau\|u^n - u_h^n\|_0^p$ and $||u - u_h||_{L^2} = \sum_{n=1}^N \tau\|u^n - u_h^n\|_0$ are numerical results with and without parallel algorithm, $||u - u_h||^p_{L^\infty} = \sum_{n=1}^N \tau\|u^n - u_h^n\|^p_{L^\infty}$ and $||u - u_h||_{L^\infty} = \sum_{n=1}^N \tau\|u^n - u_h^n\|_{L^\infty}$ are defined same.

| Accuracy comparison $\gamma = 1$ no noise | | | | |
|---|---|---|---|---|
| # partition | $\lvert\lvert u - u_h\rvert\rvert^p_{L^2}$ | $\lvert\lvert u - u_h\rvert\rvert^p_{L^\infty}$ | $\lvert\lvert u - u_h\rvert\rvert_{L^2}$ | $\lvert\lvert u - u_h\rvert\rvert_{L^\infty}$ |
| 8 | $9.46 \times 10^{-2}$ | $1.80 \times 10^{-1}$ | $9.46 \times 10^{-2}$ | $1.80 \times 10^{-2}$ |
| 16 | $9.46 \times 10^{-2}$ | $1.80 \times 10^{-1}$ | $9.46 \times 10^{-2}$ | $1,80 \times 10^{-1}$ |
| 32 | $9.46 \times 10^{-2}$ | $1.80 \times 10^{-1}$ | $9.46 \times 10^{-2}$ | $1.80 \times 10^{-1}$ |

# 5. DATA ASSIMILATION FOR STOKES-DARCY EQUATION

## 5.1. BACKGROUND FOR THE STOKES-DARCY MODEL

The Stokes-Darcy model is receiving more attentions nowadays due to its potential applications to a variety of flow phenomena, for instance, the hydrological system where surface water percolates through rock and sand [69–71], petroleum extraction [72–80], and industrial filtration [81, 82]. In recent decades, a significant effort has been on studying this sophisticated interface system both theoretically and numerically [41, 42, 83–98]. However, these existing works were dedicated to the idealized model, i.e., the relevant input data, such as initial condition, boundary condition, sink/source term, and diffusion coefficients, are entirely provided for the model prediction. In real implementations, some of these input data literally remain unknown or in uncertainty [99–101]. Therefore, one of the challenging problems is to identify a set of faithful needed data such that the forecast of the target flow can be performed reliably. This is where the data assimilation comes in [102].

We consider a free flow in a bounded domain $\Omega_f$ and a porous media flow in another bounded domain $\Omega_p$. These two flows are coupled together in the domain $\Omega$ through the interface $\Gamma = \bar{\Omega}_p \cap \bar{\Omega}_f$ such that $\bar{\Omega} = \bar{\Omega}_p \cup \bar{\Omega}_f$. We also let $\Gamma_p = \partial\Omega_p \setminus \Gamma$ and $\Gamma_f = \partial\Omega_f \setminus \Gamma$. A Stokes-Darcy model can be used to describe this coupled fluid phenomena. Then the porous media flow is governed by the Darcy equation:

$$
\begin{aligned}
&\frac{\partial \phi}{\partial t} - \nabla \cdot (\mathbb{K}\nabla\phi) = f_p \quad \text{in } \Omega_p \times (0,T], \\
&\phi(\cdot, 0) = \phi_0 \quad \text{in } \Omega_p, \\
&\phi = 0 \quad \text{on } \Gamma_p,
\end{aligned}
\tag{5.1}
$$

where $\phi$ denotes the hydraulic head, $\mathbb{K}$ is the hydraulic conductivity tensor assumed to be homogeneous isotropic in this paper, i.e., $\mathbb{K} = K\mathbb{I}$ with a constant $K$, and $f_p$ is a sink /source term. On the other hand, the free flow is governed by the Stokes equation:

$$\begin{aligned} &\frac{\partial \boldsymbol{u}}{\partial t} - \nabla \cdot \mathbb{T}(\boldsymbol{u}, p) = \boldsymbol{f}_f \quad \text{in } \Omega_f \times (0, T], \\ &\nabla \cdot \boldsymbol{u} = 0 \quad \text{in } \Omega_f \times (0, T], \\ &\boldsymbol{u}(\cdot, 0) = \boldsymbol{u}_0 \quad \text{in } \Omega_f, \\ &\boldsymbol{u} = 0 \quad \text{on } \Gamma_f, \end{aligned} \tag{5.2}$$

where $\boldsymbol{u}$ denotes the fluid velocity, $\mathbb{T}(\boldsymbol{u}, p) = 2\nu\mathbb{D}(\boldsymbol{u}) - p\mathbb{I}$ is the stress tensor, $\mathbb{D}(\boldsymbol{u}) = \frac{1}{2}(\nabla \boldsymbol{u} + \nabla^T \boldsymbol{u})$ is the deformation tensor, $\nu$ is the kinematic viscosity of the fluid, $p$ is the kinematic pressure, and $\boldsymbol{f}_f$ is a general external forcing term that includes gravitational acceleration. Systems (5.1) and (5.2) interact on $\Gamma$ through the Beavers-Joseph interface conditions, see [103–112]:

$$\begin{aligned} &\boldsymbol{u} \cdot \boldsymbol{n}_f = \mathbb{K}\nabla\phi \cdot \boldsymbol{n}_p, \\ &-\boldsymbol{\tau} \cdot (\mathbb{T}(\boldsymbol{u}, p) \cdot \boldsymbol{n}_f) = \alpha\boldsymbol{\tau} \cdot (\boldsymbol{u} + \mathbb{K}\nabla\phi), \\ &-\boldsymbol{n}_f \cdot (\mathbb{T}(\boldsymbol{u}, p) \cdot \boldsymbol{n}_f) = g(\phi - z), \end{aligned} \tag{5.3}$$

where $\boldsymbol{n}_f$ and $\boldsymbol{n}_p$ denote the outer normal vectors to the fluid and the porous media regions on the interface $\Gamma$, respectively, $\boldsymbol{\tau}$ denotes the unit tangential vector to the interface $\Gamma$, $\alpha$ is a constant depending on $\nu$ and $\mathbb{K}$, $g$ is the gravitational acceleration, $z$ is a constant assumed to be 0 from now on.

For the purpose of discussing the data assimilation problem, it is necessary to appropriately understand the Stokes-Darcy model. We first define the Hilbert spaces

$$
\begin{aligned}
&X_p := \{\psi \in H^1\left(\Omega_p\right) : \ \psi = 0 \text{ on } \partial\Omega_p \setminus \Gamma\}, \\
&\boldsymbol{X}_f := \{\boldsymbol{v} \in \boldsymbol{H}^1\left(\Omega_f\right) : \ \boldsymbol{v} = (v_1, v_2)^T = \boldsymbol{0} \text{ on } \partial\Omega_f \setminus \Gamma\}, \\
&\boldsymbol{X} := X_p \times \boldsymbol{X}_f, \ \boldsymbol{X}_{\text{div}}^f := \{\boldsymbol{v} \in \boldsymbol{X}_f : \ \nabla \cdot \boldsymbol{v} = 0\}, \ \boldsymbol{X}_{\text{div}} := X_p \times \boldsymbol{X}_{\text{div}}^f, \\
&Q := L^2\left(\Omega_f\right), \ \boldsymbol{L}^2(\Omega) := L^2\left(\Omega_p\right) \times \boldsymbol{L}^2\left(\Omega_f\right)
\end{aligned}
$$

and the corresponding norms

$$
\begin{aligned}
&\|\psi\|_{X_p} := \|\psi\|_{H^1(\Omega_p)}, \ \|\boldsymbol{v}\|_{\boldsymbol{X}_f} := \|\boldsymbol{v}\|_{\boldsymbol{H}^1(\Omega_f)} = \left(\|v_1\|^2_{H^1(\Omega_f)} + \|v_2\|^2_{H^1(\Omega_f)}\right)^{\frac{1}{2}}, \\
&\|\boldsymbol{V}\|_{\boldsymbol{X}} := \left(\|\boldsymbol{v}\|^2_{\boldsymbol{X}_f} + \|\psi\|^2_{X_p}\right)^{\frac{1}{2}}, \ \|\boldsymbol{v}\|_{\boldsymbol{X}^f_{\text{div}}} := \|\boldsymbol{v}\|_{\boldsymbol{H}^1(\Omega_f)}, \\
&\|\boldsymbol{V}\|_{\boldsymbol{X}_{\text{div}}} := \left(\|\boldsymbol{v}\|^2_{\boldsymbol{X}_f} + \|\psi\|^2_{X_p}\right)^{\frac{1}{2}}, \ \|\boldsymbol{V}\|_{\boldsymbol{L}^2(\Omega)} := \left(\|\psi\|^2_{L^2(\Omega_p)} + \|\boldsymbol{v}\|^2_{\boldsymbol{L}^2(\Omega_f)}\right)^{\frac{1}{2}}.
\end{aligned}
$$

For a domain $D$, $(\cdot,\cdot)_D$ denotes the $L^2$ inner product on $D$, $(\cdot,\cdot)_H$ denotes the inner product for other Hilbert spaces $H(D)$. Depending on the context, $\langle\cdot,\cdot\rangle$ can represent the inner product on the interface $\Gamma$, or a general duality between a Banach space and its dual space. For simplicity, let $\|\cdot\|_0$ denote all the $L^2$ norms, $H^m(D)$ denote the Sobolev space $W^{m,2}(D)$. Besides, considering the temporal-spatial function spaces, let $L^p(0,T;\mathcal{B}) = W^{0,p}(0,T;\mathcal{B})$ and $H^m(0,T;\mathcal{B}) = W^{m,2}(0,T;\mathcal{B})$, where $\mathcal{B}$ is a generic Banach space. We use these notations to define the following bilinear forms and linear functionals:

$$
\begin{aligned}
a_p(\phi,\psi) &= (K\nabla\phi,\nabla\psi)_{\Omega_p},\ \forall\phi,\psi\in X_p,\\
a_f(\boldsymbol{u},\boldsymbol{v}) &= 2\nu\left(\mathbb{D}(\boldsymbol{u}),\mathbb{D}(\boldsymbol{v})\right)_{\Omega_f}\ \forall\boldsymbol{u},\boldsymbol{v}\in\boldsymbol{X}_f,\\
a(\boldsymbol{U},\boldsymbol{V}) &= a_f(\boldsymbol{u},\boldsymbol{v})+a_p(\phi,\psi)+\langle g\phi,\boldsymbol{v}\cdot\boldsymbol{n}_f\rangle-\langle\boldsymbol{u}\cdot\boldsymbol{n}_f,\psi\rangle\\
&\quad+\alpha\langle P_\tau(\boldsymbol{u}+\mathbb{K}\nabla\phi),P_\tau\boldsymbol{v}\rangle\quad\forall\boldsymbol{U}=(\phi,\boldsymbol{u})^T\in\boldsymbol{X},\ \forall\boldsymbol{V}=(\psi,\boldsymbol{v})^T\in\boldsymbol{X},\\
b_f(\boldsymbol{v},p) &= -(\nabla\cdot\boldsymbol{v},p)_{\Omega_f},\ \forall\boldsymbol{v}\in\boldsymbol{X}_f,\ \forall p\in Q,\\
b(\boldsymbol{V},p) &= b_f(\boldsymbol{v},p),\ \forall\boldsymbol{V}\in\boldsymbol{X},\ \forall p\in Q,\\
\langle\boldsymbol{F},\boldsymbol{V}\rangle &= (f_p,\psi)_{\Omega_p}+(\boldsymbol{f}_f,\boldsymbol{v})_{\Omega_f}\quad\forall\boldsymbol{F}=(f_p,\boldsymbol{f}_f)^T\in\boldsymbol{X}',\ \forall\boldsymbol{V}\in\boldsymbol{X},\\
\langle\frac{\partial\boldsymbol{U}}{\partial t},\boldsymbol{V}\rangle &= \langle\frac{\partial\phi}{\partial t},\psi\rangle+\langle\frac{\partial\boldsymbol{u}}{\partial t},\boldsymbol{v}\rangle\quad\forall\boldsymbol{U}\in\boldsymbol{H}^1(0,T;\boldsymbol{X}'),\ \forall\boldsymbol{V}\in\boldsymbol{X},
\end{aligned}
\tag{5.4}
$$

where $P_\tau$ denotes the projection onto the tangent space on $\Gamma$, i.e., $P_\tau\boldsymbol{u}=(\boldsymbol{u}\cdot\boldsymbol{\tau})\boldsymbol{\tau}$. For $\langle P_\tau(\mathbb{K}\nabla\phi),P_\tau\boldsymbol{v}\rangle$ in (5.4), we need the trace space defined as $\mathbf{H}_{\mathbf{00}}^{\mathbf{1/2}}(\Gamma):=\boldsymbol{X}_f|_\Gamma$, which is a non-closed subspace of $\mathbf{H}_{\mathbf{0}}^{\mathbf{1/2}}(\Gamma)$ and has continuous zero extension to $\mathbf{H}_{\mathbf{0}}^{\mathbf{1/2}}(\partial\Omega_f)$, $\langle P_\tau(\mathbb{K}\nabla\phi),P_\tau\boldsymbol{v}\rangle$ is then interpreted as a duality between $(\mathbf{H}_{\mathbf{00}}^{\mathbf{1/2}}(\Gamma))'$ and $\mathbf{H}_{\mathbf{00}}^{\mathbf{1/2}}(\Gamma)$. See [106] and references cited therein for more details.

We use bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ to define linear operators $A$, $A^*$, $B$, and $B^*$:

$$
a(\boldsymbol{U},\boldsymbol{V})=\langle A\boldsymbol{U},\boldsymbol{V}\rangle=\langle\boldsymbol{U},A^*\boldsymbol{V}\rangle,\quad b(\boldsymbol{V},p)=\langle B\boldsymbol{V},p\rangle=\langle\boldsymbol{V},B^*p\rangle,
\tag{5.5}
$$

where $A\in\mathscr{L}(\boldsymbol{X},\boldsymbol{X}')$, $B\in\mathscr{L}(\boldsymbol{X},Q')$, $A^*\in\mathscr{L}(\boldsymbol{X},\boldsymbol{X}')$ and $B^*\in\mathscr{L}(Q,\boldsymbol{X}')$ are the adjoint operator of $A$ and $B$, and $\mathscr{L}$ is the set of linear and continuous operators for the relevant spaces.

Testing systems (5.1) and (5.2) with $(\psi, \boldsymbol{v}, q)^T \in \boldsymbol{X} \times Q$ and incorporating the three interface conditions in (5.3), we obtain the weak formulation of the Stokes-Darcy model:

$$\begin{cases} \langle \dfrac{\partial \phi}{\partial t}, \psi \rangle + a_p(\phi, \psi) - \langle \boldsymbol{u} \cdot \boldsymbol{n}_f, \psi \rangle = \langle f_p, \psi \rangle \quad \forall \psi \in X_p, \\ \langle \dfrac{\partial \boldsymbol{u}}{\partial t}, \boldsymbol{v} \rangle + a_f(\boldsymbol{u}, \boldsymbol{v}) + b_f(\boldsymbol{v}, p) + \langle g\phi, \boldsymbol{v} \cdot \boldsymbol{n}_f \rangle + \alpha \langle P_\tau(\boldsymbol{u} + \mathbb{K}\nabla\phi), P_\tau \boldsymbol{v} \rangle \\ = \langle \boldsymbol{f}_f, \boldsymbol{v} \rangle \quad \forall \boldsymbol{v} \in \boldsymbol{X}_f, \\ b_f(\boldsymbol{u}, q) = 0 \quad \forall q \in Q, \\ \phi(\cdot, 0) = \phi_0 \quad \text{in } L^2(\Omega_p), \ \ \boldsymbol{u}(\cdot, 0) = \boldsymbol{u}_0 \quad \text{in } \boldsymbol{L}^2(\Omega_f). \end{cases} \tag{5.6}$$

By definitions in (5.4)-(5.5) and denoting $\boldsymbol{U_0} = (\phi_0, \boldsymbol{u}_0)^T$, (5.6) is equivalent to the following expression:

$$\begin{cases} \langle \dfrac{\partial \boldsymbol{U}}{\partial t}, \boldsymbol{V} \rangle + a(\boldsymbol{U}, \boldsymbol{V}) + b(\boldsymbol{V}, p) = \langle \boldsymbol{F}, \boldsymbol{V} \rangle \qquad \forall \boldsymbol{V} \in \boldsymbol{X}, \\ b(\boldsymbol{U}, q) = 0 \quad \forall q \in Q, \\ \boldsymbol{U}(\cdot, 0) = \boldsymbol{U_0} \quad \text{in } \boldsymbol{L}^2(\Omega). \end{cases} \tag{5.7}$$

An operator form of (5.7) can be written as:

$$\begin{cases} \dfrac{\partial \boldsymbol{U}}{\partial t} + A\boldsymbol{U} + B^* p = \boldsymbol{F} \quad \text{in } \boldsymbol{X}', \\ B\boldsymbol{U} = 0 \quad \text{in } Q', \\ \boldsymbol{U}(\cdot, 0) = \boldsymbol{U_0} \quad \text{in } \boldsymbol{L}^2(\Omega). \end{cases} \tag{5.8}$$

If we consider $B\boldsymbol{U} = 0 \in Q'$ above as a constraint and restrict the discussion in space $\boldsymbol{X}_{div}$, a more concise form of (5.8) is

$$\begin{cases} \dfrac{\partial \boldsymbol{U}}{\partial t} + A\boldsymbol{U} = \boldsymbol{F} \quad \text{in } \boldsymbol{X}'_{\text{div}}, \\ \boldsymbol{U}(\cdot, 0) = \boldsymbol{U_0} \quad \text{in } \boldsymbol{L}^2(\Omega). \end{cases} \tag{5.9}$$

Recall that (5.6), (5.7), (5.8), and (5.9) sit in an equivalent class, the well-posedness of each is further guaranteed by the continuous inf-sup condition [106]:

$$\inf_{0\neq q\in Q} \sup_{\mathbf{0}\neq \mathbf{V}\in \mathbf{X}} \frac{b\left(\boldsymbol{V}, q\right)}{\|q\|_Q \, \|\boldsymbol{V}\|_{\boldsymbol{X}}} \geq \beta, \quad \beta \text{ is a positive constant.} \tag{5.10}$$

For each $\boldsymbol{U_0} \in \boldsymbol{L}^2(\Omega)$ and $\boldsymbol{F} \in \boldsymbol{L}^2(0,T;\boldsymbol{X}')$, the coupled Stokes-Darcy system (5.8) admits a unique solution $(\boldsymbol{U}, p) \in \boldsymbol{L}^2(0,T;\boldsymbol{X}) \cap \boldsymbol{H}^1(0,T;\boldsymbol{X}') \times L^2(0,T;Q)$ (cf. [113]). Thus, we can use formulation (5.8) to define the operator $M : \boldsymbol{L}^2(0,T;\boldsymbol{X}) \cap \boldsymbol{H}^1(0,T;\boldsymbol{X}') \times L^2(0,T;Q) \times \boldsymbol{L}^2(\Omega) \to \boldsymbol{L}^2(0,T;\boldsymbol{X}') \times L^2(0,T;Q') \times \boldsymbol{L}^2(\Omega)$

$$M\begin{pmatrix} \boldsymbol{U} \\ p \\ \boldsymbol{U_0} \end{pmatrix} = \begin{pmatrix} \frac{\partial \boldsymbol{U}}{\partial t} + A\boldsymbol{U} + B^* p - \boldsymbol{F} \\ B\boldsymbol{U} \\ \boldsymbol{U}(\cdot, 0) - \boldsymbol{U_0} \end{pmatrix}.$$

With a simple calculus of variation, one can see that the Fréchet derivative operator $M'$ is a bijective mapping, the surjective of $M'$ is thereafter self-contained. These basic formulations and properties will allow us to use Lagrange multiplier rule later to find the optimal solution for cost functionals constrained by the Stokes-Darcy equation. In addition, throughout this Section, $C$, $C_i$, $C_i^j$ and $C_{i,j,k}$ are generic positive constants that are independent of the mesh parameter $h$ and the time step $\tau$, and are not necessarily the same at each occurrence.

## 5.2. MODELING THE DATA ASSIMILATION PROBLEM

**5.2.1. Mathematical Formulation and Wellposedness Analysis.** Let $\boldsymbol{Y}_{ad}$ be an admissible set for the initial value that could be either $\boldsymbol{L}^2(\Omega)$ or a closed convex subset of $\boldsymbol{L}^2(\Omega)$ in which we look for a solution to our data assimilation problem stated as: given $T > 0$, $\gamma > 0$, and the distributed observation $\widehat{\boldsymbol{U}} = (\widehat{\phi}, \widehat{\boldsymbol{u}})^T \in L^2(0,T;\boldsymbol{L}^2(\Omega))$, the variational data assimilation of the Stokes-Darcy model is:

$$\min_{\boldsymbol{U_0}\in \boldsymbol{Y}_{ad}} J(\boldsymbol{U_0}) = \frac{1}{2}\int_0^T \|\widehat{\boldsymbol{U}} - \boldsymbol{U}(\boldsymbol{U_0})\|_0^2 dt + \frac{\gamma}{2}\|\boldsymbol{U_0}\|_0^2 \quad \text{subject to (5.8)} \tag{5.11}$$

where the mapping $\boldsymbol{U}(\boldsymbol{U_0}) : \boldsymbol{L}^2(\Omega) \to \boldsymbol{L}^2(0,T;\boldsymbol{X}) \cap \boldsymbol{H}^1(0,T;\boldsymbol{X}')$ is defined as the solution of (5.8) with initial condition $\boldsymbol{U_0}$. The minimization of $\frac{1}{2}\int_0^T \|\widehat{\boldsymbol{U}} - \boldsymbol{U}(\boldsymbol{U_0})\|_0^2 dt$ in (5.11) is the primary goal, which drives the state variable $\boldsymbol{U}(\boldsymbol{U_0})$ close to the distributed observation $\widehat{\boldsymbol{U}}$ via adjusting the initial data $\boldsymbol{U_0}$. The second term $\frac{\gamma}{2}\|\boldsymbol{U_0}\|_0^2$ is a $\boldsymbol{L}^2$-Tikhonov regularization. The parameter $\gamma$ measures the relative importance of the minimization between terms $\int_0^T \|\widehat{\boldsymbol{U}} - \boldsymbol{U}(\boldsymbol{U_0})\|_0^2 dt$ and $\|\boldsymbol{U_0}\|_0^2$.

**Remark 14** *Problem* (5.11) *can also be written as:*

$$\min_{\boldsymbol{U_0}\in \boldsymbol{Y_{ad}}} J(\boldsymbol{U_0}) = \frac{1}{2}\int_0^T \|\hat{\boldsymbol{U}} - \boldsymbol{U}(\boldsymbol{U_0})\|_0^2 dt + \frac{\gamma}{2}\|\boldsymbol{U_0}\|_0^2 \tag{5.12}$$

*subject to*

$$\begin{cases} \dfrac{\partial \boldsymbol{U}}{\partial t} + A\boldsymbol{U} = \boldsymbol{F} \quad \text{in } \boldsymbol{X}'_{\text{div}}, \\ \boldsymbol{U}(\cdot,0) = \boldsymbol{U_0} \quad \text{in } \boldsymbol{L}^2(\Omega). \end{cases} \tag{5.13}$$

Provided that $\boldsymbol{F}$, $\widehat{\boldsymbol{U}} \in L^2(0,T;\boldsymbol{L}^2(\Omega))$ and $\partial\Omega$, $\Gamma$ are regular enough, we have the following wellposedeness results.

**Theorem 18** *There exists a unique solution $\boldsymbol{U_0^\star} \in \boldsymbol{Y}_{ad}$ for the data assimilation problem* (5.11)*. Furthermore, the solution $\boldsymbol{U_0^\star}$ can be characterized by*

$$\begin{aligned} J'(\boldsymbol{U_0^\star})(\boldsymbol{Z}_0 - \boldsymbol{U_0^\star}) = &\int_0^T \int_\Omega (\boldsymbol{U}(\boldsymbol{U_0^\star}) - \widehat{\boldsymbol{U}})(\boldsymbol{U}(\boldsymbol{Z}_0) - \boldsymbol{U}(\boldsymbol{U_0^\star}))dxdydt \\ &+ \gamma \int_\Omega \boldsymbol{U_0^\star}(\boldsymbol{Z}_0 - \boldsymbol{U_0^\star})dxdy \geq 0 \quad \forall \boldsymbol{Z}_0 \in \boldsymbol{Y}_{ad}. \end{aligned} \tag{5.14}$$

**Proof:**

Since $J(\boldsymbol{U_0})$ is nonnegative and thus bounded from below, the infimum exists. Let $\{\boldsymbol{U}_0^n\} \in \boldsymbol{Y}_{ad}$ be a minimizing sequence such that

$$J(\boldsymbol{U}_0^n) \longrightarrow \inf_{\boldsymbol{U_0} \in \boldsymbol{Y}_{ad}} J(\boldsymbol{U_0}).$$

The coercivity of $J(\boldsymbol{U_0})$ granted by term $\frac{\gamma}{2}\|\boldsymbol{U_0}\|^2$ leads to the $\boldsymbol{L}^2$ boundedness of the sequence $\{\boldsymbol{U}_0^n\}$. By the well-posedness results the Stokes-Darcy equation, $\{\boldsymbol{U}(\boldsymbol{U}_0^n)\}$ is bounded in $\boldsymbol{W}(0,T) = \boldsymbol{L}^2(0,T;\boldsymbol{X}_{\text{div}}) \cap \boldsymbol{H}^1(0,T;(\boldsymbol{X}_{\text{div}})')$. Since the closed convex set in $\boldsymbol{L}^2(\Omega)$ is weakly closed and Hilbert spaces are weakly compact, the Eberlin-Šmulian theorem implies there exists a pair of subsequence $(\{\boldsymbol{U}_0^{n^k}\}, \{\boldsymbol{U}(\boldsymbol{U}_0^{n^k})\})$ such that

$$\begin{aligned} \boldsymbol{U}_0^{n^k} &\to \boldsymbol{U_0^\star} \in \boldsymbol{Y}_{ad} \ \text{ weakly,} \\ \boldsymbol{U}(\boldsymbol{U}_0^{n^k}) &\to \boldsymbol{U}^\star \in \boldsymbol{L}^2(0,T;\boldsymbol{X}_{\text{div}}) \ \text{ weakly,} \\ \boldsymbol{U}(\boldsymbol{U}_0^{n^k}) &\to \boldsymbol{U}^\star \in \boldsymbol{H}^1(0,T;(\boldsymbol{X}_{\text{div}})') \ \text{ weakly.} \end{aligned} \tag{5.15}$$

The next step is to show $\boldsymbol{U}^\star = \boldsymbol{U}(\boldsymbol{U}_0^\star)$. First, the continuity of the bilinear form $a(\boldsymbol{U},\boldsymbol{V})$ leads to

$$\begin{aligned} &a(\boldsymbol{U}(\boldsymbol{U}_0^{n^k}),\boldsymbol{V}) = \langle A\boldsymbol{U}(\boldsymbol{U}_0^{n^k}),\boldsymbol{V}\rangle = \langle \boldsymbol{U}(\boldsymbol{U}_0^{n^k}), A^*\boldsymbol{V}\rangle \\ &\to \langle \boldsymbol{U}^\star, A^*\boldsymbol{V}\rangle = \langle A\boldsymbol{U}^\star,\boldsymbol{V}\rangle = a(\boldsymbol{U}^\star,\boldsymbol{V}) \end{aligned} \tag{5.16}$$

as $k \to \infty$, where $A^*$ is the adjoint operator of $A$.

Considering the weak formulation:

$$\begin{aligned} &\int_0^T \langle \frac{\partial \boldsymbol{U}(\boldsymbol{U}_0^{n^k})}{\partial t}, \boldsymbol{V}\rangle dt + \int_0^T a(\boldsymbol{U}(\boldsymbol{U}_0^{n^k}),\boldsymbol{V})dt + (\boldsymbol{U}(\boldsymbol{U}_0^{n^k})(\cdot,0),\boldsymbol{V}_0) \\ &= \int_0^T \langle \boldsymbol{F},\boldsymbol{V}\rangle dt + (\boldsymbol{U}_0^{n^k},\boldsymbol{V}_0) \quad \forall (\boldsymbol{V},\boldsymbol{V}_0) \in \boldsymbol{X}_{\text{div}} \times \boldsymbol{L}^2(\Omega). \end{aligned}$$

Based on the convergence results in (5.15) and (5.16), we have

$$\int_0^T \langle \frac{\partial \boldsymbol{U}^\star}{\partial t}, \boldsymbol{V}\rangle dt + \int_0^T a(\boldsymbol{U}^*, \boldsymbol{V}) dt + (\boldsymbol{U}^\star(\cdot, 0), \boldsymbol{V}_0) = \int_0^T \langle \boldsymbol{F}, \boldsymbol{V}\rangle dt + (\boldsymbol{U}_0^\star, \boldsymbol{V}_0),$$

which indicates that $\boldsymbol{U}^\star = \boldsymbol{U}(\boldsymbol{U}_0^\star)$ via the definition of $\boldsymbol{U}(\boldsymbol{U}_0^\star)$.

Now by the weakly lower semi-continuity of the cost functional $J(\boldsymbol{U}_0)$, we deduce

$$\begin{aligned} J(\boldsymbol{U}_{\boldsymbol{0}}^\star) &= \frac{\gamma}{2}\|\boldsymbol{U}_0^\star\|_0^2 + \frac{1}{2}\int_0^T \|\boldsymbol{U}(\boldsymbol{U}_{\boldsymbol{0}}^\star) - \widehat{\boldsymbol{U}}\|_0^2 dt \\ &\leq \liminf_{k\to\infty} J(\boldsymbol{U}_0^{n^k}) = \inf_{\boldsymbol{U}_0 \in \boldsymbol{Y}_{ad}} J(\boldsymbol{U}_0) \leq J(\boldsymbol{U}_0^\star). \end{aligned}$$

Hence,

$$J(\boldsymbol{U}_0^\star) = \inf_{\boldsymbol{U}_0 \in \boldsymbol{Y}_{ad}} J(\boldsymbol{U}_0),$$

where $\boldsymbol{U}_0^\star$ is the minimizer we need.

By the linear property of the Stokes-Darcy model, one can find out that $F(\boldsymbol{U}_0)$ is Fréchet differentiable and its second order derivative can be calculated as follows:

$$J''(\boldsymbol{U}_0)(\boldsymbol{Z}_0, \boldsymbol{Z}_0) = \int_0^T \int_\Omega \boldsymbol{U}^2(\boldsymbol{Z}_0) dxdydt + \gamma \int_\Omega \boldsymbol{Z}_0^2 dxdy \geq \gamma \|\boldsymbol{Z}_0\|_0^2 \quad \forall \boldsymbol{Z}_0 \in \boldsymbol{Y}_{ad}.$$

Based on the standard argument for convex minimization we know the minimizer $\boldsymbol{U}_0^\star$ is unique. Further, $\boldsymbol{U}_0^\star$ can be characterized by

$$\begin{aligned} J'(\boldsymbol{U}_0^\star)(\boldsymbol{Z}_0 - \boldsymbol{U}_0^\star) = &\int_0^T \int_\Omega (\boldsymbol{U}(\boldsymbol{U}_0^\star) - \widehat{\boldsymbol{U}})(\boldsymbol{U}(\boldsymbol{Z}_0) - \boldsymbol{U}(\boldsymbol{U}_0^\star)) dxdydt \\ &+ \gamma \int_\Omega \boldsymbol{U}_0^\star (\boldsymbol{Z}_0 - \boldsymbol{U}_0^\star) dxdy \geq 0 \quad \forall \boldsymbol{Z}_0 \in \boldsymbol{Y}_{ad}. \end{aligned}$$

This finishes the proof.

Next, we show that the solution of problem (5.11) is stable regarding to the perturbation of the observational data $\widehat{\boldsymbol{U}}$ and the regularization parameter $\gamma$.

**Theorem 19** *The solution of problem* (5.11) *continuously depends on the observation data* $\widehat{\boldsymbol{U}}$ *and the parameter* $\gamma$.

**Proof:** Introducing perturbations $\epsilon_1 \in \mathbb{R}$ on $\gamma$ and $\boldsymbol{\epsilon}_2 \in L^2(0,T;\boldsymbol{L}^2(\Omega))$ on $\widehat{\boldsymbol{U}}$ respectively, and letting $\bar{\boldsymbol{U}}_0$ denote the perturbed optimal solution, we then have

$$\begin{aligned}&\int_0^T \int_\Omega (\boldsymbol{U}(\bar{\boldsymbol{U}}_0) - \widehat{\boldsymbol{U}} - \boldsymbol{\epsilon}_2)(\boldsymbol{U}(\boldsymbol{Z}_0) - \boldsymbol{U}(\bar{\boldsymbol{U}}_0))dxdydt \\ &+ (\gamma + \epsilon_1) \int_\Omega \bar{\boldsymbol{U}}_0(\boldsymbol{Z}_0 - \bar{\boldsymbol{U}}_0)dxdy \geq 0 \qquad \forall \boldsymbol{Z}_0 \in \boldsymbol{Y}_{ad}.\end{aligned} \tag{5.17}$$

Taking $\boldsymbol{Z}_0 = \boldsymbol{U}_0^\star$ in (5.17) and $\boldsymbol{Z}_0 = \bar{\boldsymbol{U}}_0$ in (5.14), we obtain

$$\begin{aligned}\int_0^T \int_\Omega (\boldsymbol{U}(\bar{\boldsymbol{U}}_0) - \widehat{\boldsymbol{U}} - \boldsymbol{\epsilon}_2)(\boldsymbol{U}(\boldsymbol{U}_0^\star) - \boldsymbol{U}(\bar{\boldsymbol{U}}_0))dxdydt& \\ + (\gamma + \epsilon_1) \int_\Omega \bar{\boldsymbol{U}}_0(\boldsymbol{U}_0^\star - \bar{\boldsymbol{U}}_0)dxdy \geq 0&, \\ \int_0^T \int_\Omega (\boldsymbol{U}(\boldsymbol{U}_0^\star) - \widehat{\boldsymbol{U}})(\boldsymbol{U}(\bar{\boldsymbol{U}}_0) - \boldsymbol{U}(\boldsymbol{U}_0^\star))dxdydt + \gamma \int_\Omega \boldsymbol{U}_0^\star(\bar{\boldsymbol{U}}_0 - \boldsymbol{U}_0^\star)dxdy \geq 0&.\end{aligned}$$

Adding the two inequalities together, we have

$$\begin{aligned}&\int_0^T \int_\Omega (\boldsymbol{U}(\boldsymbol{U}_0^\star) - \boldsymbol{U}(\bar{\boldsymbol{U}}_0))^2 dxdydt + (\gamma + \epsilon_1) \int_\Omega (\boldsymbol{U}_0^\star - \bar{\boldsymbol{U}}_0)^2 dxdy \\ &\leq \int_0^T \int_\Omega \boldsymbol{\epsilon}_2(\boldsymbol{U}(\bar{\boldsymbol{U}}_0) - \boldsymbol{U}(\boldsymbol{U}_0^\star))dxdydt + \epsilon_1 \int_\Omega \boldsymbol{U}_0^\star(\boldsymbol{U}_0^\star - \bar{\boldsymbol{U}}_0)dxdy.\end{aligned} \tag{5.18}$$

Applying the Cauchy-Schwarz and Young's inequalities for the right-hand side terms in (5.18), we have

$$\int_0^T \int_\Omega \epsilon_2(\boldsymbol{U}(\bar{\boldsymbol{U}}_0) - \boldsymbol{U}(\boldsymbol{U}_0^\star))dxdydt \leq \frac{1}{2}\int_0^T \int_\Omega (\boldsymbol{U}(\boldsymbol{U}_0^\star) - \boldsymbol{U}(\bar{\boldsymbol{U}}_0))^2 dxdydt \qquad (5.19)$$

$$+ \frac{1}{2}\|\boldsymbol{\epsilon}_2\|^2_{L^2(0,T;\boldsymbol{L}^2(\Omega))},$$

$$\epsilon_1 \int_\Omega \boldsymbol{U}_0(\boldsymbol{U}_0^\star - \bar{\boldsymbol{U}}_0)dxdy \leq \frac{|\epsilon_1|}{2}\|\boldsymbol{U}_0^\star\|_0^2 + \frac{|\epsilon_1|}{2}\int_\Omega (\boldsymbol{U}_0^\star - \bar{\boldsymbol{U}}_0)^2 dxdy. \qquad (5.20)$$

Combining (5.18)-(5.20) and setting $|\epsilon_1| \leq \frac{\gamma}{3}$, we obtain the inequality

$$\frac{1}{2}\int_0^T \int_\Omega (\boldsymbol{U}(\boldsymbol{U}_0^\star) - \boldsymbol{U}(\bar{\boldsymbol{U}}_0))^2 dxdydt + \frac{\gamma}{2}\int_\Omega (\boldsymbol{U}_0^\star - \bar{\boldsymbol{U}}_0)^2 dxdy \qquad (5.21)$$

$$\leq \frac{1}{2}\|\boldsymbol{\epsilon}_2\|^2_{L^2(0,T;\boldsymbol{L}^2(\Omega))} + \frac{|\epsilon_1|}{2}\|\boldsymbol{U}_0^\star\|_0^2, \qquad (5.22)$$

which implies that the solution of problem (5.11) continuously depends on the observational data $\widehat{\boldsymbol{U}}$ and the regularization parameter $\gamma$.

**Remark 15** *Continuing on* (5.18)*, a variant of treatment on the term* $\int_0^T \int_\Omega \boldsymbol{\epsilon}_2(\boldsymbol{U}(\bar{\boldsymbol{U}}_0) - \boldsymbol{U}(\boldsymbol{U}_0^\star))dxdydt$ *in* (5.18) *will produce a different stability estimation:*

$$\frac{\gamma}{2}\int_\Omega (\boldsymbol{U}_0^\star - \bar{\boldsymbol{U}}_0)^2 dxdy \leq \frac{1}{4}\|\boldsymbol{\epsilon}_2\|^2_{L^2(0,T;\boldsymbol{L}^2(\Omega))} + \frac{|\epsilon_1|}{2}\|\boldsymbol{U}_0^\star\|_0^2. \qquad (5.23)$$

*Inequality* (5.23) *offers more information about how parameter $\gamma$ affects the stability of solution. That is, small $\gamma$ will generate ill-conditioning for the data assimilation system. This is a similar argument of the Theorem 9 in Section 3.1.*

**5.2.2. First Order Optimality System.** To find out the unique optimal solution such that the objective functional (5.11) is minimized, we apply the Lagrange multiplier rule which is apparently available due to the property of the operator $M$ (surjective of $M'$)

shown in Section 5.1. The Lagrange functional is formed as:

$$\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{U}, p, \boldsymbol{U_0}) = \frac{1}{2}\int_0^T \|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_0^2 dt + \frac{\gamma}{2}\|\boldsymbol{U_0}\|_0^2 + \langle \boldsymbol{\lambda}, M(\boldsymbol{U}, p, \boldsymbol{U_0})^T \rangle, \tag{5.24}$$

where $\boldsymbol{\lambda} \in \boldsymbol{L}^2(0,T;\boldsymbol{X}'') \cap \boldsymbol{H}^1(0,T;\boldsymbol{X}') \times L^2(0,T;Q'') \times \boldsymbol{L}^2(\Omega)'$ is a Lagrange multiplier. Since Hilbert space is reflexive, $\boldsymbol{X} \times Q$ and $\boldsymbol{X}'' \times Q''$ are therefore isometric. The element in dual space of a Hilbert space can be identified by the element in the Hilbert space itself. Hence, using the definition of operator $M$, (5.24) then can be rewritten as

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{U^*}, p^*, \boldsymbol{U^*}(\cdot,0), \boldsymbol{U}, p, \boldsymbol{U_0}) \\
&= \frac{1}{2}\int_0^T \|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_0^2 dt + \frac{\gamma}{2}\|\boldsymbol{U_0}\|_0^2 + \int_0^T \langle \frac{\partial \boldsymbol{U}}{\partial t} + A\boldsymbol{U} + B^* p - \boldsymbol{F}, \boldsymbol{U^*} \rangle dt \\
&\quad + \int_0^T \langle B\boldsymbol{U}, p^* \rangle dt + (\boldsymbol{U}(\cdot,0) - \boldsymbol{U}_0, \boldsymbol{U^*}(\cdot,0)) \\
&= \frac{1}{2}\int_0^T \|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_0^2 dt + \frac{\gamma}{2}\|\boldsymbol{U_0}\|_0^2 + \int_0^T \langle \frac{\partial \boldsymbol{U}}{\partial t}, \boldsymbol{U^*} \rangle dt + \int_0^T a\left(\boldsymbol{U}, \boldsymbol{U^*}\right) dt \\
&\quad + \int_0^T b\left(\boldsymbol{U^*}, p\right) dt + \int_0^T b\left(\boldsymbol{U}, p^*\right) dt + (\boldsymbol{U}(\cdot,0) - \boldsymbol{U_0}, \boldsymbol{U^*}(\cdot,0)) \\
&\quad - \int_0^T \langle \boldsymbol{F}, \boldsymbol{U^*} \rangle dt.
\end{aligned} \tag{5.25}$$

Variations in the Lagrange multipliers $\boldsymbol{U^*}$, $p^*$ and $\boldsymbol{U^*}(\cdot,0)$ recover the constraint equation (5.8). Variations with respect to $\boldsymbol{U}$, $p$, and $\boldsymbol{U_0}$ yield

$$\begin{aligned}
&\int_0^T (\widehat{\boldsymbol{U}} - \boldsymbol{U}, -\boldsymbol{V}) dt + \int_0^T \langle \frac{\partial \boldsymbol{V}}{\partial t}, \boldsymbol{U^*} \rangle dt + \int_0^T a\left(\boldsymbol{V}, \boldsymbol{U^*}\right) dt + \int_0^T b\left(\boldsymbol{V}, p^*\right) dt \\
&+ (\boldsymbol{V}(\cdot,0), \boldsymbol{U^*}(\cdot,0)) = 0 \quad \forall \boldsymbol{V} \in \boldsymbol{L}^2(0,T;\boldsymbol{X}) \times \boldsymbol{H}^1(0,T;\boldsymbol{X}'), \\
&\int_0^T b\left(\boldsymbol{U^*}, q\right) dt = 0 \quad \forall q \in L^2(0,T;Q), \\
&\gamma(\boldsymbol{U_0}, \boldsymbol{Z}_0) - (\boldsymbol{Z}_0, \boldsymbol{U^*}(\cdot,0)) = 0 \quad \forall \boldsymbol{Z}_0 \in \boldsymbol{L}^2(\Omega).
\end{aligned} \tag{5.26}$$

Taking integration by parts with respect to time for $\int_0^T \langle \frac{\partial \boldsymbol{V}}{\partial t}, \boldsymbol{U^*} \rangle dt$ in the first equation of (5.26), we obtain

$$\begin{aligned} &\int_0^T (\widehat{\boldsymbol{U}} - \boldsymbol{U}, -\boldsymbol{V}) dt + (\boldsymbol{V}, \boldsymbol{U^*})|_0^T - \int_0^T \langle \frac{\partial \boldsymbol{U^*}}{\partial t}, \boldsymbol{V} \rangle dt + \int_0^T a\left(\boldsymbol{V}, \boldsymbol{U^*}\right) dt \\ &+ \int_0^T b\left(\boldsymbol{V}, p^*\right) dt + (\boldsymbol{V}(\cdot, 0), \boldsymbol{U^*}(\cdot, 0)) = 0. \end{aligned} \tag{5.27}$$

Choosing $\boldsymbol{U^*}(\cdot, T) = \boldsymbol{0}$ and simplifying (5.27), we have

$$\int_0^T (\widehat{\boldsymbol{U}} - \boldsymbol{U}, -\boldsymbol{V}) dt - \int_0^T \langle \frac{\partial \boldsymbol{U^*}}{\partial t}, \boldsymbol{V} \rangle dt + \int_0^T a^*\left(\boldsymbol{U^*}, \boldsymbol{V}\right) dt + \int_0^T b\left(\boldsymbol{V}, p^*\right) dt = 0, \tag{5.28}$$

where $a^*\left(\boldsymbol{U^*}, \boldsymbol{V}\right)$ is given as

$$\begin{aligned} a^*\left(\boldsymbol{U^*}, \boldsymbol{V}\right) =& 2\nu\left(\mathbb{D}\left(\boldsymbol{u^*}\right), \mathbb{D}\left(\boldsymbol{v}\right)\right)_{\Omega_f} + (K\nabla\phi^*, \nabla\psi)_{\Omega_p} + \langle g\boldsymbol{u^*}\cdot\boldsymbol{n}_f, \psi\rangle \\ &- \langle \phi^*, \boldsymbol{v}\cdot\boldsymbol{n}_f\rangle + \alpha\langle P_\tau \boldsymbol{u^*}, P_\tau \boldsymbol{v}\rangle + \alpha\langle P_\tau \boldsymbol{u^*}, P_\tau\left(\mathbb{K}\nabla\psi\right)\rangle. \end{aligned} \tag{5.29}$$

(5.29) is essentially a consequence of swapping terms related to $\boldsymbol{V}$ and $\boldsymbol{U^*}$ of $a\left(\boldsymbol{V}, \boldsymbol{U^*}\right)$.

Summarizing all operations from (5.24)-(5.29), the optimal solution $\boldsymbol{U}_0$ is attained by solving the following coupled equation systems in the weak form:

*the forward state equation*

$$\begin{cases} \langle \frac{\partial \boldsymbol{U}}{\partial t}, \boldsymbol{V} \rangle + a\left(\boldsymbol{U}, \boldsymbol{V}\right) + b\left(\boldsymbol{V}, p\right) = \langle \boldsymbol{F}, \boldsymbol{V} \rangle \quad \forall \boldsymbol{V} \in \boldsymbol{X}, \\ b\left(\boldsymbol{U}, q\right) = 0 \quad \forall q \in Q, \\ \boldsymbol{U}(\cdot, 0) = \boldsymbol{U}_0, \quad \boldsymbol{U_0} \in \boldsymbol{L}^2(\Omega), \end{cases} \tag{5.30}$$

*the backward adjoint equation*

$$\begin{cases} -\langle \dfrac{\partial \boldsymbol{U^*}}{\partial t}, \boldsymbol{V} \rangle + a^* \left( \boldsymbol{U^*}, \boldsymbol{V} \right) + b\left( \boldsymbol{V}, p^* \right) = (\widehat{\boldsymbol{U}} - \boldsymbol{U}, \boldsymbol{V}) \quad \forall \boldsymbol{V} \in \boldsymbol{X}, \\ b\left( \boldsymbol{U^*}, q \right) = 0 \quad \forall q \in Q, \\ \boldsymbol{U^*}(\cdot, T) = \boldsymbol{0}, \end{cases} \tag{5.31}$$

and *the optimality condition*

$$\boldsymbol{U_0} = \frac{1}{\gamma} \boldsymbol{U^*}(\cdot, 0). \tag{5.32}$$

Concretely, by the definition of bilinear forms $a(\cdot,\cdot)$, $a^*(\cdot,\cdot)$ and $b(\cdot,\cdot)$, (5.30)-(5.32) are equivalent to:

*the forward state equation*

$$\begin{cases} \langle \dfrac{\partial \phi}{\partial t}, \psi \rangle + a_p\left( \phi, \psi \right) - \langle \boldsymbol{u} \cdot \boldsymbol{n_f}, \psi \rangle = \langle f_p, \psi \rangle \\ \langle \dfrac{\partial \boldsymbol{u}}{\partial t}, \boldsymbol{v} \rangle + a_f\left( \boldsymbol{u}, \boldsymbol{v} \right) + b_f\left( \boldsymbol{v}, p \right) + \langle g\phi, \boldsymbol{v} \cdot \boldsymbol{n}_f \rangle + \alpha \langle P_\tau \left( \boldsymbol{u} + \mathbb{K} \nabla \phi \right), P_\tau \boldsymbol{v} \rangle = \langle \boldsymbol{f}_f, \boldsymbol{v} \rangle \\ b_f\left( \boldsymbol{u}, q \right) = 0 \\ \phi(\cdot, 0) = \phi_0 \quad \boldsymbol{u}(\cdot, 0) = \boldsymbol{u_0} \end{cases} \tag{5.33}$$

*the backward adjoint equation*

$$\begin{cases} -\langle \dfrac{\partial \phi^*}{\partial t}, \psi \rangle + a_p\left( \phi^*, \psi \right) + \langle g\boldsymbol{u^*} \cdot \boldsymbol{n_f}, \psi \rangle + \alpha \langle P_\tau \boldsymbol{u^*}, P_\tau \left( \mathbb{K} \nabla \psi \right) \rangle = (\hat{\phi} - \phi, \psi) \\ -\langle \dfrac{\partial \boldsymbol{u^*}}{\partial t}, \boldsymbol{v} \rangle + a_f\left( \boldsymbol{u^*}, \boldsymbol{v} \right) + b_f\left( \boldsymbol{v}, p^* \right) - \langle \phi^*, \boldsymbol{v} \cdot \boldsymbol{n}_f \rangle + \alpha \langle P_\tau \boldsymbol{u^*}, P_\tau \boldsymbol{v} \rangle = (\hat{\boldsymbol{u}} - \boldsymbol{u}, \boldsymbol{v}) \\ b_f\left( \boldsymbol{u^*}, q \right) = 0 \\ \phi^*(\cdot, T) = 0, \quad \boldsymbol{U^*}(\cdot, T) = \boldsymbol{0}, \end{cases} \tag{5.34}$$

and

$$\phi_0 = \frac{1}{\gamma}\phi^*(\cdot, 0), \quad \boldsymbol{u_0} = \frac{1}{\gamma}\boldsymbol{u^*}(\cdot, 0). \tag{5.35}$$

Moreover, note that $a(\boldsymbol{U}, \boldsymbol{V}) = \langle A\boldsymbol{U}, \boldsymbol{V}\rangle = \langle \boldsymbol{U}, A^*\boldsymbol{V}\rangle$, which gives $\langle A^*\boldsymbol{U}, \boldsymbol{V}\rangle = a^*(\boldsymbol{U}, \boldsymbol{V})$. Then (5.30)-(5.32) are also equivalent to:

*the forward state equation*

$$\begin{cases} \dfrac{\partial \boldsymbol{U}}{\partial t} + A\boldsymbol{U} + B^* p = \boldsymbol{F}, \\ B\boldsymbol{U} = 0, \\ \boldsymbol{U}(\cdot, 0) = \boldsymbol{U}_0, \end{cases} \tag{5.36}$$

*the backward adjoint equation*

$$\begin{cases} -\dfrac{\partial \boldsymbol{U^*}}{\partial t} + A^*\boldsymbol{U^*} + B^* p^* = \widehat{\boldsymbol{U}} - \boldsymbol{U}, \\ B\boldsymbol{U^*} = 0, \\ \boldsymbol{U^*}(\cdot, T) = \boldsymbol{0}, \end{cases} \tag{5.37}$$

and

$$\boldsymbol{U_0} = \frac{1}{\gamma}\boldsymbol{U^*}(\cdot, 0). \tag{5.38}$$

The coupled systems (5.30)-(5.32), (5.33)-(5.35) or (5.36)-(5.38) are the first order necessary optimality system. The minimization problem (5.11) is strictly convex, thus the first order necessary condition is also sufficient.

## 5.3. NUMERICAL APPROXIMATION

**5.3.1. Finite Element Approximation.** In this section, we propose a fully discrete approximation of the data assimilation problem (5.11), which is based on a finite element discretization in space and the backward Euler scheme in time.

For spatial discretization, we consider $\boldsymbol{X}^h = X_p^h \times \boldsymbol{X}_f^h$ and $Q^h$ being pairwise well-defined finite element subspaces of $\boldsymbol{X} = X_p \times \boldsymbol{X}_f$ and $Q$, respectively. These family of spaces are parameterized by the mesh size $h$ that tends to 0, and we assume these finite element spaces satisfy the inf-sup condition, i.e., there exists a positive constant $\beta$ such that

$$\inf_{0\neq q\in Q^h} \sup_{\mathbf{0\neq V\in X^h}} \frac{b(\boldsymbol{V}, q)}{\|q\|_Q \|\boldsymbol{V}\|_X} \geq \beta. \tag{5.39}$$

As usual, we also assume the following approximation properties: there exist constants $k$ and $C$, independent of $\psi$, $\boldsymbol{v}$, $q$ and $h$, such that

$$\inf_{\psi_h\in X_p^h} \|\psi - \psi_h\|_1 \leq Ch^m \|\psi\|_{m+1} \quad \forall \psi \in H^{m+1}(\Omega_p)\ 1 \leq m \leq k+1, \tag{5.40}$$

$$\inf_{\boldsymbol{v}_h\in \boldsymbol{X}_f^h} \|\boldsymbol{v} - \boldsymbol{v}_h\|_1 \leq Ch^m \|\boldsymbol{v}\|_{m+1} \quad \forall \boldsymbol{v} \in \boldsymbol{H}^{m+1}(\Omega_f), \quad 1 \leq m \leq k+1, \tag{5.41}$$

$$\inf_{q_h\in Q^h} \|q - q_h\|_0 \leq Ch^m \|q\|_m \quad \forall q \in H^m(\Omega_f), \quad 1 \leq m \leq k+1. \tag{5.42}$$

For the time discretization we uniformly construct a time grid $0 = t_0 < t_1 < \cdots < t_n < \cdots < t_N = T$ with time step $\tau = \frac{T}{N}$. Let $I_n = (t_{n-1}, t_n]$ denote the $n$-th sub-interval, we use the finite-dimensional space

$$\boldsymbol{X}_{\tau,h} = \{\boldsymbol{V} : [0,T] \to \boldsymbol{X}^h : \boldsymbol{V}|_{I_n} \in \boldsymbol{X}^h \quad \text{is constant in time}\}.$$

Let $\boldsymbol{V}_h^n$ be the value of $\boldsymbol{V}_h \in \boldsymbol{X}_{\tau,h}$ at $t_n$ and $\boldsymbol{X}_{\tau,h}^n$ be the restriction to $I_n$ of the functions in $\boldsymbol{X}_{\tau,h}$.

Given specific $h, \tau, \gamma > 0$ and an admissible set $\boldsymbol{Y}_{ad}^h = \boldsymbol{X}^h \cap \boldsymbol{Y}_{ad}$ for the possible initial values, the fully discrete approximation of problem (5.11) is stated as

$$\min_{\boldsymbol{U}_{0,h} \in \boldsymbol{Y}_{ad}^h} J_h(\boldsymbol{U}_{0,h}) = \frac{1}{2}\tau \sum_{n=1}^{N} \|\widehat{\boldsymbol{U}}^n - \boldsymbol{U}_h^n\|_0^2 + \frac{\gamma}{2}\|\boldsymbol{U}_{0,h}\|_0^2 \tag{5.43}$$

subject to

$$\begin{cases} \dfrac{\boldsymbol{U}_h^{n+1} - \boldsymbol{U}_h^n}{\tau} + A\boldsymbol{U}_h^{n+1} + B^* p_h^{n+1} = \boldsymbol{F}^{n+1} \quad \text{in} \quad (\boldsymbol{X}^h)', \\ B\boldsymbol{U}_h^{n+1} = 0 \quad \text{in} \quad (Q^h)', \\ \boldsymbol{U}_h^0 = \boldsymbol{U}_{0,h} \quad \text{in} \quad \boldsymbol{X}^h. \end{cases} \tag{5.44}$$

Similar to the proof for the well-posedness of the continuous data assimilation problem, one can prove the well-posedness of the fully discrete data assimilation problem (5.43)-(5.44).

**Theorem 20** *Given $\tau = \frac{T}{N}$ and mesh size $h$, for every fixed regularization parameter $\gamma$, there exists an unique optimal solution $\boldsymbol{U}_{0,h}^\star \in \boldsymbol{Y}_{ad}^h$ such that the cost functional* (5.44) *is minimized. The optimal solution continuously depends on the observation data $\widehat{\boldsymbol{U}}$ and the parameter $\gamma$.*

Furthermore, one can also observe that small $\gamma$ will reduce the stability of the discrete data assimilation problem.

We expect that the optimal discrete solution of (5.43)-(5.44) converges to the solution of (5.11). That is, given a fixed $\gamma$, $\boldsymbol{U}_{0,h}^\star \to \boldsymbol{U}_0^\star$ should be attained when the time step $\tau$ and finite element mesh size $h$ tend to 0. Essentially, this is not difficult to be observed by a weakly argument, plus using a $\boldsymbol{L}^2$ projection from $\boldsymbol{Y}_{ad}$ to the finite element space $\boldsymbol{X}^h$ (cf.[114])

**Theorem 21** *For a fixed regularization parameter $\gamma$, let $\{\boldsymbol{U}_{0,h}^\star\}_{h>0}$ be the corresponding sequence of minimizers of the discrete data assimilation problems* (5.43)-(5.44). *Then $\{\boldsymbol{U}_{0,h}^\star\}_{h>0}$ converges to the continuous optimal solution $\boldsymbol{U}_0^\star$ as $h \to 0$ and $\tau \to 0$.*

**5.3.2. Derivation of the Discrete Optimality System.** Similar to the continuous VDA problem, we can derive the discrete optimality system via the Lagrange multiplier technique for computing the optimal solution $\boldsymbol{U}_{0,h}$. We formulate the discrete Lagrange functional as:

$$\begin{aligned}
&\mathcal{L}(\bar{\boldsymbol{U}}_h, \bar{p}_h, \boldsymbol{U}_{0,h}, \bar{\boldsymbol{U}}_h^*, \bar{p}_h^*, \boldsymbol{U}_h^{*0}) \\
&= \frac{1}{2}\tau \sum_{n=1}^{N} \|\widehat{\boldsymbol{U}}^n - \boldsymbol{U}_h^n\|_0^2 + \frac{\gamma}{2}\|\boldsymbol{U}_{0,h}\|_0^2 + \tau \sum_{n=0}^{N-1} \langle B\boldsymbol{U}_h^{n+1}, p_h^{*n}\rangle \\
&\quad + \tau \sum_{n=0}^{N-1} \langle \frac{\boldsymbol{U}_h^{n+1} - \boldsymbol{U}_h^n}{\tau} + A\boldsymbol{U}_h^{n+1} + B^* p_h^{n+1} - \boldsymbol{F}_{n+1}, \boldsymbol{U}_h^{*n}\rangle \\
&\quad + (\boldsymbol{U}_h^0 - \boldsymbol{U}_{0,h}, \boldsymbol{U}_h^{*0}),
\end{aligned} \tag{5.45}$$

where $\bar{\boldsymbol{U}}_h = (\boldsymbol{U}_h^0, \boldsymbol{U}_h^1, \boldsymbol{U}_h^2, \ldots, \boldsymbol{U}_h^N)$, $\bar{\boldsymbol{U}}_h^* = (\boldsymbol{U}_h^{*1}, \boldsymbol{U}_h^{*2}, \ldots, \boldsymbol{U}_h^{*N-1})$, $\bar{p}_h = (p_h^1, p_h^2, p_h^3, .$ $.., p_h^N)$ and $\bar{p}_h^* = (p_h^{*0}, p_h^{*1}, p_h^{*2}, p_h^{*3}, \ldots, p_h^{*N-1})$. By a few manipulations on $\boldsymbol{U}_h^n$, $\boldsymbol{U}_h^{*n}$ and using the adjoint notation: $\langle A\boldsymbol{U}, \boldsymbol{V}\rangle = \langle \boldsymbol{U}, A^*\boldsymbol{V}\rangle$,we reorganize (5.45) as

$$\begin{aligned}
&\mathcal{L}(\bar{\boldsymbol{U}}_h, \bar{p}_h, \boldsymbol{U}_{0,h}, \bar{\boldsymbol{U}}_h^*, \bar{p}_h^*, \boldsymbol{U}_h^{*0}) \\
&= \frac{1}{2}\tau \sum_{n=1}^{N} \|\hat{\boldsymbol{U}}^n - \boldsymbol{U}_h^n\|_0^2 + \frac{\gamma}{2}\|\boldsymbol{U}_{0,h}\|_0^2 + \tau \sum_{n=0}^{N-1} \langle B\boldsymbol{U}_h^{n+1}, p_h^{*n}\rangle \\
&+ \tau \sum_{n=0}^{N-1} \langle B^* p_h^{n+1}, \boldsymbol{U}_h^{*n}\rangle + \tau \sum_{n=0}^{N-1} \langle A\boldsymbol{U}_h^{n+1}, \boldsymbol{U}_h^{*n}\rangle - \tau \sum_{n=0}^{N-1} \langle \boldsymbol{F}_{n+1}, \boldsymbol{U}_h^{*n}\rangle \\
&+ \tau \sum_{n=0}^{N-1} \langle \frac{\boldsymbol{U}_h^{n+1} - \boldsymbol{U}_h^n}{\tau}, \boldsymbol{U}_h^{*n}\rangle + (\boldsymbol{U}_h^N, \boldsymbol{U}_h^{*N}) - (\boldsymbol{U}_h^N, \boldsymbol{U}_h^{*N}) \\
&+ (\boldsymbol{U}_h^0 - \boldsymbol{U}_{0,h}, \boldsymbol{U}_h^{*0}) \\
&= \frac{1}{2}\tau \sum_{n=1}^{N} \|\hat{\boldsymbol{U}}^n - \boldsymbol{U}_h^n\|_0^2 + \frac{\gamma}{2}\|\boldsymbol{U}_{0,h}\|_0^2 + \tau \sum_{n=1}^{N} \langle B\boldsymbol{U}_h^{n}, p_h^{*n-1}\rangle \\
&+ \tau \sum_{n=1}^{N} \langle B^* p_h^{n}, \boldsymbol{U}_h^{*n-1}\rangle + \tau \sum_{n=1}^{N} \langle A^*\boldsymbol{U}_h^{*n-1}, \boldsymbol{U}_h^{n}\rangle - \tau \sum_{n=1}^{N} \langle \boldsymbol{F}_{n}, \boldsymbol{U}_h^{*n-1}\rangle \\
&+ \tau \sum_{n=1}^{N} \langle \frac{\boldsymbol{U}_h^{*n-1} - \boldsymbol{U}_h^{*n}}{\tau}, \boldsymbol{U}_h^{n}\rangle + (\boldsymbol{U}_h^N, \boldsymbol{U}_h^{*N}) - (\boldsymbol{U}_{0,h}, \boldsymbol{U}_h^{*0}).
\end{aligned} \tag{5.46}$$

Variations in the Lagrange multipliers $\bar{\boldsymbol{U}}_h^*$, $\bar{p}_h^*$ and $\boldsymbol{U}_h^{*0}$ recover the constraint equation (5.44). Variations with respect to $\boldsymbol{U}_{0,h}$, $\boldsymbol{U}_h^n$ and $p_h^n$ yield

$$\frac{\partial \mathcal{L}(\bar{\boldsymbol{U}}_h, \bar{p}_h, \boldsymbol{U}_{0,h}, \bar{\boldsymbol{U}}_h^*, \bar{p}_h^*, \boldsymbol{U}_h^{*0})}{\partial \boldsymbol{U}_{0,h}} \boldsymbol{Z}_0^h = (\gamma \boldsymbol{U}_{0,h}, \boldsymbol{Z}_0^h) - (\boldsymbol{U}_h^{*0}, \boldsymbol{Z}_0^h) = 0 \quad \forall \boldsymbol{Z}_0^h \in \boldsymbol{Y}_{ad}^h,$$

$$\frac{\partial \mathcal{L}(\bar{\boldsymbol{U}}_h, \bar{p}_h, \boldsymbol{U}_{0,h}, \bar{\boldsymbol{U}}_h^*, \bar{p}_h^*, \boldsymbol{U}_h^{*0})}{\partial \boldsymbol{U}_h^n} \boldsymbol{V}_h = \tau \langle \frac{\boldsymbol{U}_h^{*n-1} - \boldsymbol{U}_h^{*n}}{\tau}, \boldsymbol{V}_h \rangle + \tau \langle A^* \boldsymbol{U}_h^{*n-1}, \boldsymbol{V}_h \rangle$$

$$+ \tau \langle B \boldsymbol{V}_h, p_h^{*n-1} \rangle - \tau (\widehat{\boldsymbol{U}}^n - \boldsymbol{U}_h^n, \boldsymbol{V}_h) = 0 \quad \forall \boldsymbol{V}_h \in \boldsymbol{X}^h, \ n = 1, \cdots, N-1,$$

$$\frac{\partial \mathcal{L}(\bar{\boldsymbol{U}}_h, \bar{p}_h, \boldsymbol{U}_{0,h}, \bar{\boldsymbol{U}}_h^*, \bar{p}_h^*, \boldsymbol{U}_h^{*0})}{\partial \boldsymbol{U}_h^N} \boldsymbol{V}_h = \tau \langle \frac{\boldsymbol{U}_h^{*N-1}}{\tau}, \boldsymbol{V}_h \rangle + \tau \langle A^* \boldsymbol{U}_h^{*N-1}, \boldsymbol{V}_h \rangle$$

$$+ \tau \langle B \boldsymbol{V}_h, p_h^{*N-1} \rangle - \tau (\hat{\boldsymbol{U}}^N - \boldsymbol{U}_h^N, \boldsymbol{V}_h) = 0 \quad \forall \boldsymbol{V}_h \in \boldsymbol{X}^h,$$

$$\frac{\partial \mathcal{L}(\bar{\boldsymbol{U}}_h, \bar{p}_h, \boldsymbol{U}_{0,h}, \bar{\boldsymbol{U}}_h^*, \bar{p}_h^*, \boldsymbol{U}_h^{*0})}{\partial p_h^n} q_h = \langle B^* q_h, \boldsymbol{U}_h^{*n-1} \rangle = 0 \quad \forall q_h \in Q^h, \quad n = 1, \cdots, N.$$

Using the fact $\langle B \boldsymbol{V}_h, p_h^{*n-1} \rangle = \langle B^* p_h^{*n-1}, \boldsymbol{V}_h \rangle$, we obtain the discrete optimality system, $n = 0, 1, 2, 3, \ldots, N-1$,

$$\begin{cases} \dfrac{\boldsymbol{U}_h^{n+1} - \boldsymbol{U}_h^n}{\tau} + A \boldsymbol{U}_h^{n+1} + B^* p_h^{n+1} = \boldsymbol{F}^{n+1}, \\ B \boldsymbol{U}_h^{n+1} = 0, \\ \boldsymbol{U}_h^0 = \boldsymbol{U}_{0,h}, \\ -\dfrac{\boldsymbol{U}_h^{*n+1} - \boldsymbol{U}_h^{*n}}{\tau} + A^* \boldsymbol{U}_h^{*n} + B^* p_h^{*n} = \widehat{\boldsymbol{U}}^{n+1} - \boldsymbol{U}_h^{n+1}, \\ B \boldsymbol{U}_h^{*n} = 0, \\ \boldsymbol{U}_h^{*N} = 0, \\ \boldsymbol{U}_{0,h} = \dfrac{1}{\gamma} \boldsymbol{U}_h^{*0}. \end{cases} \tag{5.47}$$

**Remark 16** *One may observe that the discrete optimality system* (5.47) *is the same as the direct full discretization of* (5.33)-(5.35). *This is because of the special symmetric property of Euler's scheme, including the explicit Euler scheme. However, such coincidence may not happen for other temporal discretization schemes, such as the Crank-Nicolson and most of the Runge-Kutta methods.*

**5.3.3. Finite Element Convergence Analysis.** In addition to the general convergence result in (21), one may be more interested in how the convergence behaves in practical simulations since it will help us properly set up discretization parameters for different scenarios. In the rest of this section, we focus on proving that, under enough smoothness assumption on $\boldsymbol{U}_0, \boldsymbol{U}$, and $\boldsymbol{U}^*$, the optimal finite element convergence rate is preserved for each of them.

Recall that the discrete optimality system (5.47) coincides with the direct full discretization of (5.33)-(5.35) in sense of the operator form, see (5.36)-(5.38). (5.47) hereby shares lots of similarities with the discretization of classical PDEs except only a few special terms. Therefore, instead of directly investigating the error equation between (5.47) and (5.33)-(5.35), we can utilize the FEM results from classical PDEs to study the convergence behavior in the data assimilation problem.

Before doing so, we need to rescale (5.33)-(5.35) such that the rescaled formulations possess crucial features for our analysis. The rescaling is achieved by multiplying the first equation in (5.33) and the second equation in (5.35) with $\eta$, respectively, the corresponding rescaled bilinear forms are as follows:

$$
\begin{aligned}
a_\eta\left(\boldsymbol{U}, \boldsymbol{V}\right) &= \eta a_f\left(\boldsymbol{u}, \boldsymbol{v}\right) + a_p\left(\phi, \psi\right) + \eta\langle g\phi, \boldsymbol{v}\cdot\boldsymbol{n}_f\rangle \\
&\quad - \langle \boldsymbol{u}\cdot\boldsymbol{n}_f, \psi\rangle + \eta\alpha\langle P_\tau\left(\boldsymbol{u} + \mathbb{K}\nabla\phi\right), P_\tau\boldsymbol{v}\rangle, \\
a_\eta^*\left(\boldsymbol{U}^*, \boldsymbol{V}\right) &= a_f\left(\boldsymbol{u}^*, \boldsymbol{v}\right) + \eta a_p\left(\phi^*, \psi\right) + \eta\langle g\boldsymbol{u}^*\cdot\boldsymbol{n}_f, \psi\rangle - \langle \phi^*, \boldsymbol{v}\cdot\boldsymbol{n}_f\rangle \\
&\quad + \alpha\langle P_\tau\boldsymbol{u}^*, P_\tau\boldsymbol{v}\rangle + \eta\alpha\langle P_\tau\boldsymbol{u}^*, P_\tau\left(\mathbb{K}\nabla\psi\right)\rangle.
\end{aligned}
$$

As stated in the following lemma, both $a_\eta\left(\boldsymbol{U},\boldsymbol{V}\right)$ and $a^*_\eta\left(\boldsymbol{U^*},\boldsymbol{V}\right)$ are coercive in the sense of a Garding type inequality, and this property will be frequently used in the convergence analysis.

**Lemma 6** *For appropriately chosen positive rescaling parameter $\eta$, there exist constants $C_{1,\eta}$, $C_{2,\eta}$, $C_{3,\eta}$, and $C_{4,\eta}$ such that $a_\eta\left(\boldsymbol{U},\boldsymbol{V}\right)$ and $a^*_\eta\left(\boldsymbol{U^*},\boldsymbol{V}\right)$ are coercive in sense of the Garding type inequality:*

$$a_\eta\left(\boldsymbol{U},\boldsymbol{U}\right)+C_{1,\eta}\|\boldsymbol{U}\|_0^2\geq C_{2,\eta}\|\boldsymbol{U}\|_{\boldsymbol{X}}^2, \tag{5.48}$$

$$a^*_\eta\left(\boldsymbol{U^*},\boldsymbol{U}^*\right)+C_{3,\eta}\|\boldsymbol{U}^*\|_0^2\geq C_{4,\eta}\|\boldsymbol{U}^*\|_{\boldsymbol{X}}^2. \tag{5.49}$$

**Proof:** We first prove the coercivity of the adjoint bilinear form $a^*_\eta\left(\boldsymbol{U^*},\boldsymbol{U^*}\right)$. According to Korn's, the Cauchy-Schwarz, Poincáre's, Young's and the trace inequalities, we deduce

$$\begin{aligned}
&a^*_\eta\left(\boldsymbol{U^*},\boldsymbol{U^*}\right)+C_{3,\eta}\|\boldsymbol{U^*}\|_0^2\\
&=C_{3,\eta}\|\boldsymbol{U^*}\|_0^2+2\nu\left(\mathbb{D}\left(\boldsymbol{U^*}\right),\mathbb{D}\left(\boldsymbol{U^*}\right)\right)_{\Omega_f}+\eta(\mathbb{K}\nabla\phi^*,\nabla\phi^*)_{\Omega_p}+\eta\langle g\boldsymbol{U^*}\cdot\boldsymbol{n}_f,\phi^*\rangle\\
&\quad-\langle\phi^*,\boldsymbol{U^*}\cdot\boldsymbol{n}_f\rangle+\alpha\langle P_\tau\boldsymbol{U^*},P_\tau\boldsymbol{U^*}\rangle+\eta\alpha\langle P_\tau\boldsymbol{U^*},P_\tau\left(\mathbb{K}\nabla\phi^*\right)\rangle\\
&\geq C_{3,\eta}\|\boldsymbol{U^*}\|_0^2+2\nu\|\mathbb{D}(\boldsymbol{U^*})\|_0^2+\eta\lambda_{\min}(\mathbb{K})\|\nabla\phi^*\|_0^2\\
&\quad-\eta gC_1\|\nabla\boldsymbol{U^*}\|_0^{\frac12}\|\boldsymbol{U^*}\|_0^{\frac12}\|\nabla\phi^*\|_0^{\frac12}\|\phi^*\|_0^{\frac12}+\alpha\|P_\tau\boldsymbol{U^*}\|^2_{\boldsymbol{L}^2(\Gamma)}\\
&\quad-C_1\|\nabla\boldsymbol{U^*}\|_0^{\frac12}\|\boldsymbol{U^*}\|_0^{\frac12}\|\nabla\phi^*\|_0^{\frac12}\|\phi^*\|_0^{\frac12}-\eta\alpha\lambda_{\max}(\mathbb{K})\|\nabla\boldsymbol{U^*}\|_0\|\nabla\phi^*\|_0\\
&\geq C_{3,\eta}\|\boldsymbol{U^*}\|_0^2+2C_0\nu\|\nabla\boldsymbol{U^*}\|_0^2+\eta\lambda_{\min}(\mathbb{K})\|\nabla\phi^*\|_0^2-\frac{\eta\lambda_{\min}(\mathbb{K})}{4}\|\nabla\phi^*\|_0^2\\
&\quad-\frac{(C_1(\eta g+1))^4}{4\eta\lambda_{\min}(\mathbb{K})}\|\phi^*\|_0^2-C_0\nu\|\nabla\boldsymbol{U^*}\|_0^2-\frac{1}{16C_0\nu}\|\boldsymbol{U^*}\|_0^2-\frac{\eta\alpha^2\lambda^2_{\max}(\mathbb{K})}{\lambda_{\min}(\mathbb{K})}\|\nabla\boldsymbol{U^*}\|_0^2\\
&\quad-\frac{\eta\lambda_{\min}(\mathbb{K})}{4}\|\nabla\phi^*\|_0^2\\
&=(C_{3,\eta}-\frac{1}{16C_0\nu})\|\boldsymbol{U^*}\|_0^2+(C_{3,\eta}-\frac{(C_1(\eta g+1))^4}{4\eta\lambda_{\min}(\mathbb{K})})\|\phi^*\|_0^2\\
&\quad+(C_0\nu-\frac{\eta\alpha^2\lambda^2_{\max}(\mathbb{K})}{\lambda_{\min}(\mathbb{K})})\|\nabla\boldsymbol{U^*}\|_0^2+\frac{\eta\lambda_{\min}(\mathbb{K})}{2}\|\nabla\phi^*\|_0^2,
\end{aligned}$$

where $C_i$ are generic constants depending on $\Omega$, or $\Gamma$, or both $\Omega$ and $\Gamma$, $\lambda_{\min}(\mathbb{K})$, $\lambda_{\max}(\mathbb{K})$ are the smallest and largest eigenvalues of matrix $\mathbb{K}$, and $\langle P_\tau \boldsymbol{u}, P_\tau(\mathbb{K}\nabla\phi)\rangle$ is understood as the duality between $\mathbf{H}^{\mathbf{1/2}}_{\mathbf{00}}(\Gamma)$ and $(\mathbf{H}^{\mathbf{1/2}}_{\mathbf{00}}(\Gamma))'$. In addition, the above boundedness of $-(\eta g+1)C_1\|\nabla\boldsymbol{U}^*\|_0^{\frac{1}{2}}\|\boldsymbol{U}^*\|_0^{\frac{1}{2}}\|\nabla\phi^*\|_0^{\frac{1}{2}}\|\phi^*\|_0^{\frac{1}{2}}$ is decomposed as follows:

$$
\begin{aligned}
&-(\eta g+1)C_1\|\nabla\boldsymbol{U}^*\|_0^{\frac{1}{2}}\|\boldsymbol{U}^*\|_0^{\frac{1}{2}}\|\nabla\phi^*\|_0^{\frac{1}{2}}\|\phi^*\|_0^{\frac{1}{2}}\\
&\geq -\frac{\|\nabla\boldsymbol{U}^*\|_0\|\boldsymbol{U}^*\|_0}{2}-\frac{((\eta g+1)C_1)^2\,\|\nabla\phi^*\|_0\|\phi^*\|_0}{2}\\
&\geq -C_0\nu\|\nabla\boldsymbol{U}^*\|_0^2-\frac{1}{16C_0\nu}\|\boldsymbol{U}^*\|_0^2-\frac{\eta\lambda_{\min}(\mathbb{K})}{4}\|\nabla\phi^*\|_0^2-\frac{(C_1(\eta g+1))^4}{4\eta\lambda_{\min}(\mathbb{K})}\|\phi^*\|_0^2.
\end{aligned}
$$

Once one chooses $\eta$ and $C_{3,\eta}$ satisfying

$$
\eta<\frac{C_0\nu\lambda_{\min}(\mathbb{K})}{\alpha^2\lambda^2_{\max}(\mathbb{K})}\ \text{ and }\ C_{3,\eta}\geq\max\{\frac{1}{16C_0\nu},\frac{(C_1(\eta g+1))^4}{4\eta\lambda_{\min}(\mathbb{K})}\},
$$

there exists a positive constant $C_{4,\eta}$ such that

$$
a^*_\eta\left(\boldsymbol{U}^*,\boldsymbol{U}^*\right)+C_{3,\eta}\|\boldsymbol{U}^*\|_0^2\geq C_{4,\eta}\|\boldsymbol{U}^*\|^2_{\boldsymbol{X}}.
$$

Moreover, proceeding argument similar to above, one can identify $\eta$ and $C_{1,\eta}$ satisfying $\eta<\frac{C_0\nu\lambda_{\min}(\mathbb{K})}{\alpha^2\lambda^2_{\max}(\mathbb{K})}$ and $C_{1,\eta}>\max\{\frac{(C_1(\eta g+1))^4}{8\eta C_0\nu},\frac{1}{8\lambda_{\min}(\mathbb{K})}\}$, then there exists a constant $C_{2,\eta}$ such that

$$
a_\eta\left(\boldsymbol{U},\boldsymbol{U}\right)+C_{1,\eta}\|\boldsymbol{U}\|_0^2\geq C_{2,\eta}\|\boldsymbol{U}\|^2_{\boldsymbol{X}}.
$$

The proof is complete by choosing $\eta<\min\{\frac{C_0\nu\lambda_{\min}(\mathbb{K})}{\alpha^2\lambda^2_{\max}(\mathbb{K})},\frac{C_0\nu\lambda_{\min}(\mathbb{K})}{\alpha^2\lambda^2_{\max}(\mathbb{K})}\}$.

The following lemma is for the continuity of $a_\eta\left(\boldsymbol{U},\boldsymbol{V}\right)$ and $a^*_\eta\left(\boldsymbol{U}^*,\boldsymbol{V}\right)$, which follow naturally from a group of standard inequalities, such as the trace, Korn's, the Cauchy-Schwarz and Poincáre's inequalities.

**Lemma 7** *$a_\eta(\boldsymbol{U},\boldsymbol{V})$ and $a^*_\eta(\boldsymbol{U^*},\boldsymbol{V})$ are continuous, i.e., there exist constants $C$ depending on $\Omega$, $\Gamma$, $\eta$, $g$, $\alpha$, $\mathbb{K}$ such that*

$$a_\eta(\boldsymbol{U},\boldsymbol{V}) \le C\|\boldsymbol{U}\|_X\|\boldsymbol{V}\|_X, \tag{5.50}$$

$$a^*_\eta(\boldsymbol{U}^*,\boldsymbol{V}) \le C\|\boldsymbol{U^*}\|_X\|\boldsymbol{V}\|_X. \tag{5.51}$$

**Proof:** We provide a sketch proof for the adjoint bilinear form $a^*_\eta(\boldsymbol{U}^*,\boldsymbol{V})$, since the analysis of the others is very similar.

$$\begin{aligned}
a^*_\eta(\boldsymbol{U}^*,\boldsymbol{V}) \le\ & 2C_3\nu\|\boldsymbol{u}^*\|_{X_f}\|\boldsymbol{v}\|_{X_f} + C_7\eta\lambda_{\max}(\mathbb{K})\|\phi^*\|_{X_p}\|\psi\|_{X_p} + C_4(1+\eta g)\|\boldsymbol{u}^*\|_{X_f}\|\psi\|_{X_p} \\
& + C_5\alpha\|\boldsymbol{u}^*\|_{X_f}\|\boldsymbol{v}\|_{X_f} + C_6\eta\alpha\lambda_{\max}(\mathbb{K})\|\boldsymbol{u}^*\|_{X_f}\|\psi\|_{X_p} \\
\le\ & C_{3,5}(\|\boldsymbol{u}^*\|^2_{\boldsymbol{X}_f} + \|\phi^*\|^2_{X_p})^{\frac{1}{2}}(\|\boldsymbol{v}\|^2_{\boldsymbol{X}_f} + \|\psi\|^2_{X_p})^{\frac{1}{2}} \\
& + C_{4,6}(\|\boldsymbol{u}^*\|^2_{\boldsymbol{X}_f} + \|\phi^*\|^2_{X_p})^{\frac{1}{2}}(\|\boldsymbol{v}\|^2_{\boldsymbol{X}_f} + \|\psi\|^2_{X_p})^{\frac{1}{2}} \\
& + C_7\eta\lambda_{\max}(\mathbb{K})(\|\boldsymbol{u}^*\|^2_{\boldsymbol{X}_f} + \|\phi^*\|^2_{X_p})^{\frac{1}{2}}(\|\boldsymbol{v}\|^2_{\boldsymbol{X}_f} + \|\psi\|^2_{X_p})^{\frac{1}{2}} \\
=\ & C\|\boldsymbol{U}^*\|_X\|\boldsymbol{V}\|_X,
\end{aligned}$$

where $C_i$ and $C_{i,j}$ are generic constants depending on $\Omega$, $\Gamma$, and

$$\begin{aligned}
C_{3,5} &= \max\{2C_3\nu, C_5\alpha\}, \quad C_{4,6} = \max\{C_4(1+\eta g), C_6\eta\alpha\lambda_{\max}(\mathbb{K})\} \\
C &= \max\{C_{3,5}, C_{4,6}, C_7\eta\lambda_{\max}(\mathbb{K})\}.
\end{aligned}$$

**Remark 17** *A consequence of Lemma 6 and Lemma 7 is the optimal FEM convergence of the backward adjoint equation equipped with a regular, non-variable force term. This can be shown by an extension of the proof in [105, Theorem 4.4].*

In the following, we will use the rescaled norms which are naturally defined as:

$$\|\boldsymbol{V}\|_{0,\eta} = (\eta\|\boldsymbol{v}\|_0^2 + \|\phi\|_0^2)^{\frac{1}{2}}, \;\; \|\boldsymbol{V}\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))} = \left(\int_0^T \|\boldsymbol{V}\|_{0,\eta}^2 dt\right)^{\frac{1}{2}},$$
$$\|\boldsymbol{V}\|_{0,\eta^*} = (\|\boldsymbol{v}\|_0^2 + \eta\|\phi\|_0^2)^{\frac{1}{2}}, \;\; \|\boldsymbol{V}\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))} = \left(\int_0^T \|\boldsymbol{V}\|_{0,\eta^*}^2 dt\right)^{\frac{1}{2}}.$$

By definition, one can easily establish the norm equivalences for $\|\cdot\|_0$, $\|\cdot\|_{0,\eta}$ and $\|\cdot\|_{0,\eta^*}$ stated in the following lemma.

**Lemma 8** *Norms $\|\cdot\|_0$, $\|\cdot\|_{0,\eta}$ and $\|\cdot\|_{0,\eta^*}$ are connected each other as:*

$$C_\eta^1\|\boldsymbol{U}\|_{0,\eta} \le \|\boldsymbol{U}\|_0 \le C_\eta^2\|\boldsymbol{U}\|_{0,\eta}, \tag{5.52}$$
$$C_\eta^1\|\boldsymbol{U}\|_{0,\eta^*} \le \|\boldsymbol{U}\|_0 \le C_\eta^2\|\boldsymbol{U}\|_{0,\eta^*}, \tag{5.53}$$
$$C_\eta^3\|\boldsymbol{U}\|_{0,\eta^*} \le \|\boldsymbol{U}\|_{0,\eta} \le C_\eta^4\|\boldsymbol{U}\|_{0,\eta^*}, \tag{5.54}$$

*where*

$$C_\eta^1 = \min\{1, \frac{1}{\sqrt{\eta}}\}, \; C_\eta^2 = \max\{1, \frac{1}{\sqrt{\eta}}\}, \; C_\eta^3 = \min\{\sqrt{\eta}, \frac{1}{\sqrt{\eta}}\}, \; C_\eta^4 = \max\{\sqrt{\eta}, \frac{1}{\sqrt{\eta}}\}.$$

Define notations

$$\langle\frac{\partial \boldsymbol{U}}{\partial t}, \boldsymbol{V}\rangle_\eta = \langle\frac{\partial \phi}{\partial t}, \psi\rangle + \eta\langle\frac{\partial \boldsymbol{u}}{\partial t}, \boldsymbol{v}\rangle, \; \langle\boldsymbol{F}, \boldsymbol{V}\rangle_\eta = \langle f_p, \psi\rangle + \eta\langle\boldsymbol{f}_f, \boldsymbol{v}\rangle,$$
$$\langle\frac{\partial \boldsymbol{U}}{\partial t}, \boldsymbol{V}\rangle_\eta^* = \eta\langle\frac{\partial \phi}{\partial t}, \psi\rangle + \langle\frac{\partial \boldsymbol{u}}{\partial t}, \boldsymbol{v}\rangle, \; \langle\boldsymbol{F}, \boldsymbol{V}\rangle_\eta^* = \eta\langle f_p, \psi\rangle + \langle\boldsymbol{f}_f, \boldsymbol{v}\rangle.$$

Then, using equivalent arguments similar to those used for (5.6), (5.7), (5.8), and (5.9), we can rewrite the continuous optimality system (5.33)-(5.35) as

$$\begin{cases} \langle\frac{\partial \boldsymbol{U}}{\partial t}, \boldsymbol{V}\rangle_\eta + a_\eta(\boldsymbol{U}, \boldsymbol{V}) = \langle\boldsymbol{F}, \boldsymbol{V}\rangle_\eta \quad \forall \boldsymbol{V} \in \boldsymbol{X}_{\text{div}}, \\ \boldsymbol{U}(\cdot, 0) = \boldsymbol{U_0} \quad \text{in} \;\; \boldsymbol{L}^2(\Omega), \end{cases} \tag{5.55}$$

$$\begin{cases} -\langle \frac{\partial \boldsymbol{U}^*}{\partial t}, \boldsymbol{V}\rangle_\eta^* + a_\eta^*\left(\boldsymbol{U}^*, \boldsymbol{V}\right) = \langle \widehat{\boldsymbol{U}} - \boldsymbol{U}, \boldsymbol{V}\rangle_\eta^* \quad \forall \boldsymbol{V} \in \boldsymbol{X}_{\text{div}}, \\ \boldsymbol{U}^*(\cdot, T) = \mathbf{0} \quad \text{in } \boldsymbol{L}^2(\Omega), \end{cases} \tag{5.56}$$

$$\boldsymbol{U}_{\mathbf{0}} = \frac{1}{\gamma}\boldsymbol{U}^*(\cdot, 0). \tag{5.57}$$

As mentioned previously, we intend to carry out the convergence analysis for the data assimilation problem by the finite element convergence results for classical PDEs. A key step is to introduce the following auxiliary equations:

$$\begin{cases} \langle \frac{\partial \boldsymbol{U}(\boldsymbol{U}_{0,h})}{\partial t}, \boldsymbol{V}\rangle_\eta + a_\eta\left(\boldsymbol{U}(\boldsymbol{U}_{0,h}), \boldsymbol{V}\right) = \langle \boldsymbol{F}, \boldsymbol{V}\rangle_\eta \quad \forall \boldsymbol{V} \in \boldsymbol{X}_{\text{div}}, \\ \boldsymbol{U}(\boldsymbol{U}_{0,h})(\cdot, 0) = \boldsymbol{U}_{0,h} \quad \text{in } \boldsymbol{L}^2(\Omega), \end{cases} \tag{5.58}$$

$$\begin{cases} -\langle \frac{\partial \boldsymbol{U}^*(\boldsymbol{U}_{0,h})}{\partial t}, \boldsymbol{V}\rangle_\eta^* + a_\eta^*\left(\boldsymbol{U}^*(\boldsymbol{U}_{0,h}), \boldsymbol{V}\right) = \langle \widehat{\boldsymbol{U}} - \boldsymbol{U}(\boldsymbol{U}_{0,h}), \boldsymbol{V}\rangle_\eta^* \quad \forall \boldsymbol{V} \in \boldsymbol{X}_{\text{div}}, \\ \boldsymbol{U}^*(\boldsymbol{U}_{0,h})(\cdot, T) = \mathbf{0} \quad \text{in } \boldsymbol{L}^2(\Omega), \end{cases} \tag{5.59}$$

$$\begin{cases} -\langle \frac{\partial \boldsymbol{U}^*(\boldsymbol{U}_h)}{\partial t}, \boldsymbol{V}\rangle_\eta^* + a_\eta^*\left(\boldsymbol{U}^*(\boldsymbol{U}_h), \boldsymbol{V}\right) = \langle \widehat{\boldsymbol{U}} - \boldsymbol{U}_h, \boldsymbol{V}\rangle_\eta^* \quad \forall \boldsymbol{V} \in \boldsymbol{X}_{\text{div}}, \\ \boldsymbol{U}^*(\boldsymbol{U}_h)(\cdot, T) = \mathbf{0} \quad \text{in } \boldsymbol{L}^2(\Omega). \end{cases} \tag{5.60}$$

Equations (5.58) and (5.59) are used to remove the concern from the initial condition in (5.55). Equation (5.60) basically recovers the Galerkin orthogonality we lost between the continuous and discrete adjoint equations 5.56 and (5.47). Analyzing these equations in pair,we can establish the inequalities stated in the following lemma.

**Lemma 9** *Let $\boldsymbol{U}(\boldsymbol{U}_{0,h}), \boldsymbol{U}^*(\boldsymbol{U}_{0,h}), \boldsymbol{U}^*(\boldsymbol{U}_h)$ be solutions of equations ( 5.58), (5.59) and (5.60), respectively, and let $(\boldsymbol{U}, \boldsymbol{U}^*, \boldsymbol{U}_{\mathbf{0}})$ and $(\boldsymbol{U}_h, \boldsymbol{U}_h^*, \boldsymbol{U}_{0,h})$ be solutions of the continuous and discrete optimality systems (5.55)-(5.57) and (5.47), then the following* *estimates hold*

$$\|\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h})\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))} \leq C_{\Omega,\Gamma,T}\|\boldsymbol{U_0} - \boldsymbol{U}_{0,h}\|_{0,\eta} \tag{5.61}$$

$$\|\boldsymbol{U}^* - \boldsymbol{U^*}(\boldsymbol{U}_{0,h})\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))} \leq C_{\Omega,\Gamma,T}\|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))} \tag{5.62}$$

$$\sup_{0\leq t\leq T}\|\boldsymbol{U}^*(\boldsymbol{U}_h) - \boldsymbol{U^*}(\boldsymbol{U}_{0,h})\|_{0,\eta^*} \leq C_{\Omega,\Gamma,T}\|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))}, \tag{5.63}$$

$$\|\boldsymbol{U}^*(\boldsymbol{U}_h) - \boldsymbol{U^*}(\boldsymbol{U}_{0,h})\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))} \leq C_{\Omega,\Gamma,T}\|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))}. \tag{5.64}$$

**Proof:** Subtracting (5.58) from (5.55), we have

$$\begin{cases} \langle\dfrac{\partial(\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}))}{\partial t}, \boldsymbol{V}\rangle_\eta + a_\eta\left(\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}), \boldsymbol{V}\right) = \langle\boldsymbol{0}, \boldsymbol{V}\rangle_\eta \quad \forall \boldsymbol{V} \in \boldsymbol{X}_{\text{div}}, \\ (\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}))(\cdot, 0) = \boldsymbol{U_0} - \boldsymbol{U}_{0,h} \quad \text{in } \boldsymbol{L}^2(\Omega). \end{cases} \tag{5.65}$$

Taking $\boldsymbol{V} = \boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h})$ on (5.65), using the coercive inequality (5.48) and norm relation (5.52), we obtain

$$\frac{d\|\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h})\|^2_{0,\eta}}{dt} + C_{2,\eta}\|\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h})\|^2_{\boldsymbol{X}} \leq C^1_\eta\|\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h})\|^2_{0,\eta}. \tag{5.66}$$

Applying the Gronwall inequality on (5.66) leads to

$$\|(\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}))(\cdot, t)\|^2_{0,\eta} + C_{\Omega,\Gamma,T}\int_0^t \|\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h})\|^2_{\boldsymbol{X}} \leq C_{\Omega,\Gamma,T}\|\boldsymbol{U_0} - \boldsymbol{U}_{0,h}\|^2_{0,\eta}. \tag{5.67}$$

Inequality (5.67) gives us

$$\|\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h})\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))} \leq C_{\Omega,\Gamma,T}\|\boldsymbol{U_0} - \boldsymbol{U}_{0,h}\|_{0,\eta}. \tag{5.68}$$

Again, we subtract (5.59) from (5.56) to obtain

$$\begin{cases} -\langle \dfrac{\partial(\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}))}{\partial t}, \boldsymbol{V}\rangle_\eta^* + a_\eta^*\left(\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}), \boldsymbol{V}\right) \\ = \langle \boldsymbol{U}(\boldsymbol{U}_{0,h} - \boldsymbol{U}), \boldsymbol{V}\rangle_\eta^* \quad \forall \boldsymbol{V} \in \boldsymbol{X}_{\text{div}}, \\ (\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}))(\cdot, T) = \boldsymbol{0} \quad \text{in } \boldsymbol{L}^2(\Omega). \end{cases} \tag{5.69}$$

Testing (5.69) with $\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})$ and using the coercive inequality (5.49), the Cauchy-Schwarz, Young's inequalities and (5.53)-(5.54), we deduce

$$\begin{aligned} &- \frac{d\|\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})\|_{0,\eta^*}^2}{dt} + C_{4,\eta}\|\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})\|_{\boldsymbol{X}}^2 \\ &\leq C_{3,\eta}\|\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})\|_0^2 + \|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}\|_{0,\eta^*}\|\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})\|_0 \\ &\leq (\frac{C_\eta^2}{2C_\eta^3} + C_{3,\eta}(C_\eta^2)^2)\|\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})\|_{0,\eta^*}^2 + \frac{C_\eta^2}{2C_\eta^3}\|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}\|_{0,\eta}^2. \end{aligned} \tag{5.70}$$

The Gronwall's inequality immediately implies

$$\|\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))} \leq C_{\Omega,\Gamma,T}\|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))}. \tag{5.71}$$

Finally, subtracting (5.60) from (5.56), we have

$$\begin{cases} -\langle \dfrac{\partial(\boldsymbol{U}^*(\boldsymbol{U}_h) - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}))}{\partial t}, \boldsymbol{V}\rangle_\eta^* + a_\eta^*\left(\boldsymbol{U}^*(\boldsymbol{U}_h) - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}), \boldsymbol{V}\right) \\ = \langle \boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}_h, \boldsymbol{V}\rangle_\eta^* \quad \forall \boldsymbol{V} \in \boldsymbol{X}_{\text{div}}, \\ (\boldsymbol{U}^*(\boldsymbol{U}_h) - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}))(\cdot, T) = \boldsymbol{0} \quad \text{in } \boldsymbol{L}^2(\Omega). \end{cases} \tag{5.72}$$

Similarly, by choosing appropriate test function in (5.72) and applying the coercive inequality (5.49), the norm relations (5.53)-(5.54), the Cauchy-Schwarz inequality, the Young's inequality, and the Gronwall's inequality, the following estimates hold

$$\sup_{0\le t\le T} \|\boldsymbol{U}^*(\boldsymbol{U}_h) - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})\|_{0,\eta^*} \le C_{\Omega,\Gamma,T}\|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))}, \tag{5.73}$$

$$\|\boldsymbol{U}^*(\boldsymbol{U}_h) - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))} \le C_{\Omega,\Gamma,T}\|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))}. \tag{5.74}$$

The proof is completed by putting (5.68), (5.71), (5.73) and (5.74) together.

By using the triangle inequality and inequality (5.61), $\|\boldsymbol{U} - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))}$ can be estimated as:

$$\begin{aligned}
&\|\boldsymbol{U} - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))}\\
&\le \|\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h})\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))} + \|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))}\\
&\le C_{\Omega,\Gamma,T}\|\boldsymbol{U}_0 - \boldsymbol{U}_{0,h}\|_{0,\eta} + \|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))}.
\end{aligned} \tag{5.75}$$

One also can bound $\|\boldsymbol{U}^* - \boldsymbol{U}^*_h\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))}$ using inequalities (5.62), (5.64) and (5.61),

$$\begin{aligned}
&\|\boldsymbol{U}^* - \boldsymbol{U}^*_h\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega)}\\
&\le \|\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))} + \|\boldsymbol{U}^*(\boldsymbol{U}_{0,h}) - \boldsymbol{U}^*(\boldsymbol{U}_h)\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))}\\
&\quad + \|\boldsymbol{U}^*(\boldsymbol{U}_h) - \boldsymbol{U}^*_h\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))}\\
&\le C_{\Omega,\Gamma,T}(\|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))} + \|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))})\\
&\quad + \|\boldsymbol{U}^*(\boldsymbol{U}_h) - \boldsymbol{U}^*_h\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))}\\
&\le C_{\Omega,\Gamma,T}(\|\boldsymbol{U}_0 - \boldsymbol{U}_{0,h}\|_{0,\eta} + \|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))})\\
&\quad + \|\boldsymbol{U}^*(\boldsymbol{U}_h) - \boldsymbol{U}^*_h\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))}.
\end{aligned} \tag{5.76}$$

Note that $\boldsymbol{U}_h$ and $\boldsymbol{U}_h^*$ are the classical finite element approximations of $\boldsymbol{U}(\boldsymbol{U}_{0,h})$ and $\boldsymbol{U}^*(\boldsymbol{U}_h)$, which is as desired. From (5.75) and (5.76), we observe that the bounds for the error in these classical finite element approximations depend on $\|\boldsymbol{U}_0 - \boldsymbol{U}_{0,h}\|_{0,\eta}$, which is estimated in the following lemma through two given equalities $\boldsymbol{U}_0 = \frac{1}{\gamma}\boldsymbol{U}^*(\cdot, 0)$ and $\boldsymbol{U}_{0,h} = \frac{1}{\gamma}\boldsymbol{U}_h^{*0}$.

**Lemma 10** *Let $\boldsymbol{U}_0$, $\boldsymbol{U}_{0,h}$, $\boldsymbol{U}_h^{*0}$ , $\boldsymbol{U}^*(\boldsymbol{U}_{0,h})(\cdot, 0)$ be functions defined in equations (5.33), (5.47), and (5.59), the following error estimate holds:*

$$\|\boldsymbol{U}_0 - \boldsymbol{U}_{0,h}\|_{0,\eta} \leq \frac{C_{\Omega,\Gamma}}{\gamma}\|\boldsymbol{U}^*(\boldsymbol{U}_{0,h})(\cdot, 0) - \boldsymbol{U}_h^{*0}\|_{0,\eta^*}. \tag{5.77}$$

**Proof:** Using $\boldsymbol{U}_0 = \frac{1}{\gamma}\boldsymbol{U}^*(\cdot, 0)$ and $\boldsymbol{U}_{0,h} = \frac{1}{\gamma}\boldsymbol{U}_h^{*0}$ we have

$$\begin{aligned}\|\boldsymbol{U}_0 - \boldsymbol{U}_{0,h}\|_0^2 &= \frac{1}{\gamma}(\boldsymbol{U}^*(\cdot, 0) - \boldsymbol{U}_h^{*0}, \boldsymbol{U}_0 - \boldsymbol{U}_{0,h}) \\ &= \frac{1}{\gamma}(\boldsymbol{U}^*(\cdot, 0) - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})(\cdot, 0), \boldsymbol{U}_0 - \boldsymbol{U}_{0,h}) \\ &\quad + \frac{1}{\gamma}(\boldsymbol{U}^*(\boldsymbol{U}_{0,h})(\cdot, 0) - \boldsymbol{U}_h^{*0}, \boldsymbol{U}_0 - \boldsymbol{U}_{0,h}).\end{aligned} \tag{5.78}$$

Taking $\boldsymbol{V} = \boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})$ on (5.65) without the scalar $\eta$ and integrating with respect to $t$, we obtain

$$\begin{aligned}&\int_0^T (\frac{\partial(\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}))}{\partial t}, \boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}))dt \\ &+ \int_0^T a(\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}), \boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}))dt = 0.\end{aligned} \tag{5.79}$$

Integration by parts with respect to $t$ on (5.79) results in

$$\begin{aligned}&((\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}))(\cdot, T), (\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}))(\cdot, T)) \\ &- ((\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}))(\cdot, 0), (\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}))(\cdot, 0)) \\ &- \int_0^T (\frac{\partial(\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}))}{\partial t}, \boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}))dt \\ &+ \int_0^T a(\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}), \boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}))dt = 0.\end{aligned} \tag{5.80}$$

Using equation (5.69) without the scalar $\eta$ and the fact $a(\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}), \boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})) = a^*(\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}), \boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}))$, we simplify the previous equation as

$$
\begin{aligned}
&((\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}))(\cdot,0), (\boldsymbol{U}^* - \boldsymbol{U}^*(\boldsymbol{U}_{0,h}))(\cdot,0)) \\
&= -\int_0^T (\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}), \boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}))dt.
\end{aligned}
$$

The nonnegativity of $\int_0^T (\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}), \boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U}_{0,h}))dt$ and the equality in (5.78) immediately show

$$
\|\boldsymbol{U}_0 - \boldsymbol{U}_{0,h}\|_{0,\eta} \le \frac{C_{\Omega,\Gamma}}{\gamma} \|\boldsymbol{U}^*(\boldsymbol{U}_{0,h})(\cdot,0) - \boldsymbol{U}_h^{*0}\|_{0,\eta^*},
$$

where $C_{\Omega,\Gamma} = \frac{C_\eta^2}{C_\eta^1}$, this is the fact using the norm relations (5.52) and (5.53).

Using (5.77) and the triangle inequality, $\|\boldsymbol{U}_0 - \boldsymbol{U}_{0,h}\|_{0,\eta}$ can be bounded as below

$$
\begin{aligned}
&\|\boldsymbol{U}_0 - \boldsymbol{U}_{0,h}\|_{0,\eta} \le \frac{C_{\Omega,\Gamma}}{\gamma} \|\boldsymbol{U}_h^{*0} - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})(\cdot,0)\|_{0,\eta^*} \\
&\le \frac{C_{\Omega,\Gamma}}{\gamma} \|\boldsymbol{U}_h^{*0} - \boldsymbol{U}^*(\boldsymbol{U}_h)(\cdot,0)\|_{0,\eta^*} + \frac{C_{\Omega,\Gamma}}{\gamma} \|\boldsymbol{U}^*(\boldsymbol{U}_h)(\cdot,0) - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})(\cdot,0)\|_{0,\eta^*} \\
&\le \frac{C_{\Omega,\Gamma}}{\gamma} \max_{0\le i\le N-1} \|\boldsymbol{U}_h^{*i} - \boldsymbol{U}^*(\boldsymbol{U}_h)(\cdot,t_i)\|_{0,\eta^*} + \frac{C_{\Omega,\Gamma}}{\gamma} \sup_{0\le t<T} \|\boldsymbol{U}^*(\boldsymbol{U}_h) - \boldsymbol{U}^*(\boldsymbol{U}_{0,h})\|_{0,\eta^*} \\
&\le \frac{C_{\Omega,\Gamma}}{\gamma} \max_{0\le i\le N-1} \|\boldsymbol{U}_h^{*i} - \boldsymbol{U}^*(\boldsymbol{U}_h)(\cdot,t_i)\|_{0,\eta^*} + \frac{C_{\Omega,\Gamma}}{\gamma} \|\boldsymbol{U}_h - \boldsymbol{U}(\boldsymbol{U}_{0,h})\|_{L^2(0,T;\boldsymbol{L}_\eta^2(\Omega))}.
\end{aligned}
\tag{5.81}
$$

Summarizing (5.75), (5.76), (5.81) and the classical FEM error estimates [105, Theorem 4.4] and (17), we finally arrive at the estimation

$$
\begin{aligned}
&\|\boldsymbol{U}_0 - \boldsymbol{U}_{0,h}\|_{0,\eta} + \|\boldsymbol{U} - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))} + \|\boldsymbol{U}^* - \boldsymbol{U}^*_h\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))} \\
&\le C_{\Omega,\Gamma,T}\|\boldsymbol{U}_0 - \boldsymbol{U}_{0,h}\|_{0,\eta} + \|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega)} \\
&\quad + \|\boldsymbol{U}^*(\boldsymbol{U}_h) - \boldsymbol{U}^*_h\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))} \\
&\le \frac{C_{\Omega,\Gamma,T}}{\gamma} \max_{0\le i\le N-1} \|\boldsymbol{U}^{*i}_h - \boldsymbol{U}^*(\boldsymbol{U}_h)(\cdot, t_i)\|_{0,\eta} + \frac{C_{\Omega,\Gamma,T}}{\gamma}\|\boldsymbol{U}_h - \boldsymbol{U}(\boldsymbol{U}_{0,h})\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))} \\
&\quad + \|\boldsymbol{U}(\boldsymbol{U}_{0,h}) - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega)} + \|\boldsymbol{U}^*(\boldsymbol{U}_h) - \boldsymbol{U}^*_h\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega)} \\
&\le C_{\gamma,\Omega,\Gamma,T}(h^{r+1} + \tau),
\end{aligned}
$$

where $r$ is the polynomial degree of the finite element basis function.

**Theorem 22** *Let* $(\boldsymbol{U}_0, \boldsymbol{U}, \boldsymbol{U}^*)$ *and* $(\boldsymbol{U}_{0,h}, \boldsymbol{U}_h, \boldsymbol{U}^*_h)$ *be solutions of the continuous optimality system* (5.33)-(5.35) *and discrete optimality system* (5.47) *respectively. Assuming the input data are smooth enough, then the following error estimate holds*

$$
\begin{aligned}
\|\boldsymbol{U}_0 - \boldsymbol{U}_{0,h}\|_{0,\eta} + \|\boldsymbol{U} - \boldsymbol{U}_h\|_{L^2(0,T;\boldsymbol{L}^2_\eta(\Omega))} + \|\boldsymbol{U}^* - \boldsymbol{U}^*_h\|_{L^2(0,T;\boldsymbol{L}^2_{\eta^*}(\Omega))} \\
\le C_{\gamma,\Omega,\Gamma,T}(h^{r+1} + \tau),
\end{aligned}
\tag{5.82}
$$

*where* $C_{\gamma,\Omega,\Gamma,T}$ *is a constant proportional to* $\frac{1}{\gamma}$ *and also depends on* $\Omega$*,* $\Gamma$ *and* $T$*.*

The inequality in (5.82) indicates that very small regularization parameter $\gamma$ may have a negative impact on the numerical accuracy. Therefore, in practice, more refined $h$ and $\tau$ are necessarily applied to offset the impact from a small $\gamma$.

## 5.4. ITERATIVE METHODS FOR SOLVING THE DISCRETE OPTIMALITY SYSTEM

Due to the complex structure of Stokes-Darcy model and the forward-backward coupled temporal nature in the optimality system, solving the (5.47) directly results in an extreme large coupled linear system [54], thereby being very computationally expensive. Hence we propose two iterative algorithms, the conjugate gradient method and the inexact line search steepest descent method, to decouple the discrete optimality system.

**5.4.1. Matrix Formulation.** For the description of the iterative methods, we introduce the matrix formulation for the fully discrete optimality system. By definition of the operators $A$, $A^*$, $B$, $B^*$, the discrete OptS (5.47) at each time step can be written as

$$\begin{cases} \frac{1}{\tau} M_a \left( \begin{pmatrix} \phi_h^{n+1} \\ \boldsymbol{u}_h^{n+1} \\ p_h^{n+1} \end{pmatrix} - \begin{pmatrix} \phi_h^{n} \\ \boldsymbol{u}_h^{n} \\ p_h^{n} \end{pmatrix} \right) + S \begin{pmatrix} \phi_h^{n+1} \\ \boldsymbol{u}_h^{n+1} \\ p_h^{n+1} \end{pmatrix} = \begin{pmatrix} f_{p,h}^{n+1} \\ \boldsymbol{f}_{f,h}^{n+1} \\ \vec{0} \end{pmatrix}, \quad \boldsymbol{U}_h^0 = \begin{pmatrix} \phi_h^0 \\ \boldsymbol{u}_h^0 \end{pmatrix}, \\ -\frac{1}{\tau} M_a \left( \begin{pmatrix} \phi_h^{*n+1} \\ \boldsymbol{u}_h^{*n+1} \\ p_h^{*n+1} \end{pmatrix} - \begin{pmatrix} \phi_h^{*n} \\ \boldsymbol{u}_h^{*n} \\ p_h^{*n} \end{pmatrix} \right) + S^* \begin{pmatrix} \phi_h^{*n} \\ \boldsymbol{u}_h^{*n} \\ p_h^{*n} \end{pmatrix} = \begin{pmatrix} \widehat{\phi}_h^{n+1} \\ \widehat{\boldsymbol{u}}_h^{n+1} \\ \vec{0} \end{pmatrix} - M_a \begin{pmatrix} \phi_h^{n+1} \\ \boldsymbol{u}_h^{n+1} \\ \vec{0} \end{pmatrix}, \\ \boldsymbol{U}_h^{*N} = \begin{pmatrix} \phi_h^{*N} \\ \boldsymbol{u}_h^{*N} \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \vec{\boldsymbol{0}} \end{pmatrix}, \quad \boldsymbol{U}_{0,h} = \frac{1}{\gamma} \begin{pmatrix} \phi_h^{*0} \\ \boldsymbol{u}_h^{*0} \end{pmatrix}, \end{cases} \tag{5.83}$$

where $M_a, S$, and $S^*$ in (5.83) are formulated as:

$$M_a = \begin{pmatrix} M_{a\phi} & 0 & 0 \\ 0 & M_{a\boldsymbol{u}} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad M_{a\boldsymbol{u}} = \begin{pmatrix} M_{au} & 0 \\ 0 & M_{au} \end{pmatrix},$$

$$S = \begin{pmatrix} S_{a\phi} & S_{\boldsymbol{u}\phi} & 0 \\ S_{\phi\boldsymbol{u}} & S_{a\boldsymbol{u}} + S_{\boldsymbol{u}\boldsymbol{u}} & S_{p\boldsymbol{u}} \\ 0 & S_{\boldsymbol{u}p} & 0 \end{pmatrix}, \quad S^* = \begin{pmatrix} S_{a\phi} & S^*_{\boldsymbol{u}\phi} & 0 \\ S^*_{\phi\boldsymbol{u}} & S_{a\boldsymbol{u}} + S^*_{\boldsymbol{u}\boldsymbol{u}} & S_{p\boldsymbol{u}} \\ 0 & S_{\boldsymbol{u}p} & 0 \end{pmatrix}.$$

Here, the related matrices $M_{a\phi}$, $M_{au}$, $S_{a\phi}$, $S_{a\boldsymbol{u}}$, $S_{\boldsymbol{u}\phi}$, $S_{\boldsymbol{u}\boldsymbol{u}}$, $S_{\phi\boldsymbol{u}}$, $S_{p\boldsymbol{u}}$, $S_{\boldsymbol{u}p}$, $S^*_{\boldsymbol{u}\phi}$, $S^*_{\boldsymbol{u}\boldsymbol{u}}$, $S^*_{\phi\boldsymbol{u}}$ and other vectors $f^{n+1}_{p,h}$, $\boldsymbol{f}^{n+1}_{f,h}$, $\widehat{\phi}^{n+1}_h$, and $\widehat{\boldsymbol{u}}^{n+1}_h$ are assembled as follows:

$$\begin{aligned}
M_{a\phi} &= \left[\int_{\Omega_p} \psi_j \psi_i dxdy\right], M_{au} = \left[\int_{\Omega_f} v_j v_i dxdy\right], S_{a\phi} = \left[\int_{\Omega_p} \mathbb{K}\nabla\psi_j \nabla\psi_i dxdy\right], \\
S_{p\boldsymbol{u}} &= \left[-\int_{\Omega_f} q_j \nabla\cdot \boldsymbol{v}_i dxdy\right], S_{\boldsymbol{u}p} = S^T_{p\boldsymbol{u}}, S_{a\boldsymbol{u}} = \left[\int_{\Omega_f} 2\nu \mathbb{D}\left(\boldsymbol{v}_j\right) : \mathbb{D}\left(\boldsymbol{v}_i\right) dxdy\right], \\
S_{\phi\boldsymbol{u}} &= \left[\int_{\Gamma} g\psi_j \boldsymbol{v}_i \cdot \boldsymbol{n}_f dS\right] + \left[\int_{\Gamma} \alpha P_\tau \left(\mathbb{K}\nabla\psi_j\right) P_\tau \boldsymbol{v}_i dS\right], S_{\boldsymbol{u}\phi} = -\left[\int_{\Gamma} \boldsymbol{v_j} \cdot \boldsymbol{n}_f \psi_i dS\right], \\
S_{\boldsymbol{u}\boldsymbol{u}} &= \left[\int_{\Gamma} \alpha P_\tau \boldsymbol{v}_j P_\tau \boldsymbol{v}_i dS\right], S^*_{\boldsymbol{u}\phi} = \left[\int_{\Gamma} g\boldsymbol{v}_j \cdot \boldsymbol{n}_f \psi_i dS\right] + \left[\int_{\Gamma} \alpha P_\tau \boldsymbol{v}_j P_\tau \left(\mathbb{K}\nabla\psi_i\right) dS\right], \\
S^*_{\boldsymbol{u}\boldsymbol{u}} &= \left[\int_{\Gamma} \alpha P_\tau \boldsymbol{v}_j P_\tau \boldsymbol{v}_i dS\right], S^*_{\phi\boldsymbol{u}} = -\left[\int_{\Gamma} \psi_j \boldsymbol{v}_i \cdot \boldsymbol{n}_f dS\right], \\
f^{n+1}_{p,h} &= \left[\int_{\Omega_p} f_p(t_{n+1})\psi_i dxdy\right], \boldsymbol{f}^{n+1}_{f,h} = \left[\int_{\Omega_f} \boldsymbol{f}_f(t_{n+1})\boldsymbol{v}_i dxdy\right], \\
\widehat{\phi}^{n+1}_h &= \left[\int_{\Omega_p} \widehat{\phi}(t_{n+1})\psi_i dxdy\right], \widehat{\boldsymbol{u}}^{n+1}_h = \left[\int_{\Omega_f} \widehat{\boldsymbol{u}}(t_{n+1})\boldsymbol{v}_i dxdy\right],
\end{aligned}$$

where $\{\psi_i\}$, $\{\boldsymbol{v}_i\} = (v_i, v_i)^T$ and $\{q_i\}$ are basis functions of the finite element spaces $X^h_p$, $\boldsymbol{X}^h_f$ and $Q^h$, respectively.

**5.4.2. The Conjugate Gradient Method.** Motivated by the fundamental conjugate gradient method in [54, 68], to efficiently solve 5.47 we propose the following iterative method to decouple the equation system: given $\boldsymbol{U}^{(0)}_{0,h}$, $\boldsymbol{U}^{(1)}_{0,h}$ and $\epsilon$, solve the following equations sequentially until the stop criteria $\|\boldsymbol{U}_{0,h}^{(i+1)} - \boldsymbol{U}_{0,h}^{(i)}\|_0 \leq \epsilon$ (or $\|\gamma \boldsymbol{U}_{0,h}^{(i+1)} - \boldsymbol{U}_h^{*0(i+1)}\|_0 \leq \epsilon$) is satisfied:

$$\begin{cases} \dfrac{M_a}{\tau}\left(\begin{pmatrix} \phi_h^{n+1(i)} \\ \boldsymbol{u}_h^{n+1(i)} \\ p_h^{n+1(i)} \end{pmatrix} - \begin{pmatrix} \phi_h^{n(i)} \\ \boldsymbol{u}_h^{n(i)} \\ p_h^{n(i)} \end{pmatrix}\right) + S\begin{pmatrix} \phi_h^{n+1(i)} \\ \boldsymbol{u}_h^{n+1(i)} \\ p_h^{n+1(i)} \end{pmatrix} = \begin{pmatrix} f_{p,h}^{n+1} \\ \boldsymbol{f}_{f,h}^{n+1} \\ \vec{0} \end{pmatrix}, \\ \boldsymbol{U}_h^{0(i)} = \begin{pmatrix} \phi_h^{0(i)} \\ \boldsymbol{u}_h^{0(i)} \end{pmatrix}, \end{cases} \tag{5.84}$$

$$\begin{cases} -\dfrac{M_a}{\tau}\left(\begin{pmatrix} \phi_h^{*n+1(i)} \\ \boldsymbol{u}_h^{*n+1(i)} \\ p_h^{*n+1(i)} \end{pmatrix} - \begin{pmatrix} \phi_h^{*n(i)} \\ \boldsymbol{u}_h^{*n(i)} \\ p_h^{*n(i)} \end{pmatrix}\right) + S^*\begin{pmatrix} \phi_h^{*n(i)} \\ \boldsymbol{u}_h^{*n(i)} \\ p_h^{*n(i)} \end{pmatrix} = \begin{pmatrix} \widehat{\phi}_h^{n+1} \\ \widehat{\boldsymbol{u}}_h^{n+1} \\ \vec{0} \end{pmatrix} - M_a\begin{pmatrix} \phi_h^{n+1(i)} \\ \boldsymbol{u}_h^{n+1(i)} \\ \vec{0} \end{pmatrix}, \\ \boldsymbol{U}_h^{*N(i)} = \begin{pmatrix} \phi_h^{*N(i)} \\ \boldsymbol{u}_h^{*N(i)} \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \vec{\boldsymbol{0}} \end{pmatrix}, \end{cases} \tag{5.85}$$

$$\boldsymbol{U}_h^{0(i+1)} = \boldsymbol{U}_h^{0(i)} + \zeta^{i+1}E^i(\boldsymbol{U}_h^{*0(i)} - \gamma\boldsymbol{U}_h^{0(i)}) + \eta^{i+1}C^i(\boldsymbol{U}_h^{0(i)} - \boldsymbol{U}_h^{0(i-1)}), \tag{5.86}$$

where $n = 0, 1, 2, 3, \ldots, N-1$ is the time moment, $i = 0, 1, 2, 3, \ldots$ is the iteration step, $\boldsymbol{U}_h^{0(i)}, \boldsymbol{U}_h^{n(i)}$, and $\boldsymbol{U}_h^{*n(i)}$ are iterative sequences, $E^i$ and $C^i$ are two symmetric positive definite matrices, $\zeta^{i+1}$ and $\eta^{i+1}$ are parameters updated at each iteration.

Following the ideas in [54, 55] we use $E^i$ and $C^i$ as identity matrices, $\zeta^{i+1}$ and $\eta^{i+1}$ are then updated as

$$\zeta^{i+1} = \frac{1}{q^{i+1}},\ \eta^{i+1} = \frac{e^i}{q^{i+1}}, \tag{5.87}$$

where

$$e^i = \begin{cases} 0 & i = 0, \\ q^i \frac{\|\boldsymbol{\chi}^i\|_0^2}{\|\boldsymbol{\chi}^{i-1}\|_0^2} & i > 0, \end{cases} \qquad q^{i+1} = \frac{\|\boldsymbol{\chi}^i\|_L^2}{\|\boldsymbol{\chi}^i\|_0^2} - e^i, \quad i = 0, 1, 2, 3, \cdots.$$

Here $\boldsymbol{\chi}^i = \gamma \boldsymbol{U}_h^{0(i)} - \boldsymbol{U}_h^{*0(i)}$ and $\|\boldsymbol{\chi}^i\|_L = (L\boldsymbol{\chi}^i, \boldsymbol{\chi}^i)_\Omega^{\frac{1}{2}}$. The operator $L$ acting on $\boldsymbol{\chi}^i$ is defined as follows

$$\begin{cases} \dfrac{M_a}{\tau}\left(\begin{pmatrix} \phi_{L,h}^{n+1} \\ \boldsymbol{u}_{L,h}^{n+1} \\ p_{L,h}^{n+1} \end{pmatrix} - \begin{pmatrix} \phi_{L,h}^{n} \\ \boldsymbol{u}_{L,h}^{n} \\ p_{L,h}^{n} \end{pmatrix}\right) + S \begin{pmatrix} \phi_{L,h}^{n+1} \\ \boldsymbol{u}_{L,h}^{n+1} \\ p_{L,h}^{n+1} \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \vec{0} \\ \vec{0} \end{pmatrix}, \\ \boldsymbol{U}_{L,h}^0 = \boldsymbol{\chi}^i, \end{cases} \tag{5.88}$$

$$\begin{cases} -\dfrac{M_a}{\tau}\left(\begin{pmatrix} \phi_{L,h}^{*n+1} \\ \boldsymbol{u}_{L,h}^{*n+1} \\ p_{L,h}^{*n+1} \end{pmatrix} - \begin{pmatrix} \phi_{L,h}^{*n} \\ \boldsymbol{u}_{L,h}^{*n} \\ p_{L,h}^{*n} \end{pmatrix}\right) + S^* \begin{pmatrix} \phi_{L,h}^{*n} \\ \boldsymbol{u}_{L,h}^{*n} \\ p_{L,h}^{*n} \end{pmatrix} = -M_a \begin{pmatrix} \phi_{L,h}^{n+1} \\ \boldsymbol{u}_{L,h}^{n+1} \\ \vec{0} \end{pmatrix}, \\ \boldsymbol{U}_{L,h}^{*N} = \begin{pmatrix} \vec{0} \\ \vec{\boldsymbol{0}} \end{pmatrix}, \end{cases} \tag{5.89}$$

$$L\boldsymbol{\chi}^i = \gamma \boldsymbol{\chi}^i - \begin{pmatrix} \phi_{L,h}^{*0} \\ \boldsymbol{u}_{L,h}^{*0} \end{pmatrix}. \tag{5.90}$$

We summarize the above iterative scheme as:

**Algorithm 12** *Step 0 (Initialization): Specify a convergence tolerance $\epsilon$, guess two initial functions $\boldsymbol{U}_h^{0(i)}$ and $\boldsymbol{U}_h^{1(i)}$, and then start the iteration at step $i = 1$.*

*Step 1 (Forward phase): Use $\boldsymbol{U}_h^{0(i)}$ as the initial condition to solve* (5.84) *forward for $\boldsymbol{U}_h^{n(i)}$, $n = 1, 2, 3, \cdots$ .*

*Step 2 (Backward phase): Pass $\boldsymbol{U}_h^{n(i)}$, $n = 1, 2, 3, \cdots$ to* (5.85) *and solve* (5.85) *backward for $\boldsymbol{U}_h^{*0(i)}$.*

*Step 3 (Computing operator L):*

*(1) Set* $\boldsymbol{\chi}^i = \gamma \boldsymbol{U}_h^{0(i)} - \boldsymbol{U}_h^{*0(i)}$ *as initial value to solve equation* (5.88) *forward to obtain* $\begin{pmatrix} \phi_{L,h}^n \\ \boldsymbol{u}_{L,h}^n \\ p_{L,h}^n \end{pmatrix}$*;*

*(2) Pass* $\begin{pmatrix} \phi_{L,h}^{*n} \\ \boldsymbol{u}_{L,h}^{*n} \\ p_{L,h}^{*n} \end{pmatrix}$ *to* (5.89) *and solve equation* (5.89) *backward for attaining* $\begin{pmatrix} \phi_{L,h}^{*0} \\ \boldsymbol{u}_{L,h}^{*0} \\ p_{L,h}^{*0} \end{pmatrix}$*;*

*(3) Compute* $L\boldsymbol{\chi}^i = \gamma \boldsymbol{\chi}^i - \begin{pmatrix} \phi_{L,h}^{*0} \\ \boldsymbol{u}_{L,h}^{*0} \end{pmatrix}$.

*Step 4 (Update phase): Calculate* $\zeta^{i+1}, \eta^{i+1}$ *by using* (5.87) *and then update*

$$\boldsymbol{U}_h^{0(i+1)} = \boldsymbol{U}_h^{0(i)} + \zeta^{i+1} E^i (\boldsymbol{U}_h^{*0(i)} - \gamma \boldsymbol{U}_h^{0(i)}) + \eta^{i+1} C^i (\boldsymbol{U}_h^{0(i)} - \boldsymbol{U}_h^{0(i-1)}).$$

*Step 5 (Criteria for stopping the iteration): Compute* $\|\boldsymbol{U}_h^{0(i+1)} - \boldsymbol{U}_h^{0(i)}\|_0$*. If* $\|\boldsymbol{U}_h^{0(i+1)} - \boldsymbol{U}_h^{0(i)}\|_0 \leq \epsilon$ *then stop. Otherwise go back to Step 1 and continue.*

**5.4.3. The Inexact Line Search Steepest descent Method.** The conjugate gradient method described above has a descent convergence rate and solves the discrete optimality system (5.47) effectively in most cases. However, its descent direction is sensitive to the stability of the data assimilation problem which can hinder the convergence of the conjugate gradient method for a problem with a low stability which might be caused by a small regularization parameter $\gamma$ in the cost functional (5.43).

This shortcoming motivates us to propose a steepest descent method in [52, 53] that gains more stability at the cost of a lower convergence rate. To begin with, we need to calculate the derivative of the cost functional (5.43) and find out its representation in the admissible set,

$$J_h'(\boldsymbol{U}_{0,h})\boldsymbol{Z}_h = \tau \sum_{n=1}^{N} (\widehat{\boldsymbol{U}}^n - \boldsymbol{U}_h^n, (\boldsymbol{U}_h^n)' \boldsymbol{Z}_h) + (\gamma \boldsymbol{U}_{0,h}, \boldsymbol{Z}_h) \quad \forall \boldsymbol{Z}_h \in \boldsymbol{Y}_{ad}^h. \tag{5.91}$$

Note that $(\boldsymbol{U}_h^n)'\boldsymbol{Z}_h$ is essentially equal to $\mathcal{U}_h^n$ which is the solution of the following discretized equation

$$\frac{\mathcal{U}_h^{n+1}-\mathcal{U}_h^n}{\tau}+A\mathcal{U}_h^{n+1}+B^*p_h^{n+1}=\mathbf{0},\quad B\mathcal{U}_h^{n+1}=0,\quad \mathcal{U}_h^0=\boldsymbol{Z}_h. \tag{5.92}$$

To compute $J_h'(\boldsymbol{U}_{0,h})\boldsymbol{Z}_h$, we introduce the adjoint variables $\begin{pmatrix}\boldsymbol{U}_h^{*n}\\ p_h^{*n}\end{pmatrix}_{n=0}^{N-1}=\begin{pmatrix}\phi_h^{*n}\\ \boldsymbol{u}_h^{*n}\\ p_h^{*n}\end{pmatrix}_{n=0}^{N-1}$ and let $\boldsymbol{U}_h^{*N}=\begin{pmatrix}\phi_h^N\\ \boldsymbol{u}_h^N\end{pmatrix}=0$, $\begin{pmatrix}\phi_h^{*n}\\ \boldsymbol{u}_h^{*n}\\ p_h^{*n}\end{pmatrix}_{n=1}^{N-1}$ is the discrete solution of equation:

$$-\frac{\boldsymbol{U}_h^{*n+1}-\boldsymbol{U}_h^{*n}}{\tau}+A^*\boldsymbol{U}_h^{*n}+B^*p_h^{*n}=\widehat{\boldsymbol{U}}^{n+1}-\boldsymbol{U}_h^{n+1},\quad B\boldsymbol{U}_h^{*n}=0,\quad \boldsymbol{U}_h^{*N}=0. \tag{5.93}$$

Proceeding the similar technique as (5.46), the derivative of the cost functional (5.44) is obtained as

$$J_h'(\boldsymbol{U}_{0,h})\boldsymbol{Z}_h=(\gamma\boldsymbol{U}_{0,h}-\boldsymbol{U}_h^{*0},\boldsymbol{Z}_h), \tag{5.94}$$

and $\gamma\boldsymbol{U}_{0,h}-\boldsymbol{U}_h^{*0}$ is the gradient of $J_h$ at $\boldsymbol{U}_{0,h}$.

With the gradient information in (5.94) we now present the steepest descent method to solve the discrete data assimilation problem: given $\boldsymbol{U}_h^{0(i)}$ and $\epsilon$, solve the following equations sequentially until the stop criteria $\|\boldsymbol{U}_h^{0(i+1)}-\boldsymbol{U}_h^{0(i)}\|_0\le\epsilon$ (or $\|\gamma\boldsymbol{U}_{0,h}^{(i+1)}-\boldsymbol{U}_h^{*0(i+1)}\|_0\le\epsilon$) is satisfied:

$$\begin{cases}\dfrac{M_a}{\tau}\left(\begin{pmatrix}\phi_h^{n+1(i)}\\ \boldsymbol{u}_h^{n+1(i)}\\ p_h^{n+1(i)}\end{pmatrix}-\begin{pmatrix}\phi_h^{n(i)}\\ \boldsymbol{u}_h^{n(i)}\\ p_h^{n(i)}\end{pmatrix}\right)+S\begin{pmatrix}\phi_h^{n+1(i)}\\ \boldsymbol{u}_h^{n+1(i)}\\ p_h^{n+1(i)}\end{pmatrix}=\begin{pmatrix}f_{p,h}^{n+1}\\ \boldsymbol{f}_{f,h}^{n+1}\\ \vec{0}\end{pmatrix},\\ \boldsymbol{U}_h^{0(i)}=\begin{pmatrix}\phi_h^{0(i)}\\ \boldsymbol{u}_h^{0(i)}\end{pmatrix},\end{cases} \tag{5.95}$$

$$\begin{cases} -\dfrac{M_a}{\tau}\left(\begin{pmatrix}\phi_h^{*n+1(i)}\\ \boldsymbol{u}_h^{*n+1(i)}\\ p_h^{*n+1(i)}\end{pmatrix}-\begin{pmatrix}\phi_h^{*n(i)}\\ \boldsymbol{u}_h^{*n(i)}\\ p_h^{*n(i)}\end{pmatrix}\right)+S^*\begin{pmatrix}\phi_h^{*n(i)}\\ \boldsymbol{u}_h^{*n(i)}\\ p_h^{*n(i)}\end{pmatrix}=\begin{pmatrix}\widehat{\phi}_h^{n+1}\\ \widehat{\boldsymbol{u}}_h^{n+1}\\ \vec{0}\end{pmatrix}-M_a\begin{pmatrix}\phi_h^{n+1(i)}\\ \boldsymbol{u}_h^{n+1(i)}\\ \vec{0}\end{pmatrix}, \\ \boldsymbol{U}_h^{*N(i)}=\begin{pmatrix}\phi_h^{*N(i)}\\ \boldsymbol{u}_h^{*N(i)}\end{pmatrix}=\begin{pmatrix}\vec{0}\\ \vec{\boldsymbol{0}}\end{pmatrix}, \end{cases} \tag{5.96}$$

$$\boldsymbol{U}_{0,h}^{(i+1)}=\boldsymbol{U}_{0,h}^{(i)}+\eta^{i+1}(\boldsymbol{U}_h^{*0(i)}-\gamma\boldsymbol{U}_h^{0(i)}), \tag{5.97}$$

where $n=0,1,2,3,\ldots,N-1$ is time moment, $i=0,1,2,3,\ldots$ is the iteration step, $\boldsymbol{U}_{0,h}^{(i)}$, $\boldsymbol{U}_h^{n(i)}$, $\boldsymbol{U}_h^{*n(i)}$ are iterative sequences, and $\eta^{i+1}$ is a constant called the learning rate.

To reduce the iterations and improve computational efficiency, the learning rate $\eta^{i+1}$ is determined by using the inexact line search algorithm: find $\eta^{i+1}$ via repeatedly solving (5.95) with initial value

$$\boldsymbol{U}_{0,h}^{(i+1)}=\boldsymbol{U}_{0,h}^{(i)}+\eta^{i+1}(\boldsymbol{U}_h^{*0(i)}-\gamma\boldsymbol{U}_{0,h}^{(i)}) \quad \text{by updating} \quad \eta^{i+1}=\rho\eta^{i+1},$$

until the following inequality is satisfied

$$J_h(\boldsymbol{U}_{0,h}^{(i+1)})\leq J_h(\boldsymbol{U}_{0,h}^{(i)})+\delta\eta^{i+1}\langle J_h'(\boldsymbol{U}_{0,h}^{(i)}),\boldsymbol{U}_h^{*0(i)}-\gamma\boldsymbol{U}_{0,h}^{(i)}\rangle, \tag{5.98}$$

where $\eta^{i+1}$ is typically initialized as a constant equal or greater than 1, $\delta$ and $\rho$ are chosen between $(0,1)$.

We summarize this inexact line search descent algorithm as follows:

**Algorithm 13** *Step 0 (Initialization): Specify a convergence tolerance $\epsilon$, guess initial function $\boldsymbol{U}_{0,h}^{(0)}$, and start the iteration step $i=1$.*

*Step 1 (Forward phase): Use $\boldsymbol{U}_{0,h}^{(i)}$ as initial condition to solve equation* (5.95) *forward for $\boldsymbol{U}_h^{n(i)}, n=1,2,3,\ldots,N$.*

*Step 2 (Backward phase): Pass* $\boldsymbol{U}_h^{n(i)}, n = 1, 2, 3, \ldots, N$ *to equation* (5.96) *and solve equation* (5.96) *backward for* $\boldsymbol{U}_h^{*0(i)}$.

*Step 3 (Inexact line search for* $\eta^{i+1}$*):*

*(1) Initialize a constant* $\eta^{i+1} \geq 1$*, set* $0 < \rho < 1$ *and* $0 < \delta < 1$*;*

*(2) Use* $\boldsymbol{U}_{0,h}^{(i+1)} = \boldsymbol{U}_{0,h}^{(i)} + \eta^{i+1}(\boldsymbol{U}_h^{*0(i)} - \gamma \boldsymbol{U}_{0,h}^{(i)})$ *as initial value to solve equation* (5.95) *forward to obtain* $\boldsymbol{U}_h^n$ *for computing* $F_h(\boldsymbol{U}_{0,h}^{(i+1)})$*;*

*(3) Update* $\eta^{i+1} = \rho\eta^{i+1}$ *until inequality* (5.98) *is satisfied;*

*(4) Output* $\eta^{i+1}$.

*Step 4 (Update phase): Use* $\eta^{i+1}$ *from Step 3 and then update*

$$\boldsymbol{U}_{0,h}^{(i+1)} = \boldsymbol{U}_{0,h}^{(i)} + \eta^{i+1}(\boldsymbol{U}_h^{*0(i)} - \gamma \boldsymbol{U}_{0,h}^{(i)}).$$

*Step 5 (Criteria for stopping the iteration): Compute* $\|\boldsymbol{U}_h^{*0(i)} - \gamma \boldsymbol{U}_{0,h}^{(i)}\|$*, if* $\|\boldsymbol{U}_h^{*0(i)} - \gamma \boldsymbol{U}_{0,h}^{(i)}\| \leq \epsilon$ *then stop; otherwise, go back to Step 1 and continue.*

**Remark 18** *If the admissible set is in a box constraint:* $\boldsymbol{Y}_{ad}^h = \{\boldsymbol{U}_{0,h} \in \boldsymbol{L}^2(\Omega) : \boldsymbol{a} \leq \boldsymbol{U}_{0,h} \leq \boldsymbol{b}\}$*, for both of the conjugate gradient and steepest descent methods, we need to project the* $\boldsymbol{U}_{0,h}^{(i+1)}$ *(at each iteration of the* $Update\ phase$*) onto* $\boldsymbol{Y}_{ad}^h$*, i.e., the update in* $Step$ 4 *is replaced as:* $\boldsymbol{U}_{0,h}^{(i+1)} = \max\{\boldsymbol{a}, \min\{\boldsymbol{b}, \boldsymbol{U}_{0,h}^{(i+1)}\}\}$*. This is then called the projected gradient method.*

**Remark 19 (Application of the Incremental POD)** *The incremental POD technique can be used in the Forward phase for both conjugate gradient method steepest descent method to save computer memory.*

## 5.5. NUMERICAL EXPERIMENTS

This section presents numerical results to demonstrate the optimal convergence established in Section 5.3 and the performance of the state prediction using the algorithms developed in Section 5.4. The Taylor-Hood finite element method is applied for the space discretization.

**5.5.1. Verification of the Finite Element Convergence Rate.** In this example, we let $\mathbb{K} = \mathbb{I}$, $\alpha = 1, g = 1$, $\Omega_p = (0,\pi) \times (0,1)$, $\Omega_f = (0,\pi) \times (-1,0)$, $\Gamma : x = 0$, and $\boldsymbol{U}|_{\partial\Omega} = 0$. Based on the numerical example in [115], whose analytic solutions satisfy the Beavers-Joseph interface conditions, we choose the following initial functions and source term functions:

$$
\begin{aligned}
\boldsymbol{W}_0 =& ((2-\pi\sin(\pi x))(-y+\cos(\pi(1-y))), x^2y^2+e^{-y}, (-2/3)xy^3+2-\pi\sin(\pi x))^T,\\
f_p =& \cos(2\pi t)(\pi^2(2\cos(\pi(1-y)) - 2\pi\sin(\pi x)\cos(\pi(1-y)) + \pi y\sin(\pi x)))\\
&- 2\pi\sin(2\pi t)(2-\pi\sin(\pi x))(-y+\cos(\pi(1-y))),\\
f_1 =& \cos(2\pi t)(-2y^2-2x^2-e^{-y}+\pi^2\cos(\pi x)\cos(2\pi y))\\
&- 2\pi\sin(2\pi t)(x^2y^2+e^{-y})\sin(2\pi t)(-2\pi),\\
f_2 =& \cos(2\pi t)(4xy-\pi^3\sin(\pi x)+2\pi(2-\pi\sin(\pi x))\sin(2\pi y))\\
&- 2\pi\sin(2\pi t)(\frac{2}{3}xy^3+2-\pi\sin(\pi x)).
\end{aligned}
$$

To construct a set of smooth observation data satisfying both the interface conditions and homogeneous boundary conditions, we numerically solve the Stokes-Darcy model with $h = 1/64$, $\tau = 1/4000$, initial function $\boldsymbol{W}_0$, and source term $\boldsymbol{F} = (f_p, f_1, f_2)^T$ in the time interval $[0, 0.75]$. Then the numerical solution in the time interval $[0.25, 0.75]$ is considered as the observation data $\widehat{\boldsymbol{U}}$.

Table 5.1. The finite element convergence rate of the recovered initial condition $\phi_0$ of the Stokes-Darcy equation.

| $\gamma$ | $\|\|\phi_0 - \phi_{0,\frac{1}{8}}\|\|_0$ | $\|\|\phi_0 - \phi_{0,\frac{1}{16}}\|\|_0$ | rate | $\|\|\phi_0 - \phi_{0,\frac{1}{32}}\|\|_0$ | rate |
|---|---|---|---|---|---|
| 1 | $1.16 \times 10^{-2}$ | $1.30 \times 10^{-3}$ | 3.15 | $1.20 \times 10^{-4}$ | 3.44 |
| $\frac{1}{5}$ | $4.19 \times 10^{-2}$ | $4.90 \times 10^{-3}$ | 3.09 | $4.70 \times 10^{-4}$ | 3.38 |
| $\frac{1}{50}$ | $9.83 \times 10^{-2}$ | $1.22 \times 10^{-2}$ | 3.01 | $1.30 \times 10^{-3}$ | 3.23 |
| $\frac{1}{200}$ | $1.14 \times 10^{-1}$ | $1.41 \times 10^{-2}$ | 3.02 | $1.50 \times 10^{-3}$ | 3.23 |

Table 5.2. The finite element convergence rate of the recovered initial condition $\boldsymbol{u}_0$ of the Stokes-Darcy equation.

| $\gamma$ | $\|\|\boldsymbol{u}_0 - \boldsymbol{u}_{0,\frac{1}{8}}\|\|_0$ | $\|\|\boldsymbol{u}_0 - \boldsymbol{u}_{0,\frac{1}{16}}\|\|_0$ | rate | $\|\|\boldsymbol{u}_0 - \boldsymbol{u}_{0,\frac{1}{32}}\|\|_0$ | rate |
|---|---|---|---|---|---|
| 1 | $2.10 \times 10^{-3}$ | $2.42 \times 10^{-3}$ | 3.12 | $2.91 \times 10^{-4}$ | 3.06 |
| $\frac{1}{5}$ | $1.03 \times 10^{-2}$ | $1.25 \times 10^{-3}$ | 3.04 | $1.41 \times 10^{-4}$ | 3.15 |
| $\frac{1}{50}$ | $4.94 \times 10^{-2}$ | $6.00 \times 10^{-2}$ | 3.04 | $6.64 \times 10^{-3}$ | 3.17 |
| $\frac{1}{200}$ | $6.69 \times 10^{-2}$ | $8.65 \times 10^{-2}$ | 2.95 | $8.80 \times 10^{-3}$ | 3.29 |

Table 5.3. Relative finite element errors according to $\gamma$ of the Stokes-Darcy equation. $R_h^{\phi_0} = \frac{\|\|\phi_0 - \phi_{0,h}\|\|_0}{\|\|\phi_0\|\|_0}$, $R_h^{\boldsymbol{u}_0} = \frac{\|\|\boldsymbol{u}_0 - \boldsymbol{u}_{0,h}\|\|_0}{\|\|\boldsymbol{u}_0\|\|_0}$.

| $\gamma$ | $R_{\frac{1}{8}}^{\phi_0}$ | $R_{\frac{1}{16}}^{\phi_0}$ | $R_{\frac{1}{32}}^{\phi_0}$ | $R_{\frac{1}{8}}^{\boldsymbol{u}_0}$ | $R_{\frac{1}{16}}^{\boldsymbol{u}_0}$ | $R_{\frac{1}{32}}^{\boldsymbol{u}_0}$ |
|---|---|---|---|---|---|---|
| 1 | 0.2184 | 0.0247 | 0.0023 | 0.0972 | 0.0119 | 0.0013 |
| $\frac{1}{5}$ | 0.2341 | 0.0278 | 0.0027 | 0.1167 | 0.0143 | 0.0016 |
| $\frac{1}{50}$ | 0.2496 | 0.0314 | 0.0033 | 0.1649 | 0.0208 | 0.0023 |
| $\frac{1}{200}$ | 0.2524 | 0.0313 | 0.0034 | 0.1811 | 0.0236 | 0.0027 |

For the data assimilation problem, we use the mesh sizes of $1/8$, $1/16$, $1/32$, $1/64$ and time step sizes of $1/16$, $1/128$, $1/1024$, $1/4000$ to produce numerical solutions, based on the conjugate gradient method. For each $\gamma$, the numerical solution with $h = 1/64, \tau = 1/4000$ is considered to replace the analytical solution when computing the numerical errors. Tables 5.1-5.3 illustrate the convergence performance. From tables 5.1 and 5.2, we can see that the $\boldsymbol{L}^2$ norm errors for $\phi$ and $\boldsymbol{U}$ appear to converge optimally. In addition, the relative errors in table 5.3 become larger when $\gamma$ decreases, which is consistent with the conclusion that the coefficient $C_{\gamma,\Omega,\Gamma,T}$ in Theorem 22 is proportional to $\frac{1}{\gamma}$.

**5.5.2. Data Assimilation Performance.** We now investigate the performance of the state forecast for Stokes-Darcy model utilizing the data assimilation methods developed in this Section. Let $\mathbb{K} = \mathbb{I}$, $\alpha = 1, g = 1$, $\Omega_p = (0,\pi) \times (0,1)$, $\Omega_f = (0,\pi) \times (-1,0)$, $\Gamma : x = 0, \boldsymbol{U}|_{\partial\Omega} = 0$, and

$$\boldsymbol{F} = (\pi \sin(x) + \cos(y) + 3/2,\ \ x^2 + y + \cos(y) + 1,\ \ \sin(y) + 2x + y + 2)^T,$$

$$\phi_0 = \begin{cases} (2 - \pi\sin(\pi x))(-y + \cos(\pi(1-y))) + y + \sin(x), y \geq \frac{1}{2}, \\ (2 - \pi\sin(\pi x))(-y + \cos(\pi(1-y))) + y + \cos(x), y \leq -\frac{1}{2}, \\ (2 - \pi\sin(\pi x))(-y + \cos(\pi(1-y))), \text{otherwise} , \end{cases}$$

$$u_{10} = \begin{cases} (x^2y^2 + exp(-y)) + 1 + \sin(y), y \geq \frac{1}{2}, \\ (x^2y^2 + exp(-y)) + \frac{1}{2} + \cos(y), y \leq -\frac{1}{2}, \\ (x^2y^2 + exp(-y)), \text{otherwise} , \end{cases}$$

$$u_{20} = \begin{cases} -\frac{2}{3}xy^3 + 2 - \pi\sin(\pi x) + \frac{1}{2} + \sin(y), y \geq \frac{1}{2}, \\ -\frac{2}{3}xy^3 + 2 - \pi\sin(\pi x) + 1 + \cos(y) + x, y \leq -\frac{1}{2}, \\ -\frac{2}{3}xy^3 + 2 - \pi\sin(\pi x), \text{otherwise} . \end{cases}$$

Set $h = 1/20$, $\tau = 1/100$. To construct a set of non-smooth observation data satisfying both the interface conditions and homogeneous boundary conditions, we numerically solve the Stokes-Darcy model with initial function $\boldsymbol{U}_0 = (\phi_0, u_{10}, u_{20})^T$ and source term $\boldsymbol{F} = (f_p, f_1, f_2)^T$ in the time interval $[0, 1]$. Then the numerical solution in the time interval $[1/20, 1]$ is considered as the observation data $\widehat{\boldsymbol{U}}_1$ without noise. The observations $\widehat{\boldsymbol{U}}_2$ with noise is produced by adding perturbation with normal distribution $N(0, \frac{1}{50})$ to $\widehat{\boldsymbol{U}}_1$. The $\widetilde{\boldsymbol{L}^2}$ and $\boldsymbol{L}^\infty$ norms, which are defined as $||\boldsymbol{U}||_{\widetilde{\boldsymbol{L}^2}} = (\sum_{n=1}^{N} \tau \|\boldsymbol{U}^n\|_0^2)^{\frac{1}{2}}$ and $||\boldsymbol{U}||_{\boldsymbol{L}^\infty} = \sup_{1 \leq n \leq N} \|\boldsymbol{U}^n\|_{\boldsymbol{L}^\infty(\Omega)}$, are used to measure the errors.

For observations without noise, table 5.4 shows the errors between the numerical solutions and the observation data $\widehat{\boldsymbol{U}}_1$ as well as the number of iteration steps. For observations with noise, table 5.5 shows the errors between the numerical solutions and the observation data $\widehat{\boldsymbol{U}}_2$ as well as the number of iteration steps. We can see that the errors between the numerical solutions and the observations become smaller when $\gamma$ becomes smaller. Furthermore, the convergence comparison between the conjugate gradient method and steepest descent method indicates that the conjugate gradient method is preferred for the moderate $\gamma$ because of its higher convergence rate, and the steepest descent method is a prior option for a small $\gamma$ due to its better stability. All of these agree well with our expectation and validate the methods proposed in this Section for solving the data assimilation problem of Stokes-Darcy model.

Table 5.4. Data assimilation result without noise for the Stokes-Darcy equation. The $\widetilde{\boldsymbol{L}^2}$- and $\boldsymbol{L}^\infty$- norm errors between $\widehat{\boldsymbol{U}}_1$ and the numerical solution $\boldsymbol{U}_h$, NI=Number of Iteration.

| | The conjugate gradient method | | | The steepest descent method | | |
|---|---|---|---|---|---|---|
| $\gamma$ | $\|\|\boldsymbol{U}_h - \widehat{\boldsymbol{U}}_1\|\|_{\widetilde{\boldsymbol{L}^2}}$ | $\|\|\boldsymbol{U}_h - \widehat{\boldsymbol{U}}_1\|\|_{L^\infty}$ | NI | $\|\|\boldsymbol{U}_h - \widehat{\boldsymbol{U}}_1\|\|_{\widetilde{\boldsymbol{L}^2}}$ | $\|\|\boldsymbol{U}_h - \widehat{\boldsymbol{U}}_1\|\|_{L^\infty}$ | NI |
| 1 | 0.2441 | 0.3764 | 6 | 0.2441 | 0.3764 | 6 |
| $\frac{1}{10}$ | 0.1725 | 0.2713 | 9 | 0.1725 | 0.2713 | 63 |
| $\frac{1}{200}$ | 0.0299 | 0.0519 | 18 | 0.0298 | 0.0517 | 432 |
| $\frac{1}{2000}$ | 0.0039 | 0.0058 | 49 | 0.0038 | 0.0057 | 713 |
| $\frac{1}{10000}$ | \ | \ | $\infty$ | 0.0022 | 0.0043 | 1135 |

Table 5.5. Data assimilation result with noise for the Stokes-Darcy equation. The $\widetilde{\boldsymbol{L}^2}$- and $\boldsymbol{L}^\infty$-norm errors between $\widehat{\boldsymbol{U}}_2$ and the numerical solution $\boldsymbol{U}_h$, NI=Number of Iteration.

| | The conjugate gradient method | | | The steepest descent method | | |
|---|---|---|---|---|---|---|
| $\gamma$ | $\|\|\boldsymbol{U}_h - \widehat{\boldsymbol{U}}_2\|\|_{\widetilde{\boldsymbol{L}^2}}$ | $\|\|\boldsymbol{U}_h - \widehat{\boldsymbol{U}}_2\|\|_{L^\infty}$ | NI | $\|\|\boldsymbol{U}_h - \widehat{\boldsymbol{U}}_2\|\|_{\widetilde{\boldsymbol{L}^2}}$ | $\|\|\boldsymbol{U}_h - \widehat{\boldsymbol{U}}_2\|\|_{L^\infty}$ | NI |
| 1 | 0.2503 | 0.3845 | 6 | 0.2503 | 0.3845 | 6 |
| $\frac{1}{10}$ | 0.1806 | 0.2812 | 9 | 0.1806 | 0.2812 | 63 |
| $\frac{1}{200}$ | 0.0432 | 0.0684 | 18 | 0.0431 | 0.0683 | 433 |
| $\frac{1}{2000}$ | 0.0198 | 0.0272 | 52 | 0.0196 | 0.0270 | 716 |
| $\frac{1}{10000}$ | \ | \ | $\infty$ | 0.0180 | 0.0230 | 1139 |

**5.5.3. Data Assimilation Results using Incremental POD.** Besides the classical gradient methods, we provide the data assimilation results using incremental POD data compression in this section. We use the same model parameters in Section 5.5.2 to generate the observations without introducing noise. The time step size and mesh size are given $\tau = 1/400$ and $h = 1/20$, respectively. The regularization parameter $\gamma$ is $1/10$.

Table 5.6 shows that the use of incremental POD saves computer storage around 90%, which effectively solves the memory issues in the gradient descent method from the data assimilation problem. In Table 5.7, simulation accuracy is not affected by using the approximated data when relatively small POD truncation is applied. This is might because the gradient method itself is correcting the information deviation to rule out the sacrifice from the incremental POD procedure. It might be also due to the total information loss from incremental POD in our numerical experiment is minimal. More investigations are deserved for this promising POD behavior in optimizations.

Table 5.6. Memory saved from the incremental POD with relatively large truncation for the data assimilation of the Stokes-Darcy equation. the Gradient convergence tolerance is $10^{-3}$, the POD truncation thresholds are all $10^{-10}$.

| Storage Saved During Gradient descent Iterations | | | |
|---|---|---|---|
| | Original Data size | Compressed Data size | Storage Saved |
| $\phi$ | $1681 \times 400$ | $1681 \times 21$ | 89% |
| $u_1$ | $1681 \times 400$ | $1681 \times 20$ | 90% |
| $u_2$ | $1681 \times 400$ | $1681 \times 20$ | 90% |

Table 5.7. Data assimilation comparison between the use and no use of the incremental POD for the data assimilation of the Stokes-Darcy equation. $D^O$: observations; $D_h$: numerical results without using POD; $D_h^C$ : numerical results using POD truncation thresholds $10^{-10}$.

| Error Comparison to Observations | | |
|---|---|---|
| | $\|\|D^O - D_h\|\|_{L^2(0,T;L^2(\Omega))}$ | $\|\|D^O - D_h^C\|\|_{L^2(0,T;L^2(\Omega))}$ |
| $\phi$ | 0.03827 | 0.03827 |
| $u_1$ | 0.05158 | 0.05158 |
| $u_2$ | 0.04234 | 0.04234 |

# 6. CONCLUSIONS

Based on the optimal control theory, we proposed the variational data assimilation method to improve the state prediction of a dynamical system with interface conditions. In this dissertation, we made contributions for dealing with such problem in multiple perspectives. First, up to our knowledge, this is one of the pioneer works on data assimilation for interface problems. Second, based on a weak interpretation of the dynamical system, we rigorously formulated the data assimilation into an optimization problem, and established the existence, uniqueness, and stability of the optimal solution. We derived the first order optimality system by dual method and Lagrange multiplier rule. Third, we present a fully discrete approximation of the continuous data assimilation with finite element methods, and demonstrated the optimal finite element convergence rate via employing skillful numerical techniques. In addition, besides the implementation of classical gradient descent methods, we develop the time parallel algorithm and proper orthogonal decomposition methods to optimize the computational resource during the data assimilation procedure.

The promising numerical performances encourage us to further investigate more realistic and complex data assimilation scenarios, such as the consideration for dual-porosity models, nonlinear governing systems, inhomogeneous Dirichlet boundary condition, inhomogeneous interface conditions, and sparse data simulation. Besides, we also keep eyes on parallel computing and POD methods, especially, the iterative nature of parallel computing may enable us to imbed the reduced basis method to greatly optimize the computational efficiency.

## REFERENCES

[1] F. A. Rihan, C. G. Collier, and I. Roulstone, "Four-dimensional variational data assimilation for Doppler radar wind data," *J. Comput. Appl. Math.*, vol. 176, no. 1, pp. 15–34, 2005.

[2] C. H. Bruneau, P. Fabrie, and F. Veersé, "Optimal control data assimilation with an atmospheric model," *Numer. Funct. Anal. Optim.*, vol. 18, no. 7-8, pp. 691–722, 1997.

[3] A. Brandt and L. Y. Zaslavsky, "Multiscale algorithm for atmospheric data assimilation," *SIAM J. Sci. Comput.*, vol. 18, no. 3, pp. 949–956, 1997.

[4] M. Tang, Y. Liu, and L. J. Durlofsky, "A deep-learning-based surrogate model for data assimilation in dynamic subsurface flow problems," *J. Comput. Phys.*, vol. 413, pp. 109456, 28, 2020.

[5] H. X. Vo and L. J. Durlofsky, "Data assimilation and uncertainty assessment for complex geological models using a new PCA-based parameterization," *Comput. Geosci.*, vol. 19, no. 4, pp. 747–767, 2015.

[6] M. Gesho, E. Olson, and E. S. Titi, "A computational study of a data assimilation algorithm for the two-dimensional Navier-Stokes equations," *Commun. Comput. Phys.*, vol. 19, no. 4, pp. 1094–1110, 2016.

[7] J. Blum, F.-X. L. Dime, and I. M. Navon, "Data assimilation for geophysical fluids. Handbook of numerical analysis. Vol. XIV. Special volume: computational methods for the atmosphere and the oceans," in *Handb. Numer. Anal., 14*, pp. 385–441, Elsevier\North-Holland, Amsterdam, 2009.

[8] B. García-Archilla, J. Novo, and E. S. Titi, "Uniform in time error estimates for a finite element method applied to a downscaling data assimilation algorithm for the Navier-Stokes equations," *SIAM J. Numer. Anal.*, vol. 58, no. 1, pp. 410–429, 2020.

[9] D. Rozier, E. Birol, E. Cosme, P. Brasseur, J. M. Brankart, and J. Verron, "A reduced-order Kalman filter for data assimilation in physical oceanography," *SIAM Rev.*, vol. 49, no. 3, pp. 449–465, 2007.

[10] V. I. Agoshkov, E. I. Parmuzin, and V. P. Shutyaev, "A numerical algorithm of variational data assimilation for reconstruction of salinity fluxes on the ocean surface," *Russian J. Numer. Anal. Math. Modelling*, vol. 23, no. 2, pp. 135–161, 2008.

[11] F.-X. Le Dimet, I. Souopgui, and H. E. Ngodock, "Sensitivity analysis applied to a variational data assimilation of a simulated pollution transport problem," *Internat. J. Numer. Methods Fluids*, vol. 83, no. 5, pp. 465–482, 2017.

[12] G. Triantafyllou, I. Hoteit, G. Korres, and G. Petihakis, "Ecosystem modeling and data assimilation of physical-biogeochemical processes in shelf and regional areas of the Mediterranean Sea," *Appl. Numer. Anal. Comput. Math.*, vol. 2, no. 2, pp. 262–280, 2005.

[13] F. Veersé, D. Auroux, and M. Fisher, "Limited-memory BFGS diagonal preconditioners for a data assimilation problem in meteorology," *Optim. Eng.*, vol. 1, no. 3, pp. 323–339, 2000.

[14] S. Reich, "Data assimilation: the Schrödinger perspective," *Acta Numer.*, vol. 28, pp. 635–711, 2019.

[15] M. Meldi and A. Poux, "A reduced order model based on Kalman filtering for sequential data assimilation of turbulent flows," *J. Comput. Phys.*, vol. 347, pp. 207–234, 2017.

[16] M. A. Iglesias, K. J. H. Law, and A. M. Stuart, "Evaluation of Gaussian approximations for data assimilation in reservoir models," *Comput. Geosci.*, vol. 17, no. 5, pp. 851–885, 2013.

[17] H. D. I. Abarbanel, P. J. Rozdeba, and S. Shirman, "Machine learning: deepest learning as statistical data assimilation problems," *Neural Comput.*, vol. 30, no. 8, pp. 2025–2055, 2018.

[18] G. Evensen, "The ensemble Kalman filter for combined state and parameter estimation: Monte Carlo techniques for data assimilation in large systems," *IEEE Control Syst. Mag.*, vol. 29, no. 3, pp. 83–104, 2009.

[19] J. Mandel, L. S. Bennethum, J. D. Beezley, J. L. Coen, C. C. C. Douglas, M. Kim, and A. Vodacek, "A wildland fire model with data assimilation," *Math. Comput. Simulation*, vol. 79, no. 3, pp. 584–606, 2008.

[20] A. Apte, C. K. R. T. Jones, A. M. Stuart, and J. Voss, "Data assimilation: mathematical and statistical perspectives," *Int. J. Numer. Methods Fluids*, vol. 56, no. 8, pp. 1033–1046, 2008.

[21] V. Mons, J.-C. Chassaing, T. Gomez, and P. Sagaut, "Reconstruction of unsteady viscous flows using data assimilation schemes," *J. Comput. Phys.*, vol. 316, pp. 255–280, 2016.

[22] S. W. Funke, M. Nordaas, O. Evju, M. S. Alnæ s, and K. A. Mardal, "Variational data assimilation for transient blood flow simulations: cerebral aneurysms as an illustrative example," *Int. J. Numer. Methods Biomed. Eng.*, vol. 35, no. 1, pp. e3152, 27, 2019.

[23] A. Gronskis, D. Heitz, and E. Mémin, "Inflow and initial conditions for direct numerical simulation based on adjoint data assimilation," *J. Comput. Phys.*, vol. 242, pp. 480–497, 2013.

[24] R. Ştefănescu, A. Sandu, and I. M. Navon, "POD/DEIM reduced-order strategies for efficient four dimensional variational data assimilation," *J. Comput. Phys.*, vol. 295, pp. 569–595, 2015.

[25] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk, "Data assimilation in reduced modeling," *SIAM/ASA J. Uncertain. Quantif.*, vol. 5, no. 1, pp. 1–29, 2017.

[26] P. Korn, "Data assimilation for the Navier-Stokes-$\alpha$ equations," *Phys. D*, vol. 238, no. 18, pp. 1957–1974, 2009.

[27] D. N. Daescu and I. M. Navon, "Efficiency of a POD-based reduced second-order adjoint model in 4D-var data assimilation," *Int. J. Numer. Methods Fluids*, vol. 53, no. 6, pp. 985–1004, 2007.

[28] D. Auroux and M. Nodet, "The back and forth nudging algorithm for data assimilation problems: theoretical results on transport equations," *ESAIM Control Optim. Calc. Var.*, vol. 18, no. 2, pp. 318–342, 2012.

[29] L. G. Rebholz and C. Zerfas, "Simple and efficient continuous data assimilation of evolution equations via algebraic nudging," *Numer. Methods Partial Differential Equations*, vol. 37, no. 3, pp. 2588–2612, 2021.

[30] A. Azouani, E. Olson, and E. S. Titi, "Continuous data assimilation using general interpolant observables," *J. Nonlinear Sci.*, vol. 24, no. 2, pp. 277–304, 2014.

[31] P. A. Markowich, E. S. Titi, and S. Trabelsi, "Continuous data assimilation for the three-dimensional Brinkman-Forchheimer-extended Darcy model," *Nonlinearity*, vol. 29, no. 4, pp. 1292–1328, 2016.

[32] E. Olson and E. S. Titi, "Determining modes for continuous data assimilation in 2D turbulence," vol. 113, pp. 799–840, 2003. Progress in statistical hydrodynamics (Santa Fe, NM, 2002).

[33] J. L. Lions, "Optimal control of systems governed by partial differential equations," *Springer-Verlag, New York-Berlin*, 1971.

[34] M. Yamamoto and J. Zou, "Conditional stability in reconstruction of initial temperatures," *More Progresses in Analysis*, 2009.

[35] F.-X. L. Dimet and V. P. Shutyaev, "On data assimilation for quasilinear parabolic problems," *Russian J. Numer. Anal. Math. Modelling*, vol. 16, no. 3, pp. 247–259, 2001.

[36] S. S. Pereverzyev, R. Pinnau, and N. Siedow, "Initial temperature reconstruction for nonlinear heat equation: application to a coupled radiative–conductive heat transfer problem," *Inverse Probl. Sci. Eng.*, vol. 16, no. 1, p. 55–67, 2008.

[37] C. Clason and P. Hepperger, “A forward approach to numerical data assimilation,” *SIAM J. Sci. Comput.*, vol. 31, no. 4, p. 3090–3115, 2009.

[38] I. Babuška, “The finite element method for elliptic equations with discontinuous coefficients,” *Computing (Arch. Elektron. Rechnen)*, vol. 5, no. 3, p. 207–213, 1970.

[39] Z. Chen and J. Zou, “Finite element methods and their convergence for elliptic and parabolic interface problems,” *Numer. Math.*, vol. 79, p. 175–202, Jan 1998.

[40] Y. Boubendir and S. Tlupova, “Stokes-Darcy boundary integral solutions using preconditioners,” *J. Comput. Phys.*, vol. 228, no. 23, pp. 8627–8641, 2009.

[41] M. Discacciati, E. Miglio, and A. Quarteroni, “Mathematical and numerical models for coupling surface and groundwater flows,” *Appl. Numer. Math.*, vol. 43, no. 1-2, pp. 57–74, 2002.

[42] W. J. Layton, F. Schieweck, and I. Yotov, “Coupling fluid flow with porous media flow,” *SIAM J. Numer. Anal.*, vol. 40, no. 6, pp. 2195–2218, 2002.

[43] B. Rivière and I. Yotov, “Locally conservative coupling of Stokes and Darcy flows,” *SIAM J. Numer. Anal.*, vol. 42, no. 5, pp. 1959–1977, 2005.

[44] T. Arbogast and H. L. Lehr, “Homogenization of a Darcy-Stokes system modeling vuggy porous media,” *Comput. Geosci.*, vol. 10, no. 3, pp. 291–302, 2006.

[45] R. Hoppe, P. Porta, and Y. Vassilevski, “Computational issues related to iterative coupling of subsurface and channel flows,” *Calcolo*, vol. 44, no. 1, pp. 1–20, 2007.

[46] M. Mu and J. Xu, “A two-grid method of a mixed Stokes-Darcy model for coupling fluid flow with porous media flow,” *SIAM J. Numer. Anal.*, vol. 45, no. 5, pp. 1801–1813, 2007.

[47] J. Galvis and M. Sarkis, “Balancing domain decomposition methods for mortar coupling Stokes-Darcy systems,” in *Domain decomposition methods in science and engineering XVI. Lect. Notes Comput. Sci. Eng., 55*, pp. 373–380, Springer, Berlin, 2007.

[48] X. Li, W. Gong, X. He, and C. C. Douglas, “A finite element method for data assimilation of second order parabolic interface equation,” Submitted, 2021.

[49] X. Li, W. Gong, X. He, and T. Lin, “Variational data assimilation for Stokes-Darcy equation,” To be Submitted, 2022.

[50] E. Kreyszig, *Introductory functional analysis with applications*. Wiley Classics Library, John Wiley & Sons, Inc., New York, 1989.

[51] E. Zeidler, *Applied functional analysis*, vol. 109 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1995. Main principles and their applications.

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

[53] J. C. De los Reyes, *Numerical PDE-constrained optimization*. SpringerBriefs in Optimization, Springer, Cham, 2015.

[54] G. I. Marchuk and V. P. Shutyaev, "Solvability and numerical algorithms for a class of variational data assimilation problems. A tribute to J. L. Lions," *ESAIM Control Optim. Calc. Var.*, vol. 8, pp. 873–883, 2002.

[55] E. I. Parmuzin and V. P. Shutyaev, "Numerical analysis of iterative methods for solving evolution data assimilation problems," *Russian J. Numer. Anal. Math. Modelling*, vol. 14, no. 3, pp. 275–289, 1999.

[56] H. Fareed, J. R. Singler, Y. Zhang, and J. Shen, "Incremental proper orthogonal decomposition for PDE simulation data," *Comput. Math. Appl.*, vol. 75, no. 6, pp. 1942–1960, 2018.

[57] M. Brand, "Incremental singular value decomposition of uncertain data with missing values," *Computer Vision — ECCV 2002 Lecture Notes in Computer Science*, p. 707–720, 2002.

[58] H. Fareed and J. R. Singler, "Error analysis of an incremental proper orthogonal decomposition algorithm for PDE simulation data," *J. Comput. Appl. Math.*, vol. 368, pp. 112525, 14, 2020.

[59] H. Fareed and J. R. Singler, "A note on incremental POD algorithms for continuous time data," *Appl. Numer. Math.*, vol. 144, pp. 223–233, 2019.

[60] M. J. Gander, "50 years of time parallel time integration," in *Multiple shooting and time domain decomposition methods*, vol. 9 of *Contrib. Math. Comput. Sci.*, pp. 69–113, Springer, Cham, 2015.

[61] J.-L. Lions, Y. Maday, and G. Turinici, "Résolution d'EDP par un schéma en temps "pararéel"," *C. R. Acad. Sci. Paris Sér. I Math.*, vol. 332, no. 7, pp. 661–668, 2001.

[62] Y. Maday, M.-K. Riahi, and J. Salomon, "Parareal in time intermediate targets methods for optimal control problems," *Control and Optimization with PDE Constraints*, p. 79–92, 2013.

[63] T. P. Mathew, M. Sarkis, and C. E. Schaerer, "Analysis of block parareal preconditioners for parabolic optimal control problems," *SIAM J. Sci. Comput.*, vol. 32, no. 3, pp. 1180–1200, 2010.

[64] X. Du, M. Sarkis, C. E. Schaerer, and D. B. Szyld, "Inexact and truncated parareal-in-time Krylov subspace methods for parabolic optimal control problems," *Electron. Trans. Numer. Anal.*, vol. 40, pp. 36–57, 2013.

[65] X.-M. He, T. Lin, and Y. Lin, "Immersed finite element methods for elliptic interface problems with non-homogeneous jump conditions," *Int. J. Numer. Anal. Model.*, vol. 8, no. 2, pp. 284–301, 2011.

[66] B. Vaughan, B. Smith, and D. Chopp, "A comparison of the extended finite element method with the immersed interface method for elliptic equations with discontinuous coefficients and singular sources," *Commun. Appl. Math. Comput. Sci.*, vol. 1, no. 1, p. 207–228, 2006.

[67] X.-M. He, T. Lin, Y. Lin, and X. Zhang, "Immersed finite element methods for parabolic equations with moving interface," *Numer. Methods Partial Differential Equations*, vol. 29, no. 2, pp. 619–646, 2013.

[68] G. I. Marchuk and V. P. Shutyaev, "Iterative algorithms for data assimilation problems," in *Frontiers in mathematical analysis and numerical methods*, pp. 197–206, World Sci. Publ., River Edge, NJ, 2004.

[69] A. Çeşmelioğlu and B. Rivière, "Primal discontinuous Galerkin methods for time-dependent coupled surface and subsurface flow," *J. Sci. Comput.*, vol. 40, no. 1-3, pp. 115–140, 2009.

[70] M. Discacciati, *Domain decomposition methods for the coupling of surface and groundwater flows*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, Switzerland, 2004.

[71] W. Layton, H. Tran, and C. Trenchea, "Analysis of long time stability and errors of two partitioned methods for uncoupling evolutionary groundwater-surface water flows," *SIAM J. Numer. Anal.*, vol. 51, no. 1, pp. 248–272, 2013.

[72] T. Arbogast and D. S. Brunson, "A computational method for approximating a Darcy-Stokes system governing a vuggy porous medium," *Comput. Geosci.*, vol. 11, no. 3, pp. 207–218, 2007.

[73] T. Arbogast and M. Gomez, "A discretization and multigrid solver for a Darcy-Stokes system of three dimensional vuggy porous media," *Comput. Geosci.*, vol. 13, no. 3, pp. 331–348, 2009.

[74] J. Chen, S. Sun, and X. Wang, "A numerical method for a model of two-phase flow in a coupled free flow and porous media system," *J. Comput. Phys.*, vol. 268, pp. 1–16, 2014.

[75] Y. Gao, D. Han, X.-M. He, and U. Rüde, "Unconditionally stable numerical methods for Cahn-Hilliard-Navier-Stokes-Darcy system with different densities and viscosities," *J. Comput. Phys.*, vol. 454, p. #110968, 2022.

[76] Y. Gao, X.-M. He, L. Mei, and X. Yang, "Decoupled, linear, and energy stable finite element method for the Cahn-Hilliard-Navier-Stokes-Darcy phase field model," *SIAM J. Sci. Comput.*, vol. 40, no. 1, pp. B110–B137, 2018.

[77] J. Hou, D. Hu, X.-M. He, and C. Qiu, "Modeling and a Robin-type decoupled finite element method for dual-porosity-Navier-Stokes system with application to flows around multistage fractured horizontal wellbore," *Comput. Meth. Appl. Mech. Eng.*, vol. 388, p. #114248, 2022.

[78] J. Hou, M. Qiu, X.-M. He, C. Guo, M. Wei, and B. Bai, "A dual-porosity-Stokes model and finite element method for coupling dual-porosity flow and free flow," *SIAM J. Sci. Comput.*, vol. 38, no. 5, pp. B710–B739, 2016.

[79] M. A. A. Mahbub, X.-M. He, N. J. Nasu, C. Qiu, Y. Wang, and H. Zheng, "A coupled multi-physics model and a decoupled stabilized finite element method for closed-loop geothermal system," *SIAM J. Sci. Comput.*, vol. 42, no. 4, pp. 951–982, 2020.

[80] M. A. A. Mahbub, X.-M. He, N. J. Nasu, C. Qiu, and H. Zheng, "Coupled and decoupled stabilized mixed finite element methods for non-stationary dual-porosity-Stokes fluid flow model," *Int. J. Numer. Meth. Eng.*, vol. 120, no. 6, pp. 803–833, 2019.

[81] V. J. Ervin, E. W. Jenkins, and S. Sun, "Coupled generalized nonlinear Stokes flow with flow through a porous medium," *SIAM J. Numer. Anal.*, vol. 47, no. 2, pp. 929–952, 2009.

[82] N. Hanspal, A. Waghode, V. Nassehi, and R. Wakeman, "Numerical analysis of coupled Stokes/Darcy flow in industrial filtrations," *Transp. Porous Media*, vol. 64, pp. 73–101, 2006.

[83] Y. Boubendir and S. Tlupova, "Domain decomposition methods for solving Stokes-Darcy problems with boundary integrals," *SIAM J. Sci. Comput.*, vol. 35, no. 1, pp. B82–B106, 2013.

[84] M. Cai, M. Mu, and J. Xu, "Numerical solution to a mixed Navier-Stokes/Darcy model by the two-grid approach," *SIAM J. Numer. Anal.*, vol. 47, no. 5, pp. 3325–3338, 2009.

[85] A. Çeşmelioğlu and B. Rivière, "Existence of a weak solution for the fully coupled Navier-Stokes/Darcy-transport problem," *J. Differential Equations*, vol. 252, no. 7, pp. 4138–4175, 2012.

[86] W. Chen, M. Gunzburger, D. Sun, and X. Wang, "Efficient and long-time accurate second-order methods for the Stokes-Darcy system," *SIAM J. Numer. Anal.*, vol. 51, no. 5, pp. 2563–2584, 2013.

[87] P. Chidyagwai and B. Rivière, "On the solution of the coupled Navier-Stokes and Darcy equations," *Comput. Methods Appl. Mech. Engrg.*, vol. 198, no. 47-48, pp. 3806–3820, 2009.

[88] M. Discacciati, P. Gervasio, A. Giacomini, and A. Quarteroni, “The interface control domain decomposition method for Stokes-Darcy coupling,” *SIAM J. Numer. Anal.*, vol. 54, no. 2, pp. 1039–1068, 2016.

[89] M. Discacciati, A. Quarteroni, and A. Valli, “Robin-Robin domain decomposition methods for the Stokes-Darcy coupling,” *SIAM J. Numer. Anal.*, vol. 45, no. 3, pp. 1246–1268, 2007.

[90] G. N. Gatica, S. Meddahi, and R. Oyarzúa, “A conforming mixed finite-element method for the coupling of fluid flow with porous media flow,” *IMA J. Numer. Anal.*, vol. 29, no. 1, pp. 86–108, 2009.

[91] V. Girault and B. Rivière, “DG approximation of coupled Navier-Stokes and Darcy equations by Beaver-Joseph-Saffman interface condition,” *SIAM J. Numer. Anal*, vol. 47, no. 3, pp. 2052–2089, 2009.

[92] D. Han, X.-M. He, Q. Wang, and Y. Wu, “Existence and weak-strong uniqueness of solutions to the Cahn-Hilliard-Navier-Stokes-Darcy system in superposed free flow and porous media,” *Nonlinear Anal.*, vol. 211, p. #112411, 2021.

[93] X.-M. He, N. Jiang, and C. Qiu, “An artificial compressibility ensemble algorithm for a stochastic Stokes-Darcy model with random hydraulic conductivity and interface conditions,” *Int. J. Numer. Meth. Eng.*, vol. 121, no. 4, pp. 712–739, 2020.

[94] M. Gunzburger, X.-M. He, and B. Li, “On Ritz projection and multi-step backward differentiation schemes in decoupling the Stokes-Darcy model,” *SIAM J. Numer. Anal.*, vol. 56, no. 1, pp. 397–427, 2018.

[95] G. Kanschat and B. Riviére, “A strongly conservative finite element method for the coupling of Stokes and Darcy flow,” *J. Comput. Phys.*, vol. 229, pp. 5933–5943, 2010.

[96] R. Li, J. Li, X.-M. He, and Z. Chen, “A stabilized finite volume element method for a coupled Stokes-Darcy problem,” *Appl. Numer. Math.*, vol. 133, pp. 2–24, 2018.

[97] M. Mu and X. Zhu, “Decoupled schemes for a non-stationary mixed Stokes-Darcy model,” *Math. Comp.*, vol. 79, no. 270, pp. 707–731, 2010.

[98] D. Vassilev and I. Yotov, “Coupling Stokes-Darcy flow with transport,” *SIAM J. Sci. Comput.*, vol. 31, no. 5, pp. 3661–3684, 2009.

[99] C. Guo, M. Wei, H. Chen, X.-M. He, and B. Bai, “Improved numerical simulation for shale gas reservoirs,” *OTC-24913, Offshore Technology Conference Asia, Kuala Lumpur, Malaysia, March 25-28, 2014*.

[100] Y. Zhang, C. Zhou, C. Qu, M. Wei, X.-M. He, and B. Bai, “Fabrication and verification of a glass-silicon-glass micro-nanofluidic model for investigating multi-phase flow in unconventional dual-porosity porous media,” *Lab Chip*, vol. 19, pp. 4071–4082, 2019.

[101] Y. Zhang, J. Geng, J. Liu, B. Bai, X.-M. He, M. Wei, and W. Deng, "Direct pore-level visualization and verification of in-situ oil-in-water pickering emulsification during polymeric nanogel flooding for EOR in a transparent three-dimensional micromodel," *Langmuir*, vol. 37, pp. 13353–13364, 2021.

[102] C. C. Douglas, X. Hu, B. Bai, X.-M. He, M. Wei, and J. Hou, "A data assimilation enabled model for coupling dual porosity flow with free flow," *17th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), Wuxi, China, October 19-23, 2018*, DOI: 10.1109/DCABES.2018.00085.

[103] G. Beavers and D. Joseph, "Boundary conditions at a naturally permeable wall," *J. Fluid Mech.*, vol. 30, pp. 197–207, 1967.

[104] Y. Cao, M. Gunzburger, X.-M. He, and X. Wang, "Robin-Robin domain decomposition methods for the steady Stokes-Darcy model with Beaver-Joseph interface condition," *Numer. Math.*, vol. 117, no. 4, pp. 601–629, 2011.

[105] Y. Cao, M. Gunzburger, X. Hu, F. Hua, X. Wang, and W. Zhao, "Finite element approximation for Stokes-Darcy flow with Beavers-Joseph interface conditions," *SIAM. J. Numer. Anal.*, vol. 47, no. 6, pp. 4239–4256, 2010.

[106] Y. Cao, M. Gunzburger, F. Hua, and X. Wang, "Coupled Stokes-Darcy model with Beavers-Joseph interface boundary condition," *Comm. Math. Sci.*, vol. 8, no. 1, pp. 1–25, 2010.

[107] X.-M. He, J. Li, Y. Lin, and J. Ming, "A domain decomposition method for the steady-state Navier-Stokes-Darcy model with Beavers-Joseph interface condition," *SIAM J. Sci. Comput.*, vol. 37, no. 5, pp. S264–S290, 2015.

[108] C. Qiu, X.-M. He, J. Li, and Y. Lin, "A domain decomposition method for the time-dependent Navier-Stokes-Darcy model with Beavers-Joseph interface condition and defective boundary condition," *J. Comput. Phys.*, vol. 411, p. #109400, 2020.

[109] Y. Liu, Y. He, X. Li, and X.-M. He, "A novel convergence analysis of Robin-Robin domain decomposition method for Stokes-Darcy system with Beavers-Joseph interface condition," *Appl. Math. Lett.*, vol. 119, p. #107181, 2021.

[110] C. Qiu, "Decoupling methods for the time-dependent Navier-Stokes-Darcy interface model," *Ph.D. Dissertation, Missouri University of Science and Technology*, 2019.

[111] L. Shan and H. Zheng, "Partitioned time stepping method for fully evolutionary Stokes-Darcy flow with Beavers-Joseph interface conditions," *SIAM J. Numer. Anal.*, vol. 51, no. 2, pp. 813–839, 2013.

[112] Z. Yang, J. Ming, X.-M. He, and M. Li, "A multigrid multilevel Monte Carlo method for Stokes-Darcy model with random hydraulic conductivity and Beavers-Joseph condition," *J. Sci. Comput.*, vol. 90, p. #68, 2022.

[113] Y. Cao, M. Gunzburger, X.-M. He, and X. Wang, “Parallel, non-iterative, multi-physics domain decomposition methods for time-dependent Stokes-Darcy systems,” *Math. Comp.*, vol. 83, no. 288, pp. 1617–1644, 2014.

[114] M. Yamamoto and J. Zou, “Simultaneous reconstruction of the initial temperature and heat radiative coefficient,” vol. 17, pp. 1181–1202, 2001. Special issue to celebrate Pierre Sabatier’s 65th birthday (Montpellier, 2000).

[115] W. Feng, X.-M. He, Z. Wang, and X. Zhang, “Non-iterative domain decomposition methods for a non-stationary Stokes-Darcy model with Beavers-Joseph interface condition,” *Appl. Math. Comput.*, vol. 219, no. 2, pp. 453–463, 2012.

# VITA

Xuejian Li was born in Jingyuan, a small wild and historical town in northwest China. He grew up and studied there until moving to Chengdu for his college education. He received his bachelor degree in Statistics at Sichuan University. After that, he worked in financial industry as a stocks data analyst and trader before 2018. From Fall 2018, he started his Ph.D studies in Mathematics at Missouri University of Science and Technology under the supervision of Dr. Xiaoming He. His research was focused on numerical methods, optimal control, and data assimilation. In May 2022, he received his PhD in Mathematics with Applied and Computational Mathematics emphasis from Missouri S&T.

Xuejian Li liked hiking and playing basketball.