Networks and
Heterogeneous Media

https://www.aimspress.com/journal/nhm

NHM, 20(3): 732–758.
DOI: 10.3934/nhm.2025031
Received: 08 March 2025
Revised: 21 May 2025
Accepted: 30 May 2025
Published: 30 June 2025

Research article

A new ensemble Monte Carlo method for a parabolic optimal control problem with random coefficient

Yan Guo [1], **Xianbing Luo**[1,*] **and Changlun Ye**[2]

[1] School of Mathematics and Statistics, Guizhou University, Guiyang 550025, China
[2] School of Mathematical Sciences, Guizhou Normal University, Guiyang 550025, China

* **Correspondence:** Email: xbluo1@gzu.edu.cn.

Abstract: A new ensemble Monte Carlo (EMC) method is proposed and applied to numerically simulate a parabolic optimal control problem with random coefficients. The state equation is discretized by the EMC method, which shares a common coefficient matrix with multiple right-hand vectors. It saves the computational cost compared with the Monte Carlo (MC) method. For this new EMC method, it is unconditionally stable and does not need to subgroup the samples in the simulation. Under natural regularity condition, some error estimates are obtained for the EMC approximation of the optimal control problem. Two numerical examples are presented to test the theoretical results.

Keywords: parabolic equation; optimal control problem; ensemble Monte Carlo; random coefficients

1. Introduction

In this paper, we consider the optimal control problem parameterized by $\omega \in \Omega$

$$\min_{\bar{q}\in Q_{ad}} J(\bar{u},\bar{q}) = \frac{1}{2} \parallel \bar{u}(\omega,t,x) - u_d(t,x) \parallel^2_{L^2(0,T,L^2(D))} + \frac{\lambda}{2} \parallel \bar{q}(t,x) \parallel^2_{L^2(0,T,L^2(D))} \tag{1.1}$$

subject to

$$\begin{cases} \partial_t \bar{u}(\omega,t,x) - \nabla \cdot (a(\omega,t,x)\nabla \bar{u}(\omega,t,x)) = \bar{q}(t,x) + \bar{f}(\omega,t,x), \text{in } (0,T]\times D, \\ \qquad \bar{u}(\omega,t,x) = 0, \quad \text{on } (0,T]\times \partial D, \\ \qquad \bar{u}(\omega,0,x) = g(x), \quad \text{in } D, \end{cases} \tag{1.2}$$

where Ω is a sample space, $D \subset \mathbb{R}^2$ is a bounded convex polygonal domain, and ∂D is the boundary of D. The differential operators $\nabla\cdot$ and ∇ are with respect to $x \in D$, and $u_d \in L^2(0,T;L^2(D))$ is given as the desired state and $\lambda > 0$ is a given regularization parameter. The admissible control set $Q_{ad} \subset L^2(0, T, L^2(D))$ is defined by

$$Q_{ad} = \{\bar{q} \in L^2(0, T, L^2(D)) : q_a \le \bar{q}(t, x) \le q_b \quad \text{a.e. in } (0, T] \times D\}. \tag{1.3}$$

For simplicity, we only consider q_a, q_b to be two constants with $q_a < q_b$.

Optimal control problems widely arise in engineering and applied sciences. For deterministic optimal control problems, there are large mathematical theories and computational approaches (see, e.g., [1–4]). However, the boundary conditions, material parameters, and external loading are often not precisely measured. Therefore stochastic models are more realistic. Just like the deterministic optimal control problems, it is very difficult to obtain the analytical solution for stochastic optimal control problems (SOCPs). Numerical solutions are a good choice for the application of these problems. The study of numerical approximation for SOCPs has attracted much interest in the last few decades (see, e.g., [5–9]).

The numerical methods of stochastic differential equation play an important role in the numerical approximation of SOCPs. Some discretization schemes, such as generalized polynomial chaos (gPC) expansions [10], sparse-grid stochastic collocation, the multi-grid method [11], the stochastic Galerkin method [12, 13], and a hybrid model reduction method [14], have been studied. For stochastic problems,the MC method is another important method. the multilevel MC finite element methods in [5] and the quasi-MC method in [7] are used to solve an elliptic optimal control problem with random coefficients. The problems (1.1) and (1.2) is similar to the problem considered in [5] or [11]. It is a parameterized optimal control problem. As stated in reference [5], the statistical properties of control q is the initial guess for the corresponding robust control.

For the stochastic evolution equations, when we use the MC method, an ensemble approach is widely used method (e.g., [15–19]). The ensemble method shares a common matrix with multiple right-hand sides (RHSs) by introducing an ensemble average of random coefficients. Thus, it greatly reduces the storage requirements and computational cost. This ensemble method sometimes must subgroup the samples to meet the stability condition.

In this paper, we improve the ensemble Monte Carlo (EMC) method in [20] to make it even more efficient by choosing $\bar{a} > a_{max}/2$ (the definitions of $\bar{a}, a_{max}$ can been seen in expressions (2.2) and (3.5)), but not the ensemble mean $\bar{a} = \frac{1}{N}\sum_i^N a(\omega_i)$. For the SOCP (1.1) and (1.2), we use the discretize-then-optimize approach. The state Eq (1.2) is first discretized by the improved EMC linear conforming finite element method. Then the existence and uniqueness of the solution of the discretized optimal control problem are analyzed. After this, we get the error estimate for the EMC approximation as follows:

$$\begin{aligned}\lambda \parallel q - q_{\tau h} \parallel_{L^2(\Omega;L^2(0,T;L^2(D)))} + \parallel u - u_{\tau h} \parallel_{L^2(\Omega;L^2(0,T;L^2(D)))} \\ + \parallel z - z_{\tau h} \parallel_{L^2(\Omega;L^2(0,T;L^2(D)))} \le C(\tau + h^2),\end{aligned}$$

where (q, u, z) and $(q_{\tau h}, u_{\tau h}, z_{\tau h})$ are the continuous and discrete optimal controls triples, respectively.

The structure of the rest of the paper is as follows. In Section 2, we present some notation, the first-order optimality conditions, and the regularities of the state and the control. In Section 3, we focus on the discretization of the SOCP (1.1) and (1.2). We discuss the ensemble scheme for the state equation and adjoint equation, and variational discretization of the control variable. In Section 4, we prove an a priori estimate for the $L^2(0, T, L^2(D))$ error of the EMC approximation for the optimal control problem. Some numerical experiments are carried out to verify the validity of this algorithm in Section 5. In Section 6, some summaries are given. The proofs of some theorems are in the Appendix.

2. The problem setting

2.1. Notation

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a complete probability space, where $\mathcal{A} \subset 2^{\Omega}$ is a σ-algebra, $\mathbb{P} : \mathcal{A} \to [0, 1]$ is a probability measure, and $X = \{\upsilon : \upsilon \in L^2(0, T; H_0^1(D)), \partial_t \upsilon \in L^2(0, T; H^{-1}(D))\}$. Given a Banach space $Y = L^2(0, T; L^2(D))$, the space $L^p(\Omega, Y)$ is the set of strongly measurable functions $(\upsilon : \Omega \to Y)$ such that $\| \upsilon \|_{L^p(\Omega,Y)} \leq \infty$, where

$$\| \upsilon \|_{L^p(\Omega,Y)} = \begin{cases} \left(\left(\int_{\Omega} \| \upsilon \|_Y^p \, d\mathbb{P}(\omega) \right) \right)^{1/p}, & \text{for } 1 \leq p < \infty, \\ \operatorname{essup}_{\omega \in \Omega} \| \upsilon(\omega) \|_Y, & \text{for } p = \infty. \end{cases}$$

We use the following notation for the different inner products and the corresponding norms:

$$(v, u) = (v, u)_{L^2(D)}, \qquad \| u \| = \| u \|_{L^2(D)};$$

$$(v, u)_I = (v, u)_{L^2(0,T;L^2(D)))}, \qquad \| v \|_I = \| v \|_{L^2(0,T;L^2(D))};$$

$$(v, u)_{\Omega} = (v, u)_{L^2(\Omega;L^2(0,T;L^2(D)))}, \qquad \| v \|_{\Omega} = \| v \|_{L^2(\Omega;L^2(0,T;L^2(D)))};$$

$$\| v \|_r = \| v \|_{H^r(D)} = \left(\sum_{|\alpha| \leq r} \| D^{\alpha} v \|^2 \right)^{\frac{1}{2}}.$$

T_h is a quasi-uniform triangulation of the domain D, where $\bar{D} = \cup_{K \in T_h} \bar{K}$. The mesh size $h := \max_{K \in T_h} h_K$, where h_K is the diameter of K. V_h denotes the finite element space.

$$V_h := \{\upsilon_h \in H_0^1(D); \upsilon_h|_K \text{ is a linear polynomials}, \forall K \in T_h\}.$$

For the time interval $(0, T]$, we use a uniform partition with the step size τ. $(0, T] = I_1 \cup I_2 \cup \cdots \cup I_N$ with the subintervals $I_n = (t_{n-1}, t_n]$ of size τ and the time points $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T$.

For convenience, here, we use the following notation:

$$(v, u)_{I_n} = (v, u)_{L^2(t_{n-1},t_n;L^2(D))}, \quad \| v \|_{I_n} = \| v \|_{L^2(t_{n-1},t_n;L^2(D))}.$$

Define $\widehat{\varphi} = \frac{1}{T} \int_0^T \varphi dt$. According to the definition, for any constant C, we get

$$(C, \varphi - \widehat{\varphi})_{L^2(0,T)} = 0, \tag{2.1}$$

and $\| \widehat{\varphi} \|_{L^2(0,T)} \leq \| \varphi \|_{L^2(0,T)}$.

In the following, C is a positive constant and has different values in different places, which is independent of h, τ, and ω. C_{ω} is also a positive constant and has different values at different locations, which is dependent on ω and independent of h, τ.

2.2. The state equation

We first make the following assumptions on the input data.

Assumption (1): There are two positive constants a_{min}, a_{max} such that, for any $x \in \bar{D}, t \in [0, T]$,

$$\mathbb{P}(a_{min} \le a(\omega, t, x) \le a_{max}) = 1. \tag{2.2}$$

Assumption (2): $\bar{q}(x) \in L^2(0, T; L^2(D))$, $g(x) \in H_0^1(D)$, $\bar{f} \in L^2(\Omega; L^2(0, T; L^2(D)))$.

Assumption (3): Almost surely (a.s.), $\omega \in \Omega$, $a(\omega, t, x) \in C([0, T], C^1(\bar{D}))$. For all $(t, x) \in [0, T] \times \bar{D}$, there is a constant θ_1 such that $\mathbb{P}(|\nabla a(\omega, t, x)| \le \theta_1) = 1$, where θ_1 is independent of ω.

Assumption (4): The coefficient $a(\omega, t, x)$ has Lipschitz continuity with respect to time t, i.e.,

$$\| a(\omega, b, x) - a(\omega, c, x) \|_{C(\bar{D})} \le L \mid b - c \mid, \;\; a.s.\ \omega \in \Omega, \tag{2.3}$$

where b, $c \in [0, T]$, and L is independent of ω.

From now on, the subscript ω denotes the dependence of ω. The variational form of (1.2) is to find $\bar{u}_\omega \in X$ such that

$$\begin{cases} (\partial_t \bar{u}_\omega, v)_I + (a_\omega \nabla \bar{u}_\omega, \nabla v)_I = (\bar{f}_\omega, v)_I + (\bar{q}, v)_I, \quad \forall v \in X, \\ \qquad\qquad\qquad\qquad \bar{u}_\omega(0) = g(x). \end{cases} \tag{2.4}$$

For any $\omega \in \Omega$, $\bar{u}_\omega$ is a weak solution of problem (1.2) if and only if $\bar{u}_\omega \in L^2(0, T; H^2(D)) \cap H^1(0, T; L^2(D)) \hookrightarrow C([0, T], H_0^1(D))$ and satisfies problem (2.4) (see Chapter 7 of [21]).

For the weak solution $\bar{u}_\omega$ of problem (1.2), we have the following result.

Theorem 2.1. *Let Assumptions (1)–(3) hold. Then, for $\omega \in \Omega$ a.s., there is a unique weak solution $\bar{u}_\omega \in L^2(0, T; H^2(D)) \cap H^1(0, T; L^2(D))$ for problem* (1.2) *and it holds that*

$$\| \nabla \bar{u}_\omega \|_I^2 \le C(\| \bar{q} \|_I^2 + \| \bar{f}_\omega \|_I^2 + \| g \|^2). \tag{2.5}$$

Proof. For a fixed $\omega \in \Omega$, Eq (1.2) is a deterministic parabolic partial differential equation (PDE). We can obtain the unique solution $\bar{u}_\omega \in L^2(0, T; H^2(D)) \cap H^1(0, T; L^2(D))$ by referring to Theorem 5 on page 384 in reference [21].

In problem (2.4), we choose $v = \bar{u}_\omega$ and get

$$(\partial_t \bar{u}_\omega, \bar{u}_\omega)_I + (a_\omega \nabla \bar{u}_\omega, \nabla \bar{u}_\omega)_I = (\bar{f}_\omega + \bar{q}, \bar{u}_\omega)_I. \tag{2.6}$$

Using Poincare's inequality, we have

$$\frac{1}{2} \| \bar{u}_\omega(T) \|^2 - \frac{1}{2} \| g(x) \|^2 + a_{\min} \| \nabla \bar{u}_\omega \|_I^2 \le C \| \bar{q} + \bar{f}_\omega \|_I \| \nabla \bar{u}_\omega \|_I, \tag{2.7}$$

where C is independent of ω. Furthermore, by Young's inequality, we can obtain the desired result of inequality (2.5).

2.3. The optimal control problem

For $\omega \in \Omega$ a.s., we consider the following optimal control problem parameterized by $\omega \in \Omega$:

$$\min_{\bar{q}\in Q_{ad}} J_\omega(\bar{q}) = \frac{1}{2} \parallel \bar{u}_\omega(\bar{q}) - u_d \parallel_I^2 + \frac{\lambda}{2} \parallel \bar{q} \parallel_I^2, \tag{2.8}$$

where $\bar{u}_\omega(\bar{q})$ is the weak solution of the parabolic problem (1.2). To begin with, we consider the existence and uniqueness of the solution to problem (2.8).

Theorem 2.2. *Suppose that Q_{ad} is nonempty. Then, for $\omega \in \Omega$ a.s., a unique global solution $q_\omega \in Q_{ad}$ exists for the problem* (2.8).

Proof. For a fixed $\omega \in \Omega$, the problem (2.8) is a deterministic infinite dimensional optimization problem. Q_{ad} is closed and convex. The cost functional J_ω is strictly convex. Due to $\lambda > 0$, it is enough to consider the minimizing sequence and argue it in a classical way to verify the existence of the global solution for problem (2.8). The uniqueness follows from the strict convexity of J_ω. For the details one can refer to [1] or [4].

The subscript ω of q_ω indicates that q_ω is the solution of problem (2.8) for a given $\omega \in \Omega$. The next result gives the first-order optimality condition of problem (2.8). For the proof, one can refer to [1].

Theorem 2.3. *A feasible control $q_\omega \in Q_{ad}$ is a solution of problem* (2.8) *if and only if a state $u_\omega \in L^2(0,T;H^2(D)) \cap H^1(0,T;L^2(D))$ and an adjoint state $z_\omega \in L^2(0,T;H^2(D)) \cap H^1(0,T;L^2(D))$ exist such that the following hold:*

$$\begin{aligned} (\partial_t u_\omega, v)_I + (a_\omega \nabla u_\omega, \nabla v)_I &= (\bar{f}_\omega + q_\omega, v)_I, \forall v \in X, \quad u_\omega(0) = g(x); && (2.9)\\ -(\partial_t z_\omega, v)_I + (a_\omega \nabla z_\omega, \nabla v)_I &= (u_\omega - u_d, v)_I, \forall v \in X, \quad z_\omega(T) = 0; && (2.10)\\ (z_\omega + \lambda q_\omega, \bar{q} - q_\omega)_I &\geq 0, \qquad \forall \bar{q} \in Q_{ad}. && (2.11)\end{aligned}$$

Now we define a map $q : \Omega \times (0,T] \times D \to R$ such that

$$q(\omega,\cdot,\cdot) := \arg\min_{\bar{q}\in Q_{ad}} J_\omega(\bar{q}), \tag{2.12}$$

where $q(\omega,\cdot,\cdot)$ is the solution of problem (2.8) for a given $\omega \in \Omega$. The next theorem explains that q is measurable.

Theorem 2.4. *For each $\omega \in \Omega$, the map $\omega \to q(\omega,\cdot,\cdot)$ is measurable.*

Proof. This can be proved analogously to the proof of Theorem 3.4 in [5].

3. The ensemble scheme approximation of optimal control problem

In this section, we first introduce a projection $\mathcal{R}_{\omega,h}$. Second, we present the discretization of the state equation and the corresponding results. Finally, we discretize the optimal control problem.

Let $\mathcal{R}_{\omega,h} : H_0^1(D) \to V_h$ be the Ritz projection operator given by

$$(a_\omega \nabla u, \nabla v_h) = (a_\omega \nabla \mathcal{R}_{\omega,h} u, \nabla v_h), \quad \forall v_h \in V_h. \tag{3.1}$$

We then have

$$\| \nabla \mathcal{R}_{\omega,h} u \| \leq \frac{a_{\max}}{a_{\min}} \| \nabla u \| . \tag{3.2}$$

For the projection operator $\mathcal{R}_{\omega,h}$, the following error estimates also hold (see, e.g., [22]).

Lemma 3.1. *Suppose that $u \in H^2(D) \cap H_0^1(D)$ for the $\mathcal{R}_{\omega,h}$ defined by Eq* (3.1), *the following estimates hold:*

$$\begin{aligned} &\| \mathcal{R}_{\omega,h} u - u \| + h \| \mathcal{R}_{\omega,h} u - u \|_1 \leq Ch^2 \| u \|_2, \\ &\| \mathcal{R}_{\omega,h} u - u \|_{H^{-1}(D)} \leq Ch^2 \| u \|_1, \end{aligned} \tag{3.3}$$

where C is a positive constant independent of h, τ, and ω.

3.1. The discretization of the state equation

Choose a positive constant $\bar{a}$ which satisfies $\bar{a} > \frac{a_{max}}{2}$. For $\omega \in \Omega$ a.s., the **new EMC** approximation for problem (2.4) is defined as follows: Find $\bar{u}^n_{\omega,h} \in V_h$ such that for $n = 1, \cdots, N$

$$\begin{aligned} \left(\frac{\bar{u}^n_{\omega,h} - \bar{u}^{n-1}_{\omega,h}}{\tau}, v_h \right) + (\bar{a} \nabla \bar{u}^n_{\omega,h}, \nabla v_h) + ((a^n_\omega - \bar{a}) \nabla \bar{u}^{n-1}_{\omega,h}, \nabla v_h) &= \left(\widehat{\bar{q}}^n + \widehat{\bar{f}_\omega}^n, v_h \right), \forall v_h \in V_h, \\ \bar{u}^0_{\omega,h} &= g_h(x), \end{aligned} \tag{3.4}$$

where $\widehat{\bar{q}}^n = \frac{1}{\tau} \int_{I_n} \bar{q} dt$, $g_h(x) = \mathcal{R}_{\omega,h} g$.

Remark 3.1. *Under Assumption (1), for the constant $\bar{a}$, it is clear that*

$$\theta = \sup_{(\omega,t,x) \in \Omega \times [0,T] \times \bar{D}} |\bar{a} - a(\omega, t, x)|, \quad \theta < \bar{a}, \tag{3.5}$$

where θ is independent of ω. For the choice of $\bar{a}$, the stability condition (Theorem 1 of [20]) holds naturally.

The resulting coefficient matrix of the problem (3.4) is only dependent on the constant $\bar{a}$. This is the key feature of the ensemble method. That is, the discrete systems share a unique coefficient matrix and multiple right-hand-side vectors. These systems can be efficiently computed by many existing algorithms, such as *LU* factorization method, GMRES (Generalized minimum residual) method, etc.

Next, we present some results about the fully discrete scheme in problem (3.4).

Theorem 3.1. *Let Assumptions (1) and (2) hold. Then, for $\omega \in \Omega$ a.s., a unique solution $\{\bar{u}^n_{\omega,h}\}_{n=1}^N$ exists for the fully discrete scheme* (3.4). *Moreover*

$$\| \bar{u}^N_{\omega,h} \|^2 + \theta k \| \nabla \bar{u}^N_{\omega,h} \|^2 + (\bar{a} - \theta) k \sum_{n=1}^N \| \nabla \bar{u}^n_{\omega,h} \|^2 \leq C \left(\| \bar{q} \|_I^2 + \| \bar{f}_\omega \|_I^2 + (\theta k + 1) \| \nabla g \|^2 \right). \tag{3.6}$$

Theorem 3.2. *Let Assumptions (1)–(4) hold. Then the numerical solution of problem* (3.4) *satisfies*

$$\begin{aligned} &\frac{1}{\tau} \sum_{n=1}^N \| \bar{u}^n_{\omega,h} - \bar{u}^{n-1}_{\omega,h} \|^2 + (2\bar{a} - a_{\max}) \sum_{n=1}^N \| \nabla (\bar{u}^n_{\omega,h} - \bar{u}^{n-1}_{\omega,h}) \|^2 \\ &+ a_{\min} \| \nabla \bar{u}^N_{\omega,h} \|^2 \leq C \left(\| \nabla g \|^2 + \| \bar{q} \|_I^2 + \| \bar{f}_\omega \|_I^2 \right). \end{aligned} \tag{3.7}$$

3.2. The discretization of the optimal control problem

Here, we discretize the problem (2.8) via the variational discretization approach developed in [23]. For $\omega \in \Omega$ a.s., the variational discretization of problem (2.8) reads as follows:

$$\min_{\bar{q}\in Q_{ad}} J_{\omega,\tau h}(\bar{q}) = \frac{1}{2}\tau \sum_{n=1}^{N} \| \bar{u}^n_{\omega,h}(\bar{q}) - \widehat{u_{\omega,d}}^{\,n} \|^2 + \frac{\lambda}{2} \| \bar{q} \|_I^2, \tag{3.8}$$

where $\left\{\bar{u}^n_{\omega,h}(\bar{q})\right\}_{n=1}^{N}$ is the solution of problem (3.4). The key idea of the variational discretiation is only to discretize the state and adjoint state space, not to discretize the control space Q_{ad}. The control is obtained from a projection of the adjoint state. Hence problem (3.8) is again an optimization problem in infinite dimensions. Thus, all techniques we used previously to study the problem (2.8) can also be used for problem (3.8). A detailed study of the variational discretization, together with its numerical implementation and comparisons to classical discretizattions, can be found in [23] or Chapter 3 of [1].

Theorem 3.3. *Suppose that Q_{ad} is nonempty. Then, for $\omega \in \Omega$ a.s., a unique global solution $q_{\omega,\tau h} \in Q_{ad}$ exists for the problem* (3.8).

Proof. Analogously to Theorem 2.2, one can show that for $\omega \in \Omega$ a.s., the problem (3.8) admits a unique global solution, which we denote it as $q_{\omega,\tau h}$.

The next theorem gives the first-order optimality condition of problem (3.8) with an ensemble scheme.

Theorem 3.4. *A feasible control $q_{\omega,\tau h} \in Q_{ad}$ is a solution of problem* (3.8) *if and only if a state $u^n_{\omega,h} \in V_h$ and an adjoint state $z^n_{\omega,h} \in V_h$ exist such that*

$$\begin{aligned} \left(\frac{u^n_{\omega,h} - u^{n-1}_{\omega,h}}{\tau}, v_h\right) + (\bar{a}\nabla u^n_{\omega,h}, \nabla v_h) + ((a^n_\omega - \bar{a})\nabla u^{n-1}_{\omega,h}, \nabla v_h) &= (\widehat{q_{\omega,\tau h}}^{\,n} + \widehat{\bar{f}_\omega}^{\,n}, v_h), \forall v_h \in V_h, \\ u^0_{\omega,h} &= g_h(x), \quad n = 1, \cdots, N, \end{aligned} \tag{3.9}$$

$$\begin{aligned} -\left(\frac{z^n_{\omega,h} - z^{n-1}_{\omega,h}}{\tau}, v_h\right) + (\bar{a}\nabla z^{n-1}_{\omega,h}, \nabla v_h) + ((a^{n+1}_\omega - \bar{a})\nabla z^n_{\omega,h}, \nabla v_h) &= (u^n_{\omega,h} - \widehat{u_d}^{\,n}, v_h), \forall v_h \in V_h, \\ z^N_{\omega,h} = 0, a^{N+1}_\omega &= 0, \quad n = 1, \cdots, N, \end{aligned} \tag{3.10}$$

$$(\bar{q} - q_{\omega,\tau h}, z_{\omega,\tau h} + \lambda q_{\omega,\tau h})_I \geq 0, \quad \forall \bar{q} \in Q_{ad}, \tag{3.11}$$

where $z_{\omega,\tau h}|_{[t_{n-1},t_n)} = z^{n-1}_{\omega,h}$.

Remark 3.2. *Given Theorem 3.4, we know that the adjoint equation is also an ensemble scheme. Therefore, for all ω, we solve the optimal control problem* (2.8) *using an optimization algorithm, such as a projected gradient algorithm, a primal-dual active set strategy, or a conjugate gradient method, etc., via a unique coefficient matrix, which greatly reduces the computational cost.*

We introduce the discrete control $q_{\tau h} : \Omega\times[0,T]\times D \to R$ whose realization $q_{\tau h}(\omega,\cdot)$ is the solution of problem (3.8) for the given $\omega \in \Omega$ and a mesh size h,τ. Precisely, let $q_{\tau h} : \Omega \times [0,T] \times D \to R$ such that

$$q_{\tau h}(\omega,\cdot) = \arg\min_{\bar{q}\in Q_{ad}} J_{\omega,\tau h}(\bar{q}), \qquad \text{a.s. } \omega \in \Omega. \tag{3.12}$$

Thus $q_{\tau h}$ is indeed measurable for $\omega \to q_{\tau h}(\omega,\cdot,\cdot)$, which can be proved analogously to the proof of Theorem 3.4 of [5].

Theorem 3.5. *Let Assumptions (1)–(3) hold and $q_{\tau h}$ be the solution of problem* (3.12). *Then for any $\bar{q} \in Q_{ad}$, the following holds:*

$$\| q_{\omega,\tau h} \|_I^2 \le C\left(\| \bar{q} \|_I^2 + (\theta\tau + 1) \| \nabla g \|^2 + \| \bar{f}_\omega \|_I^2 + \| \bar{q} \|_I^2 + \| u_d \|_I^2\right). \tag{3.13}$$

Proof. We begin by establishing the bound in inequality (3.13). From the optimality of $q_{\omega,\tau h}$ together with the estimate (3.6), it follows that for any $\bar{q} \in Q_{ad}$, the following holds:

$$\begin{aligned}\| q_{\omega,\tau h} \|_I^2 &\le J_{\omega,\tau h}(\bar{q})\\ &= \frac{\tau}{2}\sum_{n=1}^{N} \| \bar{u}_{\omega,h}^n - \widehat{u_d}^n \|^2 + \frac{\lambda}{2} \| \bar{q} \|_I^2\\ &\le \sum_{n=1}^{N} \tau \| \bar{u}_{\omega,h}^n \|^2 + \sum_{n=1}^{N} \tau \| \widehat{u_d}^n \|^2 + \frac{\lambda}{2} \| \bar{q} \|_I^2\\ &\le C\left(\| \bar{q} \|_I^2 + (\theta\tau + 1) \| \nabla g \|^2 + \| \bar{f}_\omega \|_I^2 + \| u_d \|_I^2\right),\end{aligned}$$

from which we obtain the desired result.

If Q_{ad} is bounded, or q_a, q_b is bounded, then $q_{\omega,\tau h} \in L^2(\Omega; L^2(0,T; L^2(D))$.

Corollary 3.1. *Suppose that $u_{\omega,\tau h} \in L^2(0,T; L^2(D))$, $u_d \in L^2(0,T; L^2(D))$, and Assumptions (1)–(3) are satisfied. Then the numerical solution of problem* (3.10) *satisfies*

$$\| z_{\omega,h}^0 \|^2 + a_{min} \sum_{n=1}^{N} \tau \| \nabla z_{\omega,h}^{n-1} \|^2 + (\bar{a} - \theta)\tau \| \nabla z_{\omega,h}^0 \|^2 \le C\frac{\tau}{\bar{a} - \theta} \sum_{n=1}^{N} \| u_{\omega,h}^n - \widehat{u_d}^n \|^2 .$$

Proof. According to problem (3.10), we choose $v_h = z_{\omega,h}^{n-1}$ and get

$$-\left(\frac{z_{\omega,h}^n - z_{\omega,h}^{n-1}}{\tau}, z_{\omega,h}^{n-1}\right) + (\bar{a}\nabla z_{\omega,h}^{n-1}, \nabla z_{\omega,h}^{n-1}) + ((a_\omega^n - \bar{a})\nabla z_{\omega,h}^n, \nabla z_{\omega,h}^{n-1}) = (u_{\omega,h}^n - \widehat{u_d}^n, z_{\omega,h}^{n-1}). \tag{3.14}$$

On the basis of Eq (3.14), all steps of the proof of Theorem 3.1 can be repeated similarly to obtain the desired result.

Corollary 3.2. *Suppose that $u_{\omega,\tau h} \in L^2(0,T; L^2(D))$, $u_d \in L^2(0,T; L^2(D))$, and Assumptions (1)–(4) are satisfied. Then the numerical solution of problem* (3.10) *satisfies*

$$\begin{aligned}&\frac{1}{\tau}\sum_{n=1}^{N} \| z_{\omega,h}^n - z_{\omega,h}^{n-1} \|^2 + \quad (2\bar{a} - a_{\max})\sum_{n=1}^{N} \| \nabla(z_{\omega,h}^n - z_{\omega,h}^{n-1}) \|^2 + a_{\min} \| \nabla z_{\omega,h}^0 \|^2\\ &\le C\tau\sum_{n=1}^{N} \| u_{\omega,h}^n - \widehat{u_d}^n \|^2 .\end{aligned}$$

Proof. According to problem (3.10), we choose $v_h = z^{n-1}_{\omega,h} - z^n_{\omega,h}$ and have

$$\begin{aligned} &-\left(\frac{z^n_{\omega,h} - z^{n-1}_{\omega,h}}{\tau}, z^{n-1}_{\omega,h} - z^n_{\omega,h}\right) + (\bar{a}\nabla z^{n-1}_{\omega,h}, \nabla(z^{n-1}_{\omega,h} - z^n_{\omega,h})) \\ &+((a^{n+1}_{\omega} - \bar{a})\nabla z^n_{\omega,h}, \nabla(z^{n-1}_{\omega,h} - z^n_{\omega,h})) = (u^n_{\omega,h} - \widehat{u_d}^n, z^{n-1}_{\omega,h} - z^n_{\omega,h}). \end{aligned} \tag{3.15}$$

On the basis of this representation, all steps of the proof of Theorem 3.2 can be repeated similarly to obtain the desired result.

4. Error analysis of the ensemble method for the SOCP

In the subsequent analysis, we first introduce the following three auxiliary problems. $P_0((0,T], V_h)$ denotes the space of the piecewise constant defined on $\bigcup_{i=1}^N I_i = (0,T]$ with values in V_h.

For $q_{\omega,\tau h} \in Q_{ad}$, find $u_\omega(q_{\omega,\tau h}) \in L^2(0,T;L^2(D))$ such that

$$\begin{cases} \partial_t u_\omega(q_{\omega,\tau h}) - \nabla\cdot[a_\omega \nabla u_\omega(q_{\omega,\tau h})] = q_{\omega,\tau h} + \bar{f}_\omega, & t, x \in (0,T]\times D, \\ u_\omega(q_{\omega,\tau h}) = 0, & x\in\partial D,\ t\in(0,T], \\ u_\omega(q_{\omega,\tau h})(x,0) = g(x), & x \in D. \end{cases} \tag{4.1}$$

For $u_{\omega,\tau h} \in P_0((0,T], V_h)$, find $z_\omega(u_{\omega,\tau h}) \in L^2(0,T;L^2(D))$ such that

$$\begin{cases} -\partial_t z_\omega(u_{\omega,\tau h}) - \nabla\cdot[a_\omega \nabla z_\omega(u_{\omega,\tau h})] = u_{\omega,\tau h} - u_d, & t, x \in (0,T]\times D, \\ z_\omega(u_{\omega,\tau h}) = 0, & x\in\partial D,\ t\in[0,T], \\ z_\omega(u_{\omega,\tau h})(x,T) = 0, & x \in D. \end{cases} \tag{4.2}$$

And for $u_{\omega,\tau h} \in L^2(0,T;L^2(D))$, find $z_\omega(u_\omega(q_{\omega,\tau h})) \in L^2(0,T;L^2(D))$ such that

$$\begin{cases} -\partial_t z_\omega(u_\omega(q_{\omega,\tau h})) - \nabla\cdot[a_\omega \nabla z_\omega(u_\omega(q_{\omega,\tau h}))] = u_\omega(q_{\omega,\tau h}) - u_d, & t, x \in (0,T]\times D, \\ z_\omega(u_\omega(q_{\omega,\tau h})) = 0, & x\in\partial D, t\in[0,T], \\ z_\omega(u_\omega(q_{\omega,\tau h}))(x,T) = 0, & x \in D. \end{cases} \tag{4.3}$$

Since $q_{\omega,\tau h} \in Q_{ad} \subset L^2(0,T;L^2(D))$, one concludes from Theorem 2.1 that the problem (4.1) admits a unique solution $u_\omega(q_{\omega,\tau h}) \in L^2(0,T;L^2(D))$. Similarly, the problem (4.2) admits a unique solution $z_\omega(u_{\omega,\tau h}) \in L^2(0,T;L^2(D))$ and problem (4.3) admits a unique solution $z_\omega(u_\omega(q_{\omega,\tau h})) \in L^2(0,T;L^2(D))$. Here, $u_{\omega,\tau h} \in P_0((0,T],V_h)$ and $z_{\omega,\tau h} \in P_0((0,T],V_h)$ are the fully discrete finite element approximations of $u_\omega(q_{\omega,\tau h})$ and $z_\omega(u_{\omega,\tau h})$, respectively.

Theorem 4.1. *Let $u_\omega(q_{\omega,\tau h}) \in L^2(0,T;L^2(D))$ and $u_{\omega,\tau h} \in P_0((0,T],V_h)$ be the solutions of problems* (4.1) *and* (3.9)*, respectively. Assume that Assumptions (1)–(4) hold. Then we have the following error estimate:*

$$\| u_{\omega,\tau h} - u_\omega(q_{\omega,\tau h}) \|_I \leq C_\omega(\tau + h^2). \tag{4.4}$$

Theorem 4.2. *Let $z_\omega(u_{\omega,\tau h}) \in L^2(0,T;L^2(D))$ and $z_{\omega,\tau h} \in P_0((0,T],V_h)$ be the solutions of problems* (4.2) *and* (3.10)*, respectively. Assume that Assumption (1)–(4) hold. Then we have the following error estimate*

$$\| z_{\omega,\tau h} - z_\omega(u_{\omega,\tau h}) \|_I \leq C_\omega(\tau + h^2).$$

Now we are ready to obtain the main result of this paper for the error estimates between the solutions of the continuous and discretized optimal control problems.

Theorem 4.3. *Let* $(u_\omega,\ z_\omega,\ q_\omega) \in L^2(0,T;H^2(D)) \cap H^1(0,T;L^2(D)) \times L^2(0,T;H^2(D)) \cap H^1(0,T;L^2(D)) \times Q_{ad}$ *and* $(u_{\omega,\tau h}, z_{\omega,\tau h}, q_{\omega,\tau h}) \in P_0((0,T],V_h) \times P_0((0,T],V_h) \times Q_{ad}$ *denote the solutions of the problems* (2.8) *and* (3.8), *respectively. Assume that Assumptions (1)–(4) hold. Then a positive constant* C_ω *exists such that*

$$\lambda \parallel q_\omega - q_{\omega,\tau h} \parallel_I + \parallel u_\omega - u_{\omega,\tau h} \parallel_I + \parallel z_\omega - z_{\omega,\tau h} \parallel_I \leq C_\omega(\tau + h^2). \tag{4.5}$$

Moreover

$$\lambda \parallel q - q_{kh} \parallel_\Omega + \parallel u - u_{kh} \parallel_\Omega + \parallel z - z_{kh} \parallel_\Omega \leq C(\tau + h^2). \tag{4.6}$$

5. Numerical results

In this section, to verify numerically the assertion of Theorem 4.3, we use MATLAB R2018b on a PC with an Intel(R) Core(TM)i5-9500 CPU with 3.00GHz of memory to simulate the following examples.

To obtain the expectation of the optimal control, the group is simulated by using the ensemble schemes (3.4) and (3.8). The optimal control problems are solved by the gradient projection method under the ensemble scheme. Define

$$\parallel u \parallel_\Omega \approx \sqrt{\frac{1}{M}\sum_{i=1}^{M}\sum_{n=1}^{N}\tau \parallel u_i^n \parallel^2},$$

where M denotes a set of M random samples selected by the MC sampling.

5.1. Example 1

This numerical example is a constrained problem defined on the unit square $D = [0,1] \times [0,1]$, $q_a = 1$, $q_b = 6$, $\lambda = 0.1$, $a = 1 + (1+\omega)sin(t)sin(x_1x_2)$, $\omega \sim U(0,1)$, and $g = \sin(\pi x_1)\sin(\pi x_2)$. The exact state is $u = (t+1)\sin(\pi x_1)\sin(\pi x_2)$, and the exact adjoint state is $z = 0.5\,(\exp(t)sin(\pi x_1)sin(\pi x_2) - \exp(1)sin(\pi x_1)sin(\pi x_2))$. The optimal control q, u_d, and f can be derived by simple calculation.

To test the convergence order of inequality (4.5), setting the parameter $\bar{a} = 1.23$, we list the computational results in Tables 1 and 2 only for $\omega = 0.1$. We also computed $\omega = 0.5$ and $\omega = 0.9$. The results are similar and have not been listed. In Tables 1 and 2, the results match the theory of the formula (4.5).

Table 1. Error of the control q, the state u, and the adjoint state z with a fixed space step $h = 1/2^7$ ($\omega = 0.1$).

k	$\| q - q_{\tau h} \|_I$	Rate	$\| u - u_{\tau h} \|_I$	Rate	$\| z - z_{\tau h} \|_I$	Rate
1/2	1.517138		0.063970		0.191409	
1/4	0.824601	0.880	0.040922	0.645	0.100292	0.932
1/8	0.428522	0.944	0.021942	0.899	0.050188	0.999
1/16	0.217731	0.977	0.011300	0.957	0.025065	1.002
1/32	0.109788	0.988	0.005784	0.966	0.012556	0.997

Table 2. Error of the control q, the state u, and the adjoint state z with a fixed time step $\tau = 1/10000$ ($\omega = 0.1$).

h	$\| q - q_{\tau h} \|_I$	Rate	$\| u - u_{\tau h} \|_I$	Rate	$\| z - z_{\tau h} \|_I$	Rate
1/2	1.506474		0.490086		0.171945	
1/4	0.513849	1.552	0.184501	1.409	0.061813	1.476
1/8	0.137137	1.906	0.051546	1.840	0.017118	1.852
1/16	0.035334	1.956	0.013264	1.958	0.004413	1.956
1/32	0.009070	1.962	0.003341	1.989	0.001131	1.965

To test the convergence order of inequality (4.6), we list the computational results in Tables 3 and 4. In Tables 3 and 4, the results match the theory of formula (4.6).

Table 3. Error of the control q, the state u, and the adjoint state z with a fixed time step $\tau = 1/10000$, $M = 10$.

h	$\| q - q_{\tau h} \|_\Omega$	Rate	$\| u - u_{\tau h} \|_\Omega$	Rate	$\| z - z_{\tau h} \|_\Omega$	Rate
1/2	1.50906		0.49049		0.17219	
1/4	0.51470	1.552	0.18419	1.413	0.06189	1.476
1/8	0.13734	1.906	0.05141	1.841	0.01713	1.853
1/16	0.03537	1.957	0.01324	1.958	0.00442	1.956
1/32	0.00906	1.965	0.00334	1.985	0.00113	1.967

Table 4. Error of the control q, the state u, and the adjoint state z with a fixed time step $h = 1/2^7$, $M = 10$.

k	$\| q - q_{kh} \|_\Omega$	Rate	$\| u - u_{kh} \|_\Omega$	Rate	$\| z - z_{kh} \|_\Omega$	Rate
1/2	1.49782		0.05943		0.18976	
1/4	0.78166	0.938	0.03514	0.758	0.09619	0.980
1/8	0.39998	0.967	0.01851	0.925	0.04724	1.026
1/16	0.20254	0.982	0.00947	0.966	0.02339	1.014
1/32	0.10180	0.992	0.00485	0.966	0.01167	1.003

The mean of the computational result for the control q is presented in Figure 1.

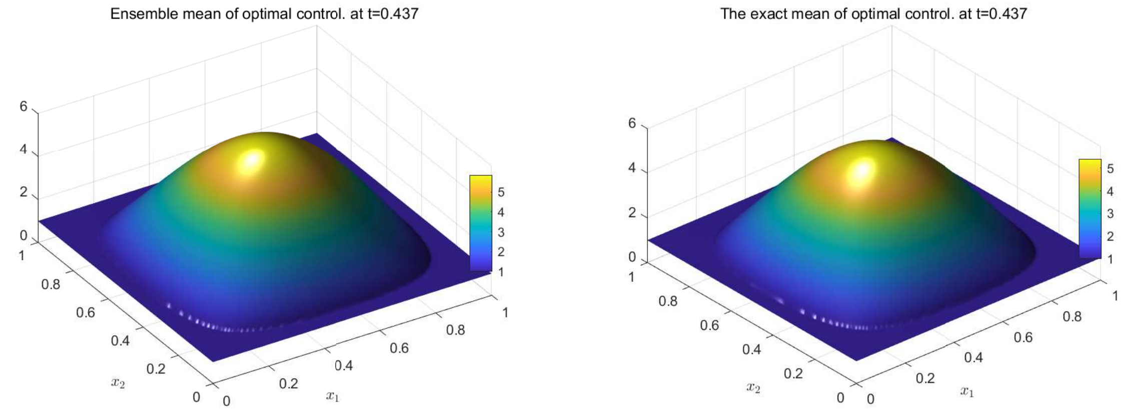

Figure 1. The computational mean of optimal control at $t = 0.4375$ derived by the EMC method (left). The exact mean of optimal control q (right) .

We also use the EMC and MC methods with the gradient projection method to simulate this numerical example. The computational results are listed in Table 5.

From Table 5, the EMC method is superior to the MC method.

Table 5. Computation time and iteration numbers for the average error with q.

M (Sampling number)		10	20	40	80
	Time (s)	45.06	89.21	179.32	359.81
MC	Iterations	30	60	120	240
	Average error	0.0616	0.0613	0.0613	0.0614
	Time(s)	6.02	11.26	22.52	44.76
EMC	Iterations	30	60	120	240
	Average error	0.0622	0.0627	0.0621	0.0624

5.2. *Example 2*

We replace the coefficient of Example 1 as follows:

$$\begin{aligned} a &= 10 + \exp\Big([\omega_1 \cos(\pi x_1) + \omega_2 \sin(\pi x_1)] \exp(-1/10) \\ &\quad +[\omega_3 \cos(\pi x_2) + \omega_4 \sin(\pi x_2)] \exp(-1/10)\Big), \end{aligned}$$

where $\omega_i \sim U(0,1), i = 1,2,3,4$, are independent and uniformly distributed random variables, where we choose $\bar{a} = 13$.

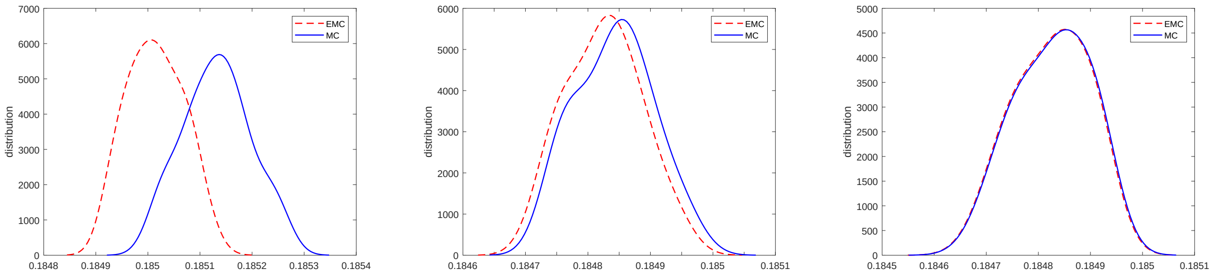

Figure 2. The density function of $J(q, u_h)$ with the control $q = 1$, $h = 1/2^3$, and sampling number $M = 100$ for the time step $\tau = 1/2^7$ (left), $\tau = 1/2^{10}$ (middle), and $\tau = 1/2^{13}$ (right).

We use kernel density estimation to approximate the probability density functions of $J(q = 1, u_h)$ with 100 samples, via the EMC and MC methods. The computation results are listed in Figure 2. From Figure 2, the probability density functions computed by the EMC and MC methods tend to be the same when $\tau \to 0$.

To test the convergence order, the corresponding computational results are listed in Tables 6 and 7. These results verify the theoretical results.

From Figure 2, and Tables 6 and 7, we find that the EMC method is an efficient stochastic numerical method.

Table 6. Error of the control q, the state u, and the adjoint state z with a fixed space step $h = 1/2^7$, where $M = 10$.

k	$\parallel q - q_{\tau h} \parallel_\Omega$	Rate	$\parallel u - u_{\tau h} \parallel_\Omega$	Rate	$\parallel z - z_{\tau h} \parallel_\Omega$	Rate
1/2	1.53054		0.02235		0.20775	
1/4	0.88568	0.789	0.01346	0.732	0.11309	0.877
1/8	0.47681	0.893	0.00724	0.894	0.05831	0.956
1/16	0.24317	0.971	0.00379	0.935	0.02956	0.980
1/32	0.12327	0.980	0.00200	0.922	0.01490	0.989

Table 7. Error of the control q, the state u, and the adjoint state z with a fixed time step $\tau = 1/10000$, where $M = 10$.

h	$\| q - q_{\tau h} \|_\Omega$	Rate	$\| u - u_{\tau h} \|_\Omega$	Rate	$\| z - z_{\tau h} \|_\Omega$	Rate
1/2	1.54878		0.48432		0.17606	
1/4	0.55555	1.479	0.18256	1.408	0.06607	1.414
1/8	0.15100	1.879	0.05139	1.829	0.01860	1.829
1/16	0.03907	1.950	0.01326	1.954	0.00482	1.948
1/32	0.01006	1.957	0.00335	1.987	0.00124	1.959

6. Conclusions

In this paper, we establish a new ensemble MC method to simulate a stochastic parabolic optimal control problem. This new ensemble scheme saves the computational cost by sharing a single coefficient with multiple right-hand sides. Compared with the methods in [20], the new ensemble avoids making subgroups for the stability condition after sampling.

This method can be applied to other stochastic evolutional problems, such as the stochastic convection diffusion problem or the stochastic convection diffusion optimal control problem. This method can also be combined with multi-level methods.

Use of AI tools declaration

The authors declare they have not used artificial intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Granted No. 12461076).

Conflict of interest

The authors declare there is no conflict of interest.

Author contributions

Yan Guo: Methodology, writing–original draft, software; Xianbing Luo: Methodology, review, editing, and supervision; Changlun Ye: Software and analysis.

References

1. M. Hinze, R. Pinnau, M. Ulbrich, S. Ulbrich, *Optimization with PDE Constraints*, Springer, New York, 2009.

2. W. Liu, H. Ma, T. Tang, N. Yan, A posteriori error estimates for discontinuous Galerkin time-stepping method for optimal control problems governed by parabolic equations, *SIAM J. Numer. Anal.*, **42** (2004), 1032–1061. https://doi.org/10.1137/S0036142902397090

3. D. Meidner, B. Vexler, A priori error estimates for space-time finite element discretization of parabolic optimal control problems. Part I: Problems without control constraints, *SIAM J. Control Optim.*, **47** (2008), 1150–1177. https://doi.org/10.1137/070694016

4. F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, American Mathematical Society, Providence, Rhode Island, 2010.

5. A. A. Ali, E. Ullmann, M. Hinze, Multilevel Monte Carlo analysis for optimal control of elliptic PDEs with random coefficients, *SIAM/ASA J. Uncertainty Quantif.*, **5**(2017), 466–492. https://doi.org/10.1137/16M109870X

6. P. Chen, U. Villa, O. Ghattas, Taylor approximation and variance reduction for PDE-constrained optimal control under uncertainty, *J. Comput. Phys.*, **385** (2019), 163–186. https://doi.org/10.1016/j.jcp.2019.01.047

7. P. A. Guth, V. Kaarnioja, F. Kuo, C. Schillings, I. H. Sloan, A quasi-Monte Carlo method for an optimal control problem under uncertainty, *SIAM/ASA J. Uncertainty Quantif.*, **9** (2021), 354–383. https://doi.org/10.1137/19M1294952

8. H. Li, M. Li, X. Luo, S. Xiang, Convergence analysis of finite element approximations for a nonlinear second order hyperbolic optimal control problems, *Networks Heterog. Media*, **19** (2024), 842–866. https://doi.org/10.3934/nhm.2024038

9. J. Yong, X. Luo, S. Sun, C. Ye, Deep mixed residual method for solving PDE-constrained optimization problems, *Comput. Math. Appl.*, **176** (2024), 510–524. https://doi.org/10.1016/j.camwa.2024.11.009

10. A. Kunoth, C. Schwab, Analytic regularity and GPC approximation for control problems constrained by linear parametric elliptic and parabolic PDEs, *SIAM J. Control Optim.*, **51** (2013), 2442–2471. https://doi.org/10.1137/110847597

11. A. Borzi, G. Winckel, Multigrid methods and sparse-grid collocation techniques for parabolic optimal control problems with random coefficients, *SIAM J. Sci. Comput.*, **31** (2009), 2172–2192. https://doi.org/10.1137/070711311

12. B. Gong, T. Sun, W. Shen, W. Liu, A priori error estimate of stochastic Galerkin method for optimal control problem governed by random parabolic PDE with constrained control, *Int. J. Comput. Methods*, **13** (2016), 1650028. https://doi.org/10.1142/S0219876216500286

13. W. Shen, L. Ge, W. Liu, Stochastic Galerkin method for optimal control problem governed by random elliptic PDE with state constraints, *J. Sci. Comput.*, **78** (2019), 1571–1600. https://doi.org/10.1007/s10915-018-0823-6

14. N. Jiang, A second-order ensemble method based on a blended backward differentiation formula timestepping scheme for time-dependent Navier-Stokes equations, *Numer. Methods PDEs*, **33** (2017), 34–61. https://doi.org/10.1002/num.22070

15. M. Gunzburger, N. Jiang, M. Schneier, An ensemble-proper orthogonal decomposition method for the nonstationary Navier-Stokes equations, *SIAM J. Numer. Anal.*, **55** (2017), 286–304. https://doi.org/10.1137/16M1056444

16. N. Jiang, W. Layton, An algorithm for fast calculation of flow ensembles, *Int. J. Uncertainty Quantif.*, **4** (2014), 273–301. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2014007691

17. M. Li, X. Luo, An MLMCE-HDG method for the convection diffusion equation with random diffusivity, *Comput. Math. Appl.*, **127** (2022), 127–143. https://doi.org/10.1016/j.camwa.2022.10.002

18. T. Yao, C. Ye, X. Luo, S. Xiang, An ensemble scheme for the numerical solution of a random transient heat equation with uncertain inputs, *Numer. Algorithms*, **94** (2023), 643–668. https://doi.org/10.1007/s11075-023-01514-z

19. J. Yong, C. Ye, X. Luo, S. Sun, Improved error estimates of ensemble Monte Carlo methods for random transient heat equations with uncertain inputs, *Comp. Appl. Math.*, **44** (2025), 58. https://doi.org/10.1007/s40314-024-03022-9

20. Y. Luo, Z. Wang, An ensemble algorithm for numerical solutions to deterministic and random parabolic PDEs, *SIAM J. Numer. Anal.*, **56** (2018), 859–876. https://doi.org/10.1137/17M1131489

21. L. C. Evans, *Partial Differential Equations (2nd edition)*, American Mathematical Society, Beijing, 2010.

22. V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, Berlin, 2006.

23. M. Hinze, A variational discretization concept in control constrained optimization: The linear quadratic case, *Comput. Optim. Appl.*, **30** (2005), 45–63. https://doi.org/10.1007/s10589-005-4559-5

24. D. A. French, J. T. King, Analysis of a robust finite element approximation for a parabolic equation with rough boundary data, *Math. Comput.*, **60** (1993), 79–104. https://doi.org/10.1090/S0025-5718-1993-1153163-1

25. W. Gong, M. Hinze, Z. Zhou, A priori error analysis for finite element approximation of parabolic optimal control problems with pointwise control, *SIAM J. Control Optim.*, **52** (2014), 97–119. https://doi.org/10.1137/110840133

26. W. Gong, M. Hinze, Z. Zhou, Finite element method and a priori error estimates for dirichlet boundary control problems governed by parabolic PDEs, *J. Sci. Comput.*, **66** (2016), 941–967. https://doi.org/10.1007/s10915-015-0051-2

27. A. Günther, M. Hinze, Elliptic control problems with gradient constraints—variational discrete versus piecewise constant controls, *Comput. Optim. Appl.*, **49** (2011), 549–566. https://doi.org/10.1007/s10589-009-9308-8

Appendix A

A.1. The proof of Theorem 3.1

Proof. (1) For every $1 \leq n \leq N$, according to the Lax–Milgram lemma, we can obtain the existence and uniqueness of the solution for problem (3.4).

(2) The proof is similar to Theorem 1 of [20]. Choosing $\upsilon_h = \bar{u}^n_{\omega,h}$ in problem (3.4), we have

$$\frac{1}{\tau}\left(\bar{u}^n_{\omega,h} - \bar{u}^{n-1}_{\omega,h}, \bar{u}^n_{\omega,h}\right) + \left(\bar{a}\nabla\bar{u}^n_{\omega,h}, \nabla\bar{u}^n_{\omega,h}\right) + \left((a^n_\omega - \bar{a})\nabla\bar{u}^{n-1}_{\omega,h}, \nabla\bar{u}^n_{\omega,h}\right) = \left(\frac{1}{\tau}\int_{I_n}(\bar{q} + \bar{f}_\omega)dt, \bar{u}^n_{\omega,h}\right).$$

Multiplying both sides by τ and using the polarization identity and expression (2.2), we get

$$\begin{aligned}&\frac{1}{2}\parallel \bar{u}^n_{\omega,h} \parallel^2 - \frac{1}{2}\parallel \bar{u}^{n-1}_{\omega,h} \parallel^2 + \frac{1}{2}\parallel \bar{u}^n_{\omega,h} - \bar{u}^{n-1}_{\omega,h} \parallel^2 + k\bar{a}\parallel \nabla\bar{u}^n_{\omega,h} \parallel^2\\ &\leq -\tau((a^n_\omega - \bar{a})\nabla\bar{u}^{n-1}_{\omega,h}, \nabla\bar{u}^n_{\omega,h}) + \left(\int_{I_n}\bar{q} + \bar{f}_\omega dt, \bar{u}^n_{\omega,h}\right).\end{aligned} \tag{A.1}$$

The Cauchy–Schwarz inequality, Young's inequality, and $\mu, \alpha > 0$ imply

$$\begin{aligned}&\tau \mid ((a^n_\omega - \bar{a})\nabla\bar{u}^{n-1}_{\omega,h}, \nabla\bar{u}^n_{\omega,h}) \mid\leq \tau\theta \parallel \nabla\bar{u}^{n-1}_{\omega,h} \parallel \quad \parallel \nabla\bar{u}^n_{\omega,h} \parallel\\ &\leq \tau\theta\left(\frac{1}{2\mu}\parallel \nabla\bar{u}^{n-1}_{\omega,h} \parallel^2 + \frac{\mu}{2}\parallel \nabla\bar{u}^n_{\omega,h} \parallel^2\right).\end{aligned} \tag{A.2}$$

Utilizing Poincare's inequality and Young's inequality, we have

$$\left|\left(\int_{I_n}\bar{q}dt, \bar{u}^n_{\omega,h}\right)\right| \leq \tau^{1/2}\parallel \bar{q} \parallel_{I_n}\parallel \bar{u}^n_{\omega,h} \parallel\leq C\left(\frac{1}{4\alpha}\parallel \bar{q} + \bar{f}_\omega \parallel^2_{I_n} + \alpha k\parallel \nabla\bar{u}^n_{\omega,h} \parallel^2\right). \tag{A.3}$$

Inserting the estimates of inequalities (A.2) and (A.3) into inequality (A.1), we have

$$\begin{aligned}&\frac{1}{2}\left(\parallel \bar{u}^n_{\omega,h} \parallel^2 - \parallel \bar{u}^{n-1}_{\omega,h} \parallel^2\right) + \tau\left[\bar{a} - \alpha - \left(\frac{\mu}{2} + \frac{1}{2\mu}\right)\theta\right]\parallel \nabla\bar{u}^n_{\omega,h} \parallel^2\\ &+\frac{\tau}{2\mu}\theta\left(\parallel \nabla\bar{u}^n_{\omega,h} \parallel^2 - \parallel \nabla\bar{u}^{n-1}_{\omega,h} \parallel^2\right) \leq \frac{C}{4\alpha}\parallel \bar{q} + \bar{f}_\omega \parallel^2_{I_n}.\end{aligned} \tag{A.4}$$

Multiplying both sides by 2, choosing $\mu = 1$ and $\alpha = \frac{\bar{a}-\theta}{2}$, and using expression (3.5), summing n from 1 to N, we can obtain

$$\begin{aligned}&\parallel \bar{u}^N_{\omega,h} \parallel^2 - \parallel \bar{u}^0_{\omega,h} \parallel^2 + (\bar{a} - \theta)\tau\sum_{n=1}^{N}\parallel \nabla\bar{u}^n_{\omega,h} \parallel^2\\ &+\sum_{n=1}^{N}\tau\theta\left(\parallel \nabla\bar{u}^n_{\omega,h} \parallel^2 - \parallel \nabla\bar{u}^{n-1}_{\omega,h} \parallel^2\right) \leq \frac{C}{\bar{a} - \theta}\parallel \bar{q} + \bar{f}_\omega \parallel^2_I.\end{aligned} \tag{A.5}$$

According to inequality (3.2), we have

$$\begin{aligned}&\parallel \nabla\bar{u}^0_{\omega,h} \parallel=\parallel \nabla\mathcal{R}_{\omega,h}g \parallel\leq \frac{a_{\max}}{a_{\min}}\parallel \nabla g \parallel,\\ &\parallel \bar{u}^0_{\omega,h} \parallel\leq\parallel \nabla\bar{u}^0_{\omega,h} \parallel\leq \frac{a_{\max}}{a_{\min}}\parallel \nabla g \parallel.\end{aligned} \tag{A.6}$$

Combined with this, from inequality (A.5), we can get

$$\| \bar{u}^{M}_{\omega,h} \|^2 + \theta\tau \| \nabla \bar{u}^{N}_{\omega,h} \|^2 + (\bar{a}-\theta)\tau \sum_{n=1}^{N} \| \nabla \bar{u}^{n}_{\omega,h} \|^2 \le \frac{C}{(\bar{a}-\theta)} \| \bar{q} + \bar{f}_\omega \|_I^2 + (\theta\tau + 1)\frac{a^2_{\max}}{a^2_{\min}} \| \nabla g \|^2, \quad \text{(A.7)}$$

which implies inequality (3.6).

A.2. The proof of Theorem 3.2

Proof. Taking $\upsilon_h = \bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h}$ in Eq (3.1), we can obtain

$$\begin{aligned} &\left(\frac{\bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h}}{\tau}, \bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h} \right) + (\bar{a}\nabla \bar{u}^{n}_{\omega,h}, \nabla(\bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h})) \\ &+ ((a^n_\omega - \bar{a})\nabla \bar{u}^{n-1}_{\omega,h}, \nabla(\bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h})) = \left(\frac{1}{\tau}\int_{I_n} (\bar{q} + \bar{f}_\omega) dt, \bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h} \right). \end{aligned} \quad \text{(A.8)}$$

Rearranging Eq (A.8), we derive

$$\frac{1}{\tau} \| \bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h} \|^2 + \bar{a} \| \nabla(\bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h}) \|^2 = (a^n_\omega \nabla \bar{u}^{n-1}_{\omega,h}, \nabla(\bar{u}^{n-1}_{\omega,h} - \bar{u}^{n}_{\omega,h})) + \left(\frac{1}{\tau}\int_{I_n} (\bar{q} + \bar{f}_\omega) dt, \bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h} \right). \quad \text{(A.9)}$$

Using the polarization identity and Young's inequality, the right-hand side of Eq (A.9) yields

$$\begin{aligned} &(a^n_\omega \nabla \bar{u}^{n-1}_{\omega,h}, \nabla(\bar{u}^{n-1}_{\omega,h} - \bar{u}^{n}_{\omega,h})) + \left(\frac{1}{\tau}\int_{I_n} (\bar{q} + \bar{f}_\omega) dt, \bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h} \right) \le \frac{1}{2} \| a^{n\,1/2}_\omega \nabla \bar{u}^{n-1}_{\omega,h} \|^2 \\ &+ \frac{1}{2} \| a^{n\,1/2}_\omega \nabla(\bar{u}^{n-1}_{\omega,h} - \bar{u}^{n}_{\omega,h}) \|^2 - \frac{1}{2} \| a^{n\,1/2}_\omega \nabla \bar{u}^{n}_{\omega,h} \|^2 + \frac{1}{2} \| \bar{q} + \bar{f}_\omega \|^2_{I_n} + \frac{1}{2\tau} \| \bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h} \|^2 . \end{aligned} \quad \text{(A.10)}$$

Using expressions (2.2) and (3.5), we obtain

$$\begin{aligned} &\frac{1}{2\tau} \| \bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h} \|^2 + \left(\bar{a} - \frac{a_{\max}}{2} \right) \| \nabla(\bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h}) \|^2 \\ &\le \frac{1}{2} \| \sqrt{a^{n-1}_\omega} \nabla \bar{u}^{n-1}_{\omega,h} \|^2 - \frac{1}{2} \| \sqrt{a^{n}_\omega} \nabla \bar{u}^{n}_{\omega,h} \|^2 + \frac{1}{2} \| \sqrt{a^{n}_\omega} \nabla \bar{u}^{n-1}_{\omega,h} \|^2 \\ &- \frac{1}{2} \| \sqrt{a^{n-1}_\omega} \nabla \bar{u}^{n-1}_{\omega,h} \|^2 + \frac{1}{2} \| \bar{q} + \bar{f}_\omega \|^2_{I_n} . \end{aligned}$$

According to inequality (2.3), we get

$$\begin{aligned} &\frac{1}{2\tau} \| \bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h} \|^2 + \left(\bar{a} - \frac{a_{\max}}{2} \right) \| \nabla(\bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h}) \|^2 \\ &+ \frac{1}{2} \| \sqrt{a^{n}_\omega} \nabla \bar{u}^{n}_{\omega,h} \|^2 - \frac{1}{2} \| \sqrt{a^{n-1}_\omega} \nabla \bar{u}^{n-1}_{\omega,h} \|^2 \le \frac{1}{2} L\tau \| \nabla \bar{u}^{n-1}_{\omega,h} \|^2 + \frac{1}{2} \| \bar{q} + \bar{f}_\omega \|^2_{I_n} . \end{aligned} \quad \text{(A.11)}$$

Summing n from 1 to N and multiplying both sides by 2, we obtain

$$\begin{aligned} &\frac{1}{\tau} \sum_{n=1}^{N} \| \bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h} \|^2 + (2\bar{a} - a_{\max}) \sum_{n=1}^{N} \| \nabla(\bar{u}^{n}_{\omega,h} - \bar{u}^{n-1}_{\omega,h}) \|^2 + a_{\min} \| \nabla \bar{u}^{N}_{\omega,h} \|^2 \\ &\le a_{\max} \| \nabla \bar{u}^{0}_{\omega,h} \|^2 + L\tau \sum_{n=1}^{N} \| \nabla \bar{u}^{n-1}_{\omega,h} \|^2 + \| \bar{q} + \bar{f}_\omega \|^2_I . \end{aligned}$$

According to Theorem 3.1, we have

$$\begin{aligned}&\frac{1}{\tau}\sum_{n=1}^{M}\parallel \bar{u}_{\omega,h}^{n}-\bar{u}_{\omega,h}^{n-1}\parallel^2+(2\bar{a}-a_{\max})\sum_{n=1}^{M}\parallel \nabla(\bar{u}_{\omega,h}^{n}-\bar{u}_{\omega,h}^{n-1})\parallel^2+a_{\min}\parallel \nabla\bar{u}_{\omega,h}^{M}\parallel^2\\ \leq&\left(\frac{\tau LC}{\bar{a}-\theta}+a_{\max}\right)\parallel \nabla\bar{u}_{\omega,h}^{0}\parallel^2+\left(\frac{L}{\bar{a}-\theta}\right)\parallel \bar{u}_{\omega,h}^{0}\parallel^2+\left(\frac{LC}{\bar{a}-\theta}+1\right)\parallel \bar{q}\parallel_I^2 .\end{aligned}$$

Combining inequalities (A.6), we can get the desired result of inequality (3.7).

A.3. The proof of Theorem 3.4

Proof. Define the discrete adjoint state $\{z_{\omega,h}^{n-1}\}_{n=1}^{N}\in V_h$ such that

$$\begin{aligned}-\left(\frac{z_{\omega,h}^{n}-z_{\omega,h}^{n-1}}{\tau},v_h\right)+(\bar{a}\nabla z_{\omega,h}^{n-1},\nabla v_h)+\left((a_\omega^{n+1}-\bar{a})\nabla z_{\omega,h}^{n},\nabla v_h\right)&=(u_{\omega,h}^{n}-\widehat{u_d}^{n},v_h),\forall v_h\in V_h,\\ z_{\omega,h}^{N}=0,a_\omega^{N+1}=0,\quad n=1,\cdots,N.&\end{aligned}\tag{A.12}$$

For every $1\leq n\leq N$, according to the Lax–Milgram lemma, we can obtain the existence and uniqueness of a solution for Eq (A.12). Due to the linearity of the state equation, the cost function $J_{\omega,\tau h}$ is strictly convex. Hence, via Theorem 1.46 of [1], a unique $q_{\omega,\tau h}\in Q_{ad}$ exists such that

$$0\leq(\lambda q_{\omega,\tau h},\bar{q}-q_{\omega,\tau h})_I+\tau\sum_{n=1}^{N}(u_{\omega,h}^{n}-\widehat{u_d}^{n},u_{\omega,h}^{n}(\bar{q})-u_{\omega,h}^{n}),\quad\forall\bar{q}\in Q_{ad}.\tag{A.13}$$

Taking $v_h=u_\omega^n(\bar{q},q_{\omega,\tau h})=u_{\omega,h}^{n}(\bar{q})-u_{\omega,h}^{n}$ in Eq (A.12) and summing n from 1 to N, we obtain

$$\begin{aligned}&\tau\sum_{n=1}^{N}\left(u_{\omega,h}^{n}-\widehat{u_d}^{n},u_\omega^n(\bar{q},q_{\omega,\tau h})\right)=\sum_{n=1}^{N}-\left(z_{\omega,h}^{n}-z_{\omega,h}^{n-1},u_\omega^n(\bar{q},q_{\omega,\tau h})\right)+\tau\sum_{n=1}^{N}\left(\bar{a}\nabla z_{\omega,h}^{n-1},\nabla u_\omega^n(\bar{q},q_{\omega,\tau h})\right)\\ &+\tau\sum_{n=1}^{N}\left((a_\omega^{n+1}-\bar{a})\nabla z_{\omega,h}^{n},\nabla u_\omega^n(\bar{q},q_{\omega,\tau h})\right)=\sum_{n=1}^{N}\left(z_{\omega,h}^{n-1},u_\omega^n(\bar{q},q_{\omega,\tau h})-u_\omega^{n-1}(\bar{q},q_{\omega,\tau h})\right)\\ &+\tau\sum_{n=1}^{N}\left(\bar{a}\nabla z_{\omega,h}^{n-1},\nabla u_\omega^n(\bar{q},q_{\omega,\tau h})\right)+\tau\sum_{n=1}^{N}\left((a_\omega^{n}-\bar{a})\nabla z_{\omega,h}^{n-1},\nabla u_\omega^{n-1}(\bar{q},q_{\omega,\tau h})\right).\end{aligned}\tag{A.14}$$

According to problem (3.4), taking $v_h=z_{\omega,h}^{n-1}$ and summing n from 1 to N, we can obtain

$$\begin{aligned}&\sum_{n=1}^{N}\left(u_\omega^n(\bar{q},q_{\omega,\tau h})-u_\omega^{n-1}(\bar{q},q_{\omega,\tau h}),z_{\omega,h}^{n-1}\right)+\tau\sum_{n=1}^{N}(\bar{a}\nabla u_\omega^n(\bar{q},q_{\omega,\tau h}),\nabla z_{\omega,h}^{n-1})\\ &+\tau\sum_{n=1}^{N}\left((a_\omega^{n}-\bar{a})\nabla u_\omega^{n-1}(\bar{q},q_{\omega,\tau h}),\nabla z_{\omega,h}^{n-1}\right)=\sum_{n=1}^{N}(\bar{q}-q_{\omega,\tau h},z_{\omega,h}^{n-1})_{I_n}=(\bar{q}-q_{\omega,\tau h},z_{\omega,\tau h})_I.\end{aligned}\tag{A.15}$$

Therefore, we obtain

$$\tau\sum_{n=1}^{N}\left(u_{\omega,h}^{n}-\widehat{u_d}^{n},u_{\omega,h}^{n}(\bar{q})-u_{\omega,h}^{n}\right)=(\bar{q}-q_{\omega,\tau h},z_{\omega,\tau h})_I.\tag{A.16}$$

Combining Eqs (A.16) and (A.13), we can obtain the desired result of inequality (3.11).

A.4. The proof of Theorem 4.1

Proof. The following proof idea comes from [24–26]. For a fixed ω, consider the dual problem

$$\begin{cases} -\partial_t\varphi_\omega - \nabla\cdot(a(\omega,t,x)\nabla\varphi_\omega) = f, & t,x \in (0,T]\times D, \\ \varphi_\omega = 0, & x\in\partial D, t\in[0,T], \\ \varphi_\omega(T) = 0, & x\in D. \end{cases}$$

Then, according to [21] for $f \in L^2(0,T;L^2(D))$ and Assumptions (1)–(3), we have $\varphi_\omega \in L^2(0,T;H^2(D)) \cap H^1(0,T;L^2(D))$ and

$$\| \varphi_\omega(0) \|_1 \leq \| f \|_I, \tag{A.17}$$

$$\| \varphi_\omega \|_{L^2(0,T;H^2(D))} + \| \partial_t\varphi_\omega \|_I \leq C \| f \|_I, \tag{A.18}$$

where C is a positive constant independent of ω, with $\varphi_\omega \in L^2(0,T;H^2(D)) \cap H^1(0,T;L^2(D)) \hookrightarrow C([0,T],H^1(D))$. We use φ_ω^n to denote $\varphi_\omega(t^n)$ $(n = 0,1,\cdots,N)$. In addition, we have

$$\begin{aligned}
&(u_\omega(q_{\omega,\tau h}) - u_{\omega,\tau h}, f)_I = (u_\omega(q_{\omega,\tau h}), -\partial_t\varphi_\omega - \nabla\cdot(a_\omega\nabla\varphi_\omega))_I - (u_{\omega,\tau h}, -\partial_t\varphi_\omega - \nabla\cdot(a_\omega\nabla\varphi_\omega))_I \\
&= (g(x),\varphi_\omega(0)) + (q_{\omega,\tau h} + \bar{f}_\omega, \varphi_\omega)_I - (u_{\omega,\tau h}, -\partial_t\varphi_\omega - \nabla\cdot(a_\omega\nabla\varphi_\omega))_I \\
&= (g(x),\varphi_\omega(0)) + (q_{\omega,\tau h} + \bar{f}_\omega, \varphi_\omega)_I - (g_h(x),\varphi_\omega(0)) - \sum_{n=1}^{N}(u_{\omega,h}^n - u_{\omega,h}^{n-1}, \varphi_\omega^{n-1}) - (a_\omega\nabla u_{\omega,\tau h}, \nabla\varphi_\omega)_I \\
&= (g(x) - g_h(x),\varphi_\omega(0)) + (q_{\omega,\tau h} + \bar{f}_\omega, \varphi_\omega)_I - \sum_{n=1}^{N}(u_{\omega,h}^n - u_{\omega,h}^{n-1}, \varphi_\omega^{n-1}) - (a_\omega\nabla u_{\omega,\tau h}, \nabla\varphi_\omega)_I \\
&= (g(x) - g_h(x),\varphi_\omega(0)) + \sum_{n=1}^{N}(\widehat{q_{\omega,\tau h}}^n + \bar{f}_\omega, \varphi_\omega)_{I_n} - \sum_{n=1}^{N}(u_{\omega,h}^n - u_{\omega,h}^{n-1}, \varphi_\omega^{n-1}) - \sum_{n=1}^{N}(a_\omega\nabla u_{\omega,\tau h}, \nabla\varphi_\omega)_{I_n}.
\end{aligned} \tag{A.19}$$

Note that the $q_{\omega,\tau h}$ is piecewise constant with respect to time but, in general, $q_{\omega,\tau h}$ is not a finite element function with respect to space (see, e.g., [1, 23, 27]). We substitute $\upsilon_h = \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^n$ (where $\widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^n = \frac{1}{\tau}\int_{I_n}\mathcal{R}_{\omega,h}\varphi_\omega dt$) into problem (3.9) and obtain

$$\frac{1}{\tau}(u_{\omega,h}^n - u_{\omega,h}^{n-1}, \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^n) + (\bar{a}\nabla u_{\omega,h}^n, \nabla\widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^n) + ((a_\omega^n - \bar{a})\nabla u_{\omega,h}^{n-1}, \nabla\widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^n) = (\widehat{q_{\omega,\tau h}}^n + \widehat{\bar{f}_\omega}^n, \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^n).$$

Summing n from 1 to N, we get

$$\begin{aligned}
&\sum_{n=1}^{N}(u_{\omega,h}^n - u_{\omega,h}^{n-1}, \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^n) + \sum_{n=1}^{N}(\bar{a}\nabla u_{\omega,h}^n, \nabla\widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^n)_{I_n} \\
&+ \sum_{n=1}^{N}((a_\omega^n - \bar{a})\nabla u_{\omega,h}^{n-1}, \nabla\widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^n)_{I_n} - \sum_{n=1}^{N}(\widehat{q_{\omega,\tau h}}^n + \widehat{\bar{f}_\omega}^n, \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^n)_{I_n} = 0.
\end{aligned} \tag{A.20}$$

Substituting Eq (A.20) into Eq (A.19) yields

$$\begin{aligned}
&(u_\omega(q_{\omega,\tau h}) - u_{\omega,\tau h}, f)_I\\
=&\underbrace{(g(x) - g_h(x), \varphi_\omega(0))}_{①} + \underbrace{\sum_{n=1}^{N}(\widehat{q_{\omega,\tau h}}^{\,n} + \bar{f}_\omega, \varphi_\omega - \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n})_{I_n}}_{②} + \underbrace{\sum_{n=1}^{N}(u^n_{\omega,h} - u^{n-1}_{\omega,h}, \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n} - \varphi^{n-1}_\omega)}_{③}\\
&+\underbrace{\sum_{n=1}^{N}(\bar{a}\nabla u^n_{kh}, \nabla\widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n})_{I_n} + \sum_{n=1}^{N}((a^n_\omega - \bar{a})\nabla u^{n-1}_{\omega,h}, \nabla\widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n})_{I_n} - \sum_{n=1}^{N}(a_\omega\nabla u_{\omega,\tau h}, \nabla\varphi_\omega)_{I_n}}_{④}.
\end{aligned}\tag{A.21}$$

By inequalities (3.3) and (A.17), we derive

$$① \leq\mid (g(x) - g_h(x), \varphi_\omega(0)) \mid\leq \|g(x) - g_h(x)\|_{H^{-1}}\ \|\varphi_\omega(0)\|_1 \leq h^2 \parallel g \parallel_1 \quad \parallel f \parallel_I . \tag{A.22}$$

According to Eq (2.1) and inequality (A.18), Theorem 3.5, and the Cauchy inequality, we can obtain

$$\begin{aligned}
② \leq& \mid \sum_{n=1}^{N}(\widehat{q_{\omega,\tau h}}^{\,n} + \bar{f}_\omega, \varphi_\omega - \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n})_{I_n} \mid\leq\mid \sum_{n=1}^{N}(\widehat{q_{\omega,\tau h}}^{\,n}, \varphi_\omega - \mathcal{R}_{\omega,h}\varphi_\omega + \mathcal{R}_{\omega,h}\varphi_\omega - \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n})_{I_n} \mid\\
&+ \mid \sum_{n=1}^{N}(\bar{f}_\omega, \varphi_\omega - \widehat{\varphi_\omega}^{\,n} + \widehat{\varphi_\omega}^{\,n} - \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n})_{I_n} \mid\\
=& \mid \sum_{n=1}^{N}(\widehat{q_{\omega,\tau h}}^{\,n}, \varphi_\omega - \mathcal{R}_{\omega,h}\varphi_\omega)_{I_n} \mid + \mid \sum_{n=1}^{N}(\bar{f}_\omega, \varphi_\omega - \widehat{\varphi_\omega}^{\,n} + \widehat{\varphi_\omega}^{\,n} - \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n})_{I_n} \mid\\
\leq& (\parallel q_{\omega,\tau h} \parallel_I + \parallel \bar{f}_\omega \parallel_I) \parallel \varphi_\omega - \mathcal{R}_{\omega,h}\varphi_\omega \parallel_I + \parallel \bar{f}_\omega \parallel_I \left(\sum_{n=1}^{N} \parallel \varphi_\omega - \widehat{\varphi_\omega}^{\,n} \parallel^2_{I_n}\right)^{1/2}\\
\leq& Ch^2(\parallel q_{\omega,\tau h} \parallel_I + \parallel \bar{f}_\omega \parallel_I) \parallel \varphi_\omega \parallel_{L^2(0,T;H^2(D))} + C\tau \parallel \bar{f}_\omega \parallel_I \parallel \partial_t\varphi_\omega \parallel_I\\
\leq& C\left[h^2(\parallel q_{\omega,\tau h} \parallel_I + \parallel \bar{f}_\omega \parallel_I) + \tau \parallel \bar{f}_\omega \parallel_I\right] \parallel f \parallel_I .
\end{aligned}\tag{A.23}$$

For the third term, we have

$$\begin{aligned}
③ =& \sum_{n=1}^{N}(u^n_{\omega,h} - u^{n-1}_{\omega,h}, \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n} - \varphi^{n-1}_\omega)\\
=& \sum_{n=1}^{N}\left(u^n_{\omega,h} - u^{n-1}_{\omega,h}, \frac{1}{\tau}\int_{I_n} \mathcal{R}_{\omega,h}\varphi_\omega - \varphi_\omega dt + \frac{1}{\tau}\int_{I_n} \varphi_\omega dt - \varphi^{n-1}_\omega\right)\\
=& \underbrace{\sum_{n=1}^{N}\left(u^n_{\omega,h} - u^{n-1}_{\omega,h}, \frac{1}{\tau}\int_{I_n} \mathcal{R}_{\omega,h}\varphi_\omega - \varphi_\omega dt\right)}_{I} + \underbrace{\sum_{n=1}^{N}\left(u^n_{\omega,h} - u^{n-1}_{\omega,h}, \frac{1}{\tau}\int_{I_n} \varphi_\omega dt - \varphi^{n-1}_\omega\right)}_{II}.
\end{aligned}$$

For the first part (I), the Cauchy inequality implies

$$I \leq \frac{1}{\tau}\left(\sum_{n=1}^{N} \parallel u^n_{\omega,h} - u^{n-1}_{\omega,h} \parallel^2\right)^{\frac{1}{2}} \left(\sum_{n=1}^{N}\left(\int_{I_n} \parallel \mathcal{R}_{\omega,h}\varphi_\omega - \varphi_\omega \parallel\right)^2\right)^{\frac{1}{2}}.$$

By Hölder's inequality, we can get

$$\int_{I_n} \| \mathcal{R}_{\omega,h}\varphi_\omega - \varphi_\omega \| \leq k^{\frac{1}{2}} \| \mathcal{R}_{\omega,h}\varphi_\omega - \varphi_\omega \|_{I_n} .$$

Therefore, we can deduce

$$I \leq \frac{1}{\tau^{1/2}} \left(\sum_{n=1}^{N} \| u_{\omega,h}^{n} - u_{\omega,h}^{n-1} \|^2 \right)^{\frac{1}{2}} \| \mathcal{R}_{\omega,h}\varphi_\omega - \varphi_\omega \|_I .$$

By Theorem 3.2, 3.5, and the error estimates of inequalities (3.3) of the Ritz projection, we have

$$I \leq C_2 h^2 \| \varphi_\omega \|_{L^2(0,T;H^2(D))} .$$

For the second part (II), standard error analysis yields (see, e.g., [22] and [24])

$$II \leq \tau^{1/2} \left(\sum_{n=1}^{N} \| u_{\omega,h}^{n} - u_{\omega,h}^{n-1} \|^2 \right)^{\frac{1}{2}} \| \partial_t \varphi_\omega \|_I .$$

Utilizing Theorem 3.1, we get

$$II \leq C_1 \tau \| \partial_t \varphi_\omega \|_I .$$

Furthermore, by inequality (A.18), we can derive

$$\begin{aligned} \textcircled{3} \leq & \max\{C_1, C_2\} h^2 \| \varphi_\omega \|_{L^2(0,T;H^2(D))} + \max\{C_1, C_2\} \tau \| \partial_t \varphi_\omega \|_I \\ \leq & C(\tau + h^2) \| f \|_I . \end{aligned} \tag{A.24}$$

For the fourth term, we have

$$\begin{aligned} \textcircled{4} = & \sum_{n=1}^{N} (\bar{a} \nabla u_{\omega,h}^{n}, \nabla \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n})_{I_n} + \sum_{n=1}^{N} ((a_\omega^n - \bar{a}) \nabla u_{\omega,h}^{n-1}, \nabla \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n})_{I_n} - \sum_{n=1}^{N} (a_\omega^n \nabla u_{\omega,\tau h}, \nabla \varphi_\omega)_{I_n} \\ & + \sum_{n=1}^{N} ((a_\omega^n - a_\omega) \nabla u_{\omega,\tau h}, \nabla \varphi_\omega)_{I_n} \\ = & \sum_{n=1}^{N} (\bar{a} \nabla u_{\omega,h}^{n}, \nabla \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n})_{I_n} - \sum_{n=1}^{N} (\bar{a} \nabla u_{\omega,h}^{n}, \nabla \mathcal{R}_{\omega,h}\varphi_\omega)_{I_n} + \sum_{n=1}^{N} (\bar{a} \nabla u_{\omega,h}^{n}, \nabla \mathcal{R}_{\omega,h}\varphi_\omega)_{I_n} \\ & + \sum_{n=1}^{N} ((a_\omega^n - \bar{a}) \nabla u_{\omega,h}^{n-1}, \nabla \widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n})_{I_n} - \sum_{n=1}^{N} ((a_\omega^n - \bar{a}) \nabla u_{\omega,h}^{n-1}, \nabla \mathcal{R}_{\omega,h}\varphi_\omega)_{I_n} + \sum_{n=1}^{N} ((a_\omega^n \\ & - \bar{a}) \nabla u_{\omega,h}^{n-1}, \nabla \mathcal{R}_{\omega,h}\varphi_\omega)_{I_n} - \sum_{n=1}^{N} (a_\omega^n \nabla u_{\omega,\tau h}, \nabla \varphi_\omega)_{I_n} + \sum_{n=1}^{N} ((a^n - a_\omega) \nabla u_{\omega,\tau h}, \nabla \varphi_\omega)_{I_n}. \end{aligned}$$

Using Eq (2.1), we can obtain

$$\sum_{n=1}^{N} (\bar{a} \nabla u_{\omega,h}^{n}, \nabla (\widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n} - \mathcal{R}_{\omega,h}\varphi_\omega))_{I_n} = 0, \quad \sum_{n=1}^{N} ((a_\omega^n - \bar{a}) \nabla u_{\omega,h}^{n-1}, \nabla (\widehat{\mathcal{R}_{\omega,h}\varphi_\omega}^{\,n} - \mathcal{R}_{\omega,h}\varphi_\omega))_{I_n} = 0.$$

Hence, we can obtain

$$\begin{aligned}
\textcircled{4} &= \sum_{n=1}^{N}(\bar{a}\nabla u_{\omega,h}^{n}, \nabla\mathcal{R}_{\omega,h}\varphi_{\omega})_{I_n} + \sum_{n=1}^{N}((a_{\omega}^{n} - \bar{a})\nabla u_{\omega,h}^{n-1}, \nabla\mathcal{R}_{\omega,h}\varphi_{\omega})_{I_n} \\
&\quad - \sum_{n=1}^{N}(a_{\omega}^{n}\nabla u_{\omega,\tau h}, \nabla\varphi_{\omega})_{I_n} + \sum_{n=1}^{N}((a_{\omega}^{n} - a_{\omega})\nabla u_{\omega,\tau h}, \nabla\varphi_{\omega})_{I_n} \\
&= \sum_{n=1}^{N}(\bar{a}\nabla u_{\omega,h}^{n}, \nabla\mathcal{R}_{\omega,h}\varphi_{\omega})_{I_n} + \sum_{n=1}^{N}((a_{\omega}^{n} - \bar{a})\nabla u_{\omega,h}^{n}, \nabla\mathcal{R}_{\omega,h}\varphi_{\omega})_{I_n} - \sum_{n=1}^{N}(a_{\omega}^{n}\nabla u_{\omega,\tau h}, \nabla\varphi_{\omega})_{I_n} \\
&\quad + \sum_{n=1}^{N}((a_{\omega}^{n} - a_{\omega})\nabla u_{\omega,\tau h}, \nabla\varphi_{\omega})_{I_n} + \sum_{n=1}^{N}((a_{\omega}^{n} - \bar{a})\nabla(u_{\omega,h}^{n-1} - u_{\omega,h}^{n}), \nabla\mathcal{R}_{\omega,h}\varphi_{\omega})_{I_n} \\
&= \sum_{n=1}^{N}(a_{\omega}^{n}\nabla u_{\omega,\tau h}, \nabla(\mathcal{R}_{\omega,h}\varphi_{\omega} - \varphi_{\omega}))_{I_n} + \sum_{n=1}^{N}((a_{\omega}^{n} - a_{\omega})\nabla u_{\omega,\tau h}, \nabla\varphi_{\omega})_{I_n} \\
&\quad + \sum_{n=1}^{N}((a_{\omega}^{n} - \bar{a})\nabla(u_{\omega,h}^{n-1} - u_{\omega,h}^{n}), \nabla\mathcal{R}_{\omega,h}\varphi_{\omega})_{I_n} \\
&= \sum_{n=1}^{N}(a_{\omega}^{n}\nabla u_{\omega,\tau h}, \nabla(\mathcal{R}_{\omega,h}\varphi_{\omega} - \varphi_{\omega}))_{I_n} + \sum_{n=1}^{N}((a_{\omega}^{n} - a_{\omega})\nabla u_{\omega,\tau h}, \nabla\varphi_{\omega})_{I_n} \\
&\quad + \sum_{n=1}^{N}((a_{\omega}^{n} - \bar{a})\nabla(u_{\omega,h}^{n-1} - u_{\omega,h}^{n}), \nabla(\mathcal{R}_{\omega,h}\varphi_{\omega} - \varphi_{\omega}))_{I_n} + \sum_{n=1}^{N}((a_{\omega}^{n} - \bar{a})\nabla(u_{\omega,h}^{n-1} - u_{\omega,h}^{n}), \nabla\varphi_{\omega})_{I_n}.
\end{aligned}$$

The property of the Ritz projection of Eq (3.1) implies

$$\sum_{n=1}^{N}(a_{\omega}^{n}\nabla u_{\omega,\tau h}, \nabla(\mathcal{R}_{\omega,h}\varphi_{\omega} - \varphi_{\omega}))_{I_n} = 0.$$

Hence, we have

$$\begin{aligned}
\textcircled{4} &= \sum_{n=1}^{N}((a_{\omega}^{n} - a_{\omega})\nabla u_{\omega,\tau h}, \nabla\varphi_{\omega})_{I_n} + \sum_{n=1}^{N}((a_{\omega}^{n} - \bar{a})\nabla(u_{\omega,h}^{n-1} - u_{\omega,h}^{n}), \nabla(\mathcal{R}_{\omega,h}\varphi_{\omega} - \varphi_{\omega}))_{I_n} \\
&\quad + \sum_{n=1}^{N}((a_{\omega}^{n} - \bar{a})\nabla(u_{\omega,h}^{n-1} - u_{\omega,h}^{n}), \nabla\varphi_{\omega})_{I_n}.
\end{aligned}$$

Using the Cauchy inequality and the property of Lipschitz continuity, we deduce

$$\left|\sum_{n=1}^{N}((a_{\omega}^{n} - a_{\omega})\nabla u_{\omega,\tau h}, \nabla\varphi_{\omega})_{I_n}\right| \leq L\tau\left(\sum_{n=1}^{N}\tau \parallel \nabla u_{\omega,h}^{n} \parallel^{2}\right)^{\frac{1}{2}} \parallel \nabla\varphi_{\omega} \parallel_{I} .$$

Further, using the Cauchy inequality, we can obtain

$$\begin{aligned}
&\left|\sum_{n=1}^{N}((a_\omega^n-\bar{a})\nabla(u_{\omega,h}^{n-1}-u_{\omega,h}^n),\nabla(\mathcal{R}_{\omega,h}\varphi_\omega-\varphi_\omega))_{I_n}\right|\\
\le&\sum_{n=1}^{N}\theta h \parallel \nabla(u_{\omega,h}^{n-1}-u_{\omega,h}^n)\parallel_{I_n}\parallel \varphi_\omega \parallel_{L^2(I_n;H^2(D))}\\
\le&\theta h\left(\sum_{n=1}^{N}\parallel \nabla(u_{\omega,h}^{n-1}-u_{\omega,h}^n)\parallel_{I_n}^2\right)^{\frac{1}{2}}\parallel \varphi_\omega \parallel_{L^2(0,T;H^2(D))}\\
=&\theta h\tau^{1/2}\left(\sum_{n=1}^{N}\parallel \nabla(u_{\omega,h}^{n-1}-u_{\omega,h}^n)\parallel^2\right)^{\frac{1}{2}}\parallel \varphi_\omega \parallel_{L^2(0,T;H^2(D))}\ .
\end{aligned}$$

By Green's formulation, we can get

$$\begin{aligned}
&\sum_{n=1}^{N}((a_\omega^n-\bar{a})\nabla(u_{\omega,h}^{n-1}-u_{\omega,h}^n),\nabla\varphi_\omega)_{I_n}\\
=&-\sum_{n=1}^{N}(u_{\omega,h}^{n-1}-u_{\omega,h}^n,\nabla(a_\omega^n-\bar{a})\cdot\nabla\varphi_\omega)_{I_n}-\sum_{n=1}^{N}(u_{\omega,h}^{n-1}-u_{\omega,h}^n,(a_\omega^n-\bar{a})\Delta\varphi_\omega)_{I_n}.
\end{aligned}$$

The Cauchy inequality, Theorem 3.2 and 3.5 imply

$$\begin{aligned}
\left|\sum_{n=1}^{N}((a_\omega^n-\bar{a})\nabla(u_{\omega,h}^{n-1}-u_{\omega,h}^n),\nabla\varphi_\omega)_{I_n}\right|\le&\tau\theta_1\left(\sum_{n=1}^{N}\frac{1}{\tau}\parallel u_{\omega,h}^n-u_{\omega,h}^{n-1}\parallel^2\right)^{\frac{1}{2}}\parallel\nabla\varphi_\omega\parallel_I\\
&+\tau\theta\left(\sum_{n=1}^{N}\frac{1}{\tau}\parallel u_{\omega,h}^n-u_{\omega,h}^{n-1}\parallel^2\right)^{\frac{1}{2}}\parallel\Delta\varphi_\omega\parallel_I\ .
\end{aligned}$$

Therefore, we can obtain

$$\begin{aligned}
\textcircled{4}\le&\theta h\tau^{\frac{1}{2}}\left(\sum_{n=1}^{N}\parallel\nabla(u_{\omega,h}^{n-1}-u_{\omega,h}^n)\parallel^2\right)^{\frac{1}{2}}\parallel\Delta\varphi_\omega\parallel_I+L\tau\left(\sum_{n=1}^{N}\tau\parallel\nabla u_{\omega,h}^n\parallel^2\right)^{\frac{1}{2}}\parallel\nabla\varphi_\omega\parallel_I\\
&+\tau\theta_1\left(\sum_{n=1}^{N}\frac{1}{\tau}\parallel u_{\omega,h}^n-u_{\omega,h}^{n-1}\parallel^2\right)^{\frac{1}{2}}\parallel\nabla\varphi_\omega\parallel_I+\tau\theta\left(\sum_{n=1}^{N}\frac{1}{\tau}\parallel u_{\omega,h}^n-u_{\omega,h}^{n-1}\parallel^2\right)^{\frac{1}{2}}\parallel\Delta\varphi_\omega\parallel_I\ .
\end{aligned}$$

Together with this, by inequality (A.18), Theorem 3.2, 3.1, 3.5, Sobolev's embedding theorem, and the Young's inequality, we have

$$\begin{aligned}
\textcircled{4}\le&C((\tau+\tau^{\frac{1}{2}}h)\parallel\Delta\varphi_\omega\parallel_I+\tau\parallel\nabla\varphi_\omega\parallel_I)\\
\le&C((\tau+\tau^{\frac{1}{2}}h)\parallel\Delta\varphi_\omega\parallel_I+\tau\parallel\varphi_\omega\parallel_{L^2(0,T;H^2(D))})\\
\le&C(\tau+h^2)\parallel f\parallel_I\ .
\end{aligned}\tag{A.25}$$

Combining Eq (A.21), inequalities (A.22)–(A.25) and (3.13) and setting $f=u_\omega(q_{\omega,\tau h})-u_{\omega,\tau h}$, we can obtain inequality (4.4).

A.5. The proof of Theorem 4.2

Proof. For a fixed ω, consider the dual problem

$$\begin{cases} \partial_t \psi_\omega - \nabla \cdot (a(\omega, t, x) \nabla \psi_\omega) = f(x), & t, x \in (0, T] \times D, \\ \psi_\omega = 0, & x \in \partial D, t \in [0, T], \\ \psi_\omega(0) = 0, & x \in D. \end{cases}$$

According to [21] for $f \in L^2(0, T; L^2(D))$ and Assumptions (1)–(3), we have $\psi_\omega \in L^2(0, T; H^2(D)) \cap H^1(0, T; L^2(D))$ and

$$\| \psi_\omega \|_{L^2(0,T;H^2(D))} + \| \partial_t \psi_\omega \|_I \leq C \| f \|_I \,. \tag{A.26}$$

Thus it is reasonable for us to use ψ_ω^n to denote $\psi_\omega(t^n)$ $(n = 0, 1, \cdots, N)$ because $\psi_\omega \in L^2(0, T; H^2(D)) \cap H^1(0, T; L^2(D)) \hookrightarrow C(\bar{I}, H^1(D))$. Furthermore, we have

$$\begin{aligned} (z_\omega(u_{\omega,\tau h}) - z_{\omega,\tau h}, f)_I &= (z_\omega(u_{\omega,\tau h}), \partial_t \psi_\omega - \nabla \cdot (a_\omega \nabla \psi_\omega))_I - (z_{\omega,\tau h}, \partial_t \psi_\omega - \nabla \cdot (a_\omega \nabla \psi_\omega))_I \\ &= (u_{\omega,\tau h} - u_{\omega,d}, \psi_\omega)_I - \sum_{n=1}^{N} (z_{\omega,\tau h}, \partial_t \psi_\omega)_{I_n} - (a_\omega \nabla z_{\omega,\tau h}, \nabla \psi_\omega)_I \\ &= (u_{\omega,\tau h} - u_{\omega,d}, \psi_\omega)_I + \sum_{n=1}^{N} (z_{\omega,h}^n - z_{\omega,h}^{n-1}, \psi_\omega^n) - (a_\omega \nabla z_{\omega,\tau h}, \nabla \psi_\omega)_I. \end{aligned} \tag{A.27}$$

Substituting $v_h = \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n$ into problem (3.10) yields

$$-\left(\frac{z_{\omega,h}^n - z_{\omega,h}^{n-1}}{\tau}, \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n \right) + (\bar{a} \nabla z_{\omega,h}^{n-1}, \nabla \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n) + ((a_\omega^n - \bar{a}) \nabla z_{\omega,h}^n, \nabla \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n) = (u_{\omega,h}^n - \widehat{u_{\omega,d}}^n, \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n).$$

Summing n from 1 to N, we get

$$\begin{aligned} &- \sum_{n=1}^{N} (z_{\omega,h}^n - z_{\omega,h}^{n-1}, \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n) + \sum_{n=1}^{N} (\bar{a} \nabla z_{\omega,h}^{n-1}, \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n)_{I_n} \\ &+ \sum_{n=1}^{N} ((a_\omega^n - \bar{a}) \nabla z_{\omega,h}^n, \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n)_{I_n} - \sum_{n=1}^{N} (u_{\omega,h}^n - \widehat{u_{\omega,d}}^n, \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n)_{I_n} = 0. \end{aligned} \tag{A.28}$$

We note that

$$\sum_{n=1}^{N} (u_{\omega,h}^n - \widehat{u_{\omega,d}}^n, \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n)_{I_n} = \sum_{n=1}^{N} (u_{\omega,\tau h} - u_{\omega,d}, \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n)_{I_n}.$$

Substituting Eq (A.28) into Eq (A.27) yields

$$\begin{aligned} (z_\omega(u_{\omega,\tau h}) - z_{\omega,\tau h}, f)_I = &\underbrace{\sum_{n=1}^{N} (u_{\omega,\tau h} - u_{\omega,d}, \psi_\omega - \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n)_{I_n}}_{\textcircled{1}} + \underbrace{\sum_{n=1}^{N} (z_{\omega,h}^n - z_{\omega,h}^{n-1}, \psi_\omega - \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n)}_{\textcircled{2}} \\ + &\underbrace{\sum_{n=1}^{N} (\bar{a} \nabla z_{\omega,h}^{n-1}, \nabla \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n)_{I_n} + \sum_{n=1}^{N} ((a_\omega^n - \bar{a}) \nabla z_{\omega,h}^n, \nabla \widehat{\mathcal{R}_{\omega,h}\psi_\omega}^n)_{I_n} - \sum_{n=1}^{N} (a_\omega \nabla z_{\omega,\tau h}, \nabla \psi_\omega)_{I_n}}_{\textcircled{3}}. \end{aligned} \tag{A.29}$$

$$\textcircled{1}=\sum_{n=1}^{N}(u_{\omega,\tau h}-u_{\omega,d},\psi_\omega-\widehat{\mathcal{R}_{\omega,h}\psi_\omega}^{\,n})_{I_n}=\sum_{n=1}^{N}(u_{\omega,\tau h}-u_{\omega,d},\psi_\omega-\widehat{\psi_\omega}^{\,n})_{I_n}+\sum_{n=1}^{N}(u_{\omega,\tau h}-u_{\omega,d},\widehat{\psi_\omega}^{\,n}-\widehat{\mathcal{R}_{\omega,h}\psi_\omega}^{\,n})_{I_n}.$$

Using the Hölder's inequality or the property of L^2 projection, we can obtain that

$$\| \widehat{\psi_\omega}^{\,n}-\widehat{\mathcal{R}_{\omega,h}\psi_\omega}^{\,n} \|_{I_n}\leq\| \psi_\omega-\mathcal{R}_{\omega,h}\psi_\omega \|_{I_n} .$$

Finally, we can deduce

$$\textcircled{1}\leq\sum_{n=1}^{N}\| u_{\omega,\tau h}-u_{\omega,d} \|_{I_n}\| \psi_\omega-\widehat{\psi_\omega}^{\,n} \|_{I_n}+\sum_{n=1}^{N}\| u_{\omega,\tau h}-u_{\omega,d} \|_{I_n}\| \psi_\omega-\mathcal{R}_{\omega,h}\psi_\omega \|_{I_n} .$$

On the basis of this representation, all steps of the proof of Theorem 4.1 can be repeated similarly to obtain the stated result.

A.6. The proof of Theorem 4.3

Proof. It follows from the continuous optimality conditions (2.11) and discrete optimality conditions (3.11) that

$$(z_\omega+\lambda q_\omega,\bar{q}-q_\omega)_I\geq 0,\qquad \forall\bar{q}\in Q_{ad}, \tag{A.30}$$

$$(z_{\omega,\tau h}+\lambda q_{\omega,\tau h},\bar{q}-q_{\omega,\tau h})_I\geq 0,\qquad \forall\bar{q}\in Q_{ad}. \tag{A.31}$$

Choosing $\bar{q}=q_{\omega,\tau h}$ in inequality (A.30), $\bar{q}=q_\omega$ in inequality (A.31), we have

$$\begin{aligned}&(z_\omega+\lambda q_\omega,q_{\omega,\tau h}-q_\omega)_I\geq 0,\\&(z_{\omega,\tau h}+\lambda q_{\omega,\tau h},q_\omega-q_{\omega,\tau h})_I\geq 0.\end{aligned}$$

Adding the two inequalities yields

$$\begin{aligned}\lambda\| q_\omega-q_{\omega,\tau h} \|_I^2\leq&(z_{\omega,\tau h}-z_\omega,q_\omega-q_{\omega,\tau h})_I\\\leq&(z_{\omega,\tau h}-z_\omega(u_\omega(q_{\omega,\tau h}))+z_\omega(u_\omega(q_{\omega,\tau h}))-z_\omega,q_\omega-q_{\omega,\tau h})_I\\=&(z_{\omega,\tau h}-z_\omega(u_\omega(q_{\omega,\tau h})),q_\omega-q_{\omega,\tau h}))_I+(z_\omega(u_\omega(q_{\omega,\tau h}))-z_\omega,q_\omega-q_{\omega,\tau h})_I\\=&(z_{\omega,\tau h}-z_\omega(u_\omega(q_{\omega,\tau h})),q_\omega-q_{\omega,\tau h}))_I+(u_\omega(q_{\omega,\tau h})-u_\omega,u_\omega-u_\omega(q_{\omega,\tau h}))_I\\\leq&(z_{\omega,\tau h}-z_\omega(u_\omega(q_{\omega,\tau h})),q_\omega-q_{\omega,\tau h}))_I\\=&(z_{\omega,\tau h}-z_\omega(u_{\omega,\tau h}),q_\omega-q_{\omega,\tau h})_I+(z_\omega(u_{\omega,\tau h})-z_\omega(u_\omega(q_{\omega,\tau h})),q_\omega-q_{\omega,\tau h})_I.\end{aligned} \tag{A.32}$$

Following problems (4.2) and (4.3), we have

$$-(\partial_t(z_\omega(u_{\omega,\tau h})-z_\omega(u_\omega(q_{\omega,\tau h}))),v)_I+(a_\omega\nabla(z_\omega(u_{\omega,\tau h})-z_\omega(u_\omega(q_{\omega,\tau h}))),\nabla v)_I=(u_{\omega,\tau h}-u_\omega(q_{\omega,\tau h}),v)_I.$$

inequality (A.18) implies

$$\| z_\omega(u_{\omega,\tau h})-z_\omega(u_\omega(q_{\omega,\tau h})) \|_I\leq C\| u_{\omega,\tau h}-u_\omega(q_{\omega,\tau h}) \|_I . \tag{A.33}$$

Using inequalities (A.32) and (A.33), Theorem 4.1 and 4.2, we can obtain

$$\begin{aligned}\lambda\| q_\omega-q_{\omega,\tau h} \|_I\leq&\| z_{\omega,\tau h}-z_\omega(u_{\omega,\tau h}) \|_I+\| u_{\omega,\tau h}-u_\omega(q_{\omega,\tau h}) \|_I\\\leq&C_\omega(\tau+h^2).\end{aligned} \tag{A.34}$$

Regarding $\| u_\omega - u_{\omega,\tau h} \|_I$, by the triangle inequality, we have

$$\| u_\omega - u_{\omega,\tau h} \|_I \leq \| u_\omega - u_\omega(q_{\omega,\tau h}) \|_I + \| u_\omega(q_{\omega,\tau h}) - u_{\omega,\tau h} \|_I \ . \tag{A.35}$$

According to problems (2.9) and (4.1), we have

$$(\partial_t(u_\omega - u_\omega(q_{\omega,\tau h})), v)_I + (a_\omega \nabla(u_\omega - u_\omega(q_{\omega,\tau h})), \nabla v)_I = (q_\omega - q_{\omega,\tau h}, v)_I.$$

Using Theorem 2.1, we can get

$$\| u_\omega - u_\omega(q_{\omega,\tau h}) \|_I \leq C_\omega \| q_\omega - q_{\omega,\tau h} \|_I \ . \tag{A.36}$$

Using inequalityies (A.34)–(A.36) and Theorem 4.1, we can derive

$$\begin{aligned} \| u_\omega - u_{\omega,\tau h} \|_I &\leq \| q_\omega - q_{\omega,\tau h} \|_I + \| u_\omega(q_{\omega,\tau h}) - u_{\omega,\tau h} \|_I, \\ &\leq C_\omega(\tau + h^2). \end{aligned} \tag{A.37}$$

Regarding $\| z_\omega - z_{\omega,\tau h} \|_I$, by the triangle inequality, we have

$$\| z_\omega - z_{\omega,\tau h} \|_I \leq \| z_\omega - z_\omega(u_{\omega,\tau h}) \|_I + \| z_\omega(u_{\omega,\tau h}) - z_{\omega,\tau h} \|_I \ . \tag{A.38}$$

Following problems (2.10) and (4.2), we have

$$-(\partial_t(z_\omega - z_\omega(u_{\omega,\tau h})), v)_I + (a_\omega \nabla(z_\omega - z_\omega(u_{\omega,\tau h})), \nabla v)_I = (u_\omega - u_{\omega,\tau h}, v)_I.$$

This, together with inequality (A.18), implies

$$\| z_\omega - z_\omega(u_{\omega,\tau h}) \|_I \leq C_\omega \| u_\omega - u_{\omega,\tau h} \|_I \ . \tag{A.39}$$

Using inequalities (A.37)–(A.39) and Theorem 4.2, we can obtain

$$\begin{aligned} \| z_\omega - z_{\omega,\tau h} \|_I &\leq \| u_\omega - u_{\omega,\tau h} \|_I + \| z_\omega(u_{\omega,\tau h}) - z_{\omega,\tau h} \|_I \\ &\leq C_\omega(\tau + h^2). \end{aligned}$$

This, together with inequalities (A.34) and (A.37), yield inequality (4.5). It is clear that inequality (4.6) follows from inequality (4.5).

AIMS Press

© 2025 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)