Applied Probability Trust (10 August 2018)

A MULTILEVEL APPROACH TOWARDS UNBIASED SAMPLING OF RANDOM ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS

XIAOOU LI,* *University of Minnesota*

JINGCHEN LIU,** *Columbia University*

SHUN XU,*** *Columbia University*

Abstract

Partial differential equation is a powerful tool to characterize various physics systems. In practice, measurement errors are often present and probability models are employed to account for such uncertainties. In this paper, we present a Monte Carlo scheme that yields unbiased estimators for expectations of random elliptic partial differential equations. This algorithm combines multilevel Monte Carlo [8] and a randomization scheme proposed by [12, 13]. Furthermore, to obtain an estimator with both finite variance and finite expected computational cost, we employ higher order approximations.

Keywords: Unbiased sampling, PDE with random coefficients, Monte Carlo methods

2010 Mathematics Subject Classification: Primary 65C05

Secondary 35R60;82B80

1. Introduction

Elliptic partial differential equation is a classic equation that is employed to describe various static physics systems. In practical life, such systems are usually not described precisely. For instance, imprecision could be due to microscopic heterogeneity or

* Postal address: School of Statistics, 224 Church Street SE, Minneapolis, MN 55455

** Postal address: Department of Statistics, 1255 Amsterdam Avenue, New York, New York, 10027, USA

*** Postal address: Department of Statistics, 1255 Amsterdam Avenue, New York, New York, 10027, USA

measurement errors of parameters. To account for this, we introduce uncertainty to the system by letting certain coefficients contain randomness. To be precise, let $U \subset R^d$ be a simply connected domain. We consider the following differential equation concerning $u : U \to R$

$$-\nabla \cdot (a(x)\nabla u(x)) = f(x) \text{ for } x \in U, \tag{1}$$

where $f(x)$ is a real-valued function and $a(x)$ is a strictly positive function. Just to clarify the notation, $\nabla u(x)$ is the gradient of $u(x)$ and "$\nabla\cdot$" is the divergence of a vector field. For each a and f, one solves u subject to certain boundary conditions that are necessary for the uniqueness of the solution. This will be discussed in the sequel. The randomness is introduced to the system through $a(x)$ and $f(x)$. Thus, the solution u as an implicit functional of a and f is a real-valued stochastic process living on U. More precisely, consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The functions a, f, and u are maps from $U \times \Omega$ to R, where the function a is in fact almost surely strictly positive. In the rest of this paper, we omit the second argument in $a(x, \omega)$, $f(x, \omega)$, and $u(x, \omega)$ and write $a(x)$, $f(x)$, and $u(x)$ instead that satisfy equation (1) and boundary conditions almost surely. Throughout this paper, we consider $d \leq 3$ that is sufficient for most physics applications.

Of interest is the distributional characteristics of $\{u(x) : x \in U\}$. The solution is typically not in an analytic form of a and f and thus closed form characterizations are often infeasible. In this paper, we study the distribution of u via Monte Carlo. Let $C(\bar{U})$ be the set of continuous functions on $\bar{U}$. For a real-valued functional

$$\mathcal{Q} : C(\bar{U}) \to R$$

satisfying certain regularity conditions, we are interested in computing

$$w_{\mathcal{Q}} = \mathbb{E}\{\mathcal{Q}(u)\} = \int \mathcal{Q}(u(\cdot, \omega))\mathbb{P}(d\omega).$$

The expectation in the above display is taken with respect to the uncertainty in the random fields $a(x)$ and $f(x)$. Such problems appear often in the studies of physics systems; see, for instance, [5, 6].

The contribution of this paper is the development of an unbiased Monte Carlo estimator of $w_{\mathcal{Q}}$ with finite variance. Furthermore, the expected computational cost

of generating one such estimator is finite. The analysis strategy is a combination of multilevel Monte Carlo and a randomization scheme. Multilevel Monte Carlo is a recent advance in simulation and approximation of continuous processes [8, 4, 9]. The randomization scheme is developed by [12, 13]. Under the current setting, a direct application of these two methods leads to either an estimator with infinite variance or infinite expected computational cost. This is mostly due to the fact that the accuracy of regular numerical methods of the partial differential equations is insufficient. More precisely, the mean squared error of a discretized Monte Carlo estimator is proportional to the square of mesh size [2, 15]. The technical contribution of this paper is to employ the finite element method with quadratic isoparametric elements to solve PDE under certain smoothness conditions of $a(x)$ and $f(x)$ and to perform careful analysis of the numerical solver for equation (1).

Physics applications. Equation (1) has been widely used in many disciplines to describe time-independent physical problems. The well-known Poisson equation or Laplace equation is a special case when $a(x)$ is a constant. In different disciplines, the solution $u(x)$ and the coefficients $a(x)$ and $f(x)$ have their specific physics meanings. When the elliptic PDE is used to describe the steady-state distribution of heat (as temperature), $u(x)$ carries the meaning of temperature at x and the coefficient $a(x)$ is the heat conductivity. In the study of electrostatics, u is the potential (or voltage) induced by electronic charges, ∇u is the electric field, and $a(x)$ is the permittivity (or resistance) of the medium. In groundwater hydraulics, the meaning of $u(x)$ is the hydraulic head (water level elevation) and $a(x)$ is the hydraulic conductivity (or permeability). The physics laws for the above three different problems to derive the same type of elliptic PDE are called Fourier's law, Gauss's law, and Darcy's law, respectively. In classical continuum mechanics, equation (1) is known as the generalized Hook's law where u describes the material deformation under the external force f. The coefficient $a(x)$ is known as the elasticity tensor.

In this paper, we consider that both $a(x)$ and $f(x)$ possibly contain randomness. We elaborate its physics interpretation in the context of material deformation application. In the model of classical continuum mechanics, the domain U is a smooth manifold denoting the physical location of the piece of material. The displacement $u(x)$ depends

on the external force $f(x)$, boundary conditions, and the elasticity tensor $\{a(x) : x \in U\}$. The elasticity coefficient $a(x)$ is modeled as a spatially varying random field to characterize the inherent heterogeneity and uncertainties in the physical properties of the material (such as the modulus of elasticity, c.f. [14, 11]). For example, metals, which lend themselves most readily to the analysis by means of the classical elasticity theory, are actually polycrystals, i.e., aggregates of an immense number of anisotropic crystals randomly oriented in space. Soils, rocks, concretes, and ceramics provide further examples of materials with very complicated structures. Thus, incorporating randomness in $a(x)$ is necessary to take into account of the heterogeneities and the uncertainties under many situations. Furthermore, there may also be uncertainty contained in the external force $f(x)$.

The rest of the paper is organized as follows. In Section 2, we present the problem settings and some preliminary materials for the main results. Section 3 presents the construction of the unbiased Monte Carlo estimator for $w_{\mathcal{Q}}$ and rigorous complexity analysis. Numerical implementations are included in Section 4. Technical proofs and detailed definition of finite element methods are included in the appendix.

2. Preliminary analysis

Throughout this paper, we consider equation (1) living on a bounded domain $U \subset R^d$ with twice differentiable boundary denoted by ∂U. To ensure the uniqueness of the solution, we consider the Dirichlet boundary condition

$$u(x) = 0, \quad \text{for } x \in \partial U. \tag{2}$$

We let both exogenous functions $f(x)$ and $a(x)$ be random processes, that is,

$$f(x, \omega) : U \times \Omega \to R \quad \text{and} \quad a(x, \omega) : U \times \Omega \to R \tag{3}$$

where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. To simplify notation, we omit the second argument and write $a(x)$ and $f(x)$. As an implicit function of the input processes $a(x)$ and $f(x)$, the solution $u(x)$ is also a stochastic process living on U. We are interested in computing the distribution of $u(x)$ via Monte Carlo. In particular, for some functional $\mathcal{Q} : C(\bar{U}) \to R$ satisfying certain regularity conditions that will be

specified in the sequel, we compute the expectation

$$w_{\mathcal{Q}} = \mathbb{E}[\mathcal{Q}(u)] \tag{4}$$

by Monte Carlo. The notation $\bar{U}$ is the closure of domain U and $C(\bar{U})$ is the set of real-valued continuous functions on $\bar{U}$.

Let $\hat{Z}$ be an estimator (possibly biased) of $\mathbb{E}\mathcal{Q}(u)$. The mean square error (MSE)

$$\mathbb{E}(\hat{Z} - w_{\mathcal{Q}})^2 = Var(\hat{Z}) + \{\mathbb{E}(\hat{Z}) - w_{\mathcal{Q}}\}^2. \tag{5}$$

consists of a bias term and a variance term. For the Monte Carlo estimator in this paper, the bias is removed via a randomization scheme combined with multilevel Monte Carlo. To start with, we present the basics of multilevel Monte Carlo and the randomization scheme.

2.1. Multilevel Monte Carlo

Consider a biased estimator of $w_{\mathcal{Q}}$ denote by Z_n. In the current context, Z_n is the estimator corresponding to some numerical solution based on certain discretization scheme, for instance, $Z_n = \mathcal{Q}(u_n)$ where u_n is the solution of the finite element method. The subscript n is a generic index of the discretization size. The detailed construction of Z_n will be provided in the sequel. As $n \to \infty$, the estimator becomes unbiased, that is, $\mathbb{E}(Z_n) \to w_{\mathcal{Q}}$. Multilevel Monte Carlo is based on the following telescope sum

$$w_{\mathcal{Q}} = \mathbb{E}(Z_0) + \sum_{i=0}^{\infty} \mathbb{E}(Z_{i+1} - Z_i). \tag{6}$$

One may choose Z_0 to be some simple constant. Without loss of generality, we choose $Z_0 \equiv 0$ and thus the first term vanishes. The advantage of writing $w_{\mathcal{Q}}$ as the telescope sum is that one is often able to construct Z_i and Z_{i+1} carefully such that they are appropriately coupled and the variance of $Y_i = Z_{i+1} - Z_i$ decreases fast as i tends infinity. The coupling is commonly done by constructing Z_{i+1} and Z_i with the same sample path ω (that is, same $a(\cdot, \omega)$ and $f(\cdot, \omega)$). The specific choice of our Y_i and Z_i in this paper is given on page 14. Let

$$\Delta_i = \mathbb{E}(Z_{i+1} - Z_i) \tag{7}$$

be estimated by

$$\hat{\Delta}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_i^{(j)}$$

where $Y_i^{(j)}$, $j = 1, ..., n_i$ are independent replicates of Y_i. The multilevel Monte Carlo estimator is

$$\hat{Z} = \sum_{i=1}^{I} \hat{\Delta}_i \tag{8}$$

where I is a large integer truncating the infinite sum (6).

2.2. An unbiased estimator via a randomization scheme

In the construction of the multilevel Monte Carlo estimator (8), the truncation level I is always finite and therefore the estimator is always biased. In what follows, we present an estimator with the bias removed. It is constructed based on the telescope sum of the multilevel Monte Carlo estimator and a randomization scheme that is originally proposed by [12, 13].

Let N be a positive-integer-valued random variable that is independent of $\{Z_i\}_{i=1,2,...}$. Let $p_n = \mathbb{P}(N = n)$ be the probability mass function of N such that $p_n > 0$ for all $n > 0$. The following identity holds trivially

$$w_{\mathcal{Q}} = \sum_{i=1}^{\infty} \mathbb{E}(Z_n - Z_{n-1}) = \sum_{n=1}^{\infty} \frac{\mathbb{E}[Z_n - Z_{n-1}; N = n]}{p_n} = \mathbb{E}\Big(\frac{Z_N - Z_{N-1}}{p_N}\Big),$$

where we use the notation $\mathbb{E}[X; B] = \mathbb{E}[X\mathbf{1}_B]$ with X, $\mathbf{1}_B$ being a random variable and the indicator function of an event B, respectively. Therefore, an unbiased estimator of $w_{\mathcal{Q}}$ is given by

$$\tilde{Z} = \frac{Z_N - Z_{N-1}}{p_N}. \tag{9}$$

Let $\tilde{Z}_i, i = 1, ..., M$ be independent copies of $\tilde{Z}$. The averaged estimator

$$\tilde{Z}_M = \frac{1}{M} \sum_{i=1}^{M} \tilde{Z}_i$$

is unbiased for $w_{\mathcal{Q}}$ with variance $Var(\tilde{Z})/M$ if finite.

We provide a complexity analysis of the estimator $\tilde{Z}$. This consists of the calculation of the variance of $\tilde{Z}$ and of the computational cost to generate $\tilde{Z}$. We start with the second moment

$$\mathbb{E}(\tilde{Z}^2) = \mathbb{E}\Big[\frac{(Z_N - Z_{N-1})^2}{p_N^2}\Big] = \sum_{n=1}^{\infty} \frac{\mathbb{E}(Z_n - Z_{n-1})^2}{p_n}. \tag{10}$$

In order to have a finite second moment, the sequence $\frac{\mathbb{E}(Z_n - Z_{n-1})^2}{p_n}$ needs to tend to zero no slower than n^{-1}. Thus, we would like to choose the random variable N such

that

$$p_n > n\mathbb{E}(Z_n - Z_{n-1})^2 \text{ for all } n \text{ sufficiently large.} \tag{11}$$

Furthermore, p_n must also satisfy the natural constraint that

$$\sum_{n=1}^{\infty} p_n = 1,$$

which suggests $p_n < n^{-1}$ for sufficiently large n. Combining with (11), we have

$$n^{-1} > p_n > n\mathbb{E}(Z_n - Z_{n-1})^2 \tag{12}$$

Notice that we have not yet specified a discretization method, thus (12) can typically be met by appropriately indexing the mesh size. For instance, in the context of solving PDE numerically, one may choose the mesh size converging to 0 at a super exponential rate with n (such as e^{-n^2}) and thus $\mathbb{E}(Z_n - Z_{n-1})^2$ decreases sufficiently fast that allows quite some flexibility in choosing p_n. Thus, constraint (12) alone can always be satisfied and it is not intrinsic to the problem. It is the combination with the following constraint that forms the key issue.

We now compute the expected computational cost for generating $\tilde{Z}$. Let c_n be the computational cost for generating $Z_n - Z_{n-1}$. Then, the expected cost is

$$C = \sum_{i=1}^{n} p_n c_n. \tag{13}$$

In order to have C finite, for n sufficiently large,

$$p_n < n^{-1} c_n^{-1}. \tag{14}$$

Based on the above calculation, if the estimator $\tilde{Z}$ has a finite variance and a finite expected computation time, then p_n must satisfy both (12) and (14), which suggests

$$\mathbb{E}(Z_n - Z_{n-1})^2 < n^{-2} c_n^{-1}. \tag{15}$$

That is, one must be able to construct a coupling between Z_n and Z_{n-1} such that (15) is in place. In Section 3, we provide detailed complexity analysis for the random elliptic PDE illustrating the challenges and presenting the solution.

2.3. Function spaces and norms

In this section, we present a list of notation that will be frequently used in later discussion. Let $U \subset R^d$ be a bounded open set. We define the following spaces of functions.

$$C^k(\bar{U}) \quad = \quad \{u : \bar{U} \to R | u \text{ is } k\text{-time continuously differentiable over } \bar{U} \ \}$$

That is, $u \in C^k(\bar{U})$ means that all the k-th partial derivatives of u are continuous over $\bar{U}$.

$$\begin{aligned} L^p(U) &= \{u : U \to R | \int_U |u(x)|^p dx < \infty\} \\ L^p_{loc}(U) &= \{u : U \to R | u \in L^p(K) \text{ for any compact subset } K \subset U\} \\ C^\infty_c(U) &= \{u : U \to R | u \text{ is infinitely differentiable with a compact support that is a subset of } U\}. \end{aligned}$$

Definition 1. For $u, w \in L^1_{loc}(U)$ and a multiple index α, we say w is the α-weak derivative of u, and write $D^\alpha u = w$ if $\int_U u D^\alpha \phi dx = (-1)^{|\alpha|} \int_U w\phi dx$ for all $\phi \in C^\infty_c(U)$, where $D^\alpha \phi$ in the above expression denote the usual α-partial derivative of ϕ.

If $u \in C^k(\bar{U})$ and $|\alpha| \leq k$, then the α-weak derivative and the usual partial derivative are the same. Therefore, we can write $D^\alpha \phi$ for both continuously differentiable and weakly differentiable functions without ambiguous.

We further define norms $\| \cdot \|_{C^k(\bar{U})}$ and $\| \cdot \|_{L^p(U)}$ on $C^k(\bar{U})$ and $L^p(U)$ respectively as follows.

$$\|u\|_{C^k(\bar{U})} = \sup_{|\alpha| \leq k, x \in \bar{U}} |D^\alpha u(x)|, \tag{16}$$

and

$$\|u\|_{L^p(U)} = \Big(\int_U |u|^p dx \Big)^{1/p}. \tag{17}$$

We proceed to the definition of Sobolev space $H^k(U)$ and $H^k_{loc}(U)$

$$H^k(U) = \{u : U \to R | D^\alpha u \in L^2(U) \text{ for all multiple index } \alpha \text{ such that } |\alpha| \leq k\}, \tag{18}$$

and $H^k_{loc}(U) = \{u : U \to R | \ u|_V \in H^k(V) \text{ for all } V \subsetneq U\}$. For $u \in H^k(U)$, the norm $\|u\|_{H^k(U)}$ and semi-norm $|u|_{H^k(U)}$ are defined as

$$\|u\|_{H^k(U)} = \Big(\sum_{|\alpha| \leq k} \|D^\alpha u\|^2_{L^2(U)} \Big)^{1/2}, \tag{19}$$

and

$$|u|_{H^k(U)} = \Big(\sum_{|\alpha|=k} \|D^\alpha u\|^2_{L^2(U)} \Big)^{1/2}. \tag{20}$$

We define the space $H_0^1(U)$ as

$$H_0^1(U) = \{u \in H^1(U) : u(x) = 0 \text{ for } x \in \partial U\}. \tag{21}$$

On the space $H_0^1(U)$ the norm $\|\cdot\|_{H^1(U)}$ and the semi-norm $|\cdot|_{H^1(U)}$ are equivalent.

2.4. Finite element method for partial differential equation

We briefly describe the finite element method for partial differential equations. The weak solution $u \in H_0^1(U)$ to (1) under the Dirichlet boundary condition (2) is defined through the following variational form

$$b(u, v) = L(v) \text{ for all } v \in H_0^1(U), \tag{22}$$

where we define the bilinear and linear forms

$$b(u, v) = \int_U a(x)\nabla u(x) \cdot \nabla v(x) dx \quad \text{and} \quad L(v) = \int_U f(x)v(x)dx,$$

and "$\cdot$" is the vector inner product. When the coefficients a and f are sufficiently smooth, say, infinitely differentiable, the weak solution u becomes a strong solution. That is, u is the solution of (1). The key step of the finite element method is to approximate the infinite dimensional space $H_0^1(U)$ by some finite dimensional linear space $V_n = \text{span}\{\phi_1, ..., \phi_{L_n}\}$, where L_n is the dimension of V_n. The approximate solution $u_n \in V_n$ is defined through the set of equations

$$b(u_n, v) = L(v) \text{ for all } v \in V_n. \tag{23}$$

Both sides of the above equations are linear in v. Then, (23) is equivalent to $b(u_n, \phi_i) = L(\phi_i)$ for $i = 1, ..., L_n$. We further write $u_n = \sum_{i=1}^{L_n} d_i\phi_i$ as a linear combination of the basis functions. Then, (23) is equivalent to solving linear equations

$$\sum_{j=1}^{L_n} d_j b(\phi_j, \phi_i) = L(\phi_i) \text{ for } i = 1, ..., L_n. \tag{24}$$

The basis functions $\phi_1, ..., \phi_{L_n}$ are often chosen such that (24) is a sparse linear system. That is, the order of the number of non-zero $b(\phi_j, \phi_i)$'s is $O(L_n)$. Such a sparse linear

system can be solved using an iterative method with a computational cost of the order $O(L_n \log(L_n))$ as $L_n \to \infty$. See Chapter 5 of [10] for more details.

Several choices of V_n have been studied for elliptic PDEs. For example, V_n may consists of all the piecewise linear functions over a triangularization of U. Such a linear element method has been adopted in [2] and [4] to construct multilevel Monte Carlo estimators for random elliptic PDEs.

In this paper, our choice V_n is a function space induced by quadratic isoparametric elements, which is suitable when U has a smooth boundary. The intuitive explanation of isoparametric elements is given in page 11, and the precise definition will be delayed to the Appendix C. The advantage of using quadratic isoparametric elements over the linear elements is twofold. First, the quadratic approximation provides better convergence rate when the solution has a higher order regularity ($\|u\|_{H^3(D)} < \infty$). Second, isoparametric triangularization provides a better approximation for the boundary ∂U, yielding a better approximation of the solution. For more details of finite element methods for elliptic PDEs, we refer the readers to the book [3] and the references therein.

3. Main results

In this section, we present the construction of $\tilde{Z}$ and its complexity analysis. We use finite element method to solve the PDE numerically and then construct Z_n. To illustrate the challenge, we start with the complexity analysis of $\tilde{Z}$ based on usual finite element method with linear basis functions, with which we show that (12) and (14) cannot be satisfied simultaneously. Thus, $\tilde{Z}$ either has infinite variance or has infinite expected computational cost. We improve upon this by means of quadratic approximation under smoothness assumptions on a and f. The estimator $\tilde{Z}$ thus can be generated in constant time and has a finite variance.

3.1. Finite element method

Piecewise linear basis functions. A popular choice of V_n is the space of piecewise linear functions defined on a triangularization $\mathcal{T}_n$ of U. In particular, $\mathcal{T}_n$ is a partition of U that is each element of $\mathcal{T}_n$ is a triangle partitioning U. The maximum edge length of

triangles is proportional to 2^{-n} and V_n is the space of all the piecewise linear functions over $\mathcal{T}_n$ that vanish on the boundary ∂U. The dimension of V_n is $L_n = O(2^{dn})$. Detailed construction of $\mathcal{T}_n$ and piecewise linear basis functions is provided in Appendix C.

Once a set of basis functions has been chosen, the coefficients d_i's are solved according to the linear equations (24) and the numerical solution is given by $u_n(x) = \sum_{i=1}^{L_n} d_i \phi_i(x)$. For each functional $\mathcal{Q}$, the biased estimator is $Z_n = \mathcal{Q}(u_n)$. It is important to notice that, for different n, u_n are computed based on the same realizations of a and f. Thus, Z_n and Z_{n-1} are coupled.

We now proceed to verifying (15) for linear basis functions. The dimension of V_n is of order $L_n = O(2^{dn})$ where $d = dim(U)$. We consider the case when $\mathcal{Q}$ is a functional that involves weak derivatives of u. For instance, $\mathcal{Q}$ could be in the form $q(|\cdot|_{H^1(U)})$ for some smooth function q and $Z = \mathcal{Q}(u)$, where $|\cdot|_{H^1(U)}$ is defined as in (20).

According to Proposition 4.2 of [2], under the conditions that $\mathbb{E}[\frac{1}{\min_{x\in U} a^p(x)}] < \infty$, $\mathbb{E}(\|a\|^p_{C^1(\bar{U})}) < \infty$, and $\mathbb{E}(\|f\|^p_{L^2(U)}) < \infty$ for all $p > 0$, $\mathbb{E}(Z_n - Z_{n-1})^2 = O(2^{-2n})$ if u_n and u_{n-1} are computed using the same sample of a and f. The condition (15) becomes $n2^{-2(n-1)} < n^{-1}2^{-dn}|\log 2^{-nd}|^{-1}$. A simple calculation yields that the above inequality holds only if $d = 1$. Therefore, *it is impossible to pick p_n such that the estimator $\tilde{Z}$ has a finite variance and a finite expected computational cost using the finite element method with linear basis functions if $d \geq 2$.* The one-dimensional case is not of great interest given that u can be solved explicitly. To establish (15) for higher dimensions, we need a faster convergence rate of the PDE numerical solver.

Quadratic isoparametric elements. We improve the accuracy of the finite element method by means of quadratic isoparametric elements, whose precise definition is given in Appendix C, under smoothness conditions on $a(x)$ and $f(x)$. Classic results (e.g. [3, Chapter VI]) show that if the solution u of the PDE is smooth enough and U has a smooth boundary ∂U, then the accuracy of the finite element method can be improved by means of isoparametric elements. We obtain similar results for random coefficients.

In this paper, we let V_n be defined as in (78) (Appendix C) with a mesh size $O(2^{-n})$. We explain the space V_n intuitively. In general, the construction of V_n consists of two steps: 1) partition the space U into small and curved triangles. We will refer this partition as $\mathcal{T}_n$, whose precise definition is given in Appendix C. 2) For each $T \in \mathcal{T}_n$,

we need to define a linear space of functions over T, denoted by P_T. Then, we put the spaces P_T for $T \in \mathcal{T}_n$ together and define $V_n = \{v \in C(\bar{U}) : v|_{\partial U} = 0 \text{ and } v|_T \in P_T \text{ for } T \in \mathcal{T}_n\}$. Step 1) is usually done by certain mesh generating algorithm and step 2) is done through isoparametric mapping of a reference element. We provide a graphical illustration of the construction in the next example.

Example 1. Let $U = B(0,1) = \{(x,y) : x^2 + y^2 < 1\}$. For simplicity, we restrict our illustration to a subset $U' = B(0,1) \cap R^2_+$. The analysis on $U \setminus U'$ can be done similarly. The left panel of Figure 1 shows a possible choice of the partition when $n = 1$. The right panel of Figure 1 shows a refinement of the partition when $n = 2$. In this example, if $T \in \mathcal{T}_n$ does not have an edge (possibly curved) lying on the boundary of U (e.g. black region in Figure 1), then T is a triangle; if an edge of T lies on the boundary (e.g. gray region in Figure 1), then T has a curved boundary along ∂U. We can see that if we only allow a partition using straight triangles, it is not possible to have U exactly covered due to its curved boundary.

Now we explain how to define a linear space on each $T \in \mathcal{T}_n$. Typically, this is done by the so-called isoparametric mapping from a reference element. The procedure is as follows. First, we take the simplex $\hat{T} = \{(x,y) : x, y \geq 0, x+y \leq 1\}$ to be the reference element, and define the space $\hat{P}$ to be the space containing all quadratic functions over $\hat{T}$.

Then, for each $T \in \mathcal{T}_n$ shown in Figure 1, there is an invertible quadratic function $F_T : \hat{T} \to R^2$ such that $T = F_T(\hat{T})$. Now, we define a linear space P_T over T as $P_T = \{v : T \to \mathbb{R} : v(x) = \hat{v}(F_K^{-1}(x)) \text{ for some } \hat{v} \in \hat{P}\}$. Of note, when T is a triangle, the linear space P_T contains all quadratic functions over T; when T is curved, then P_T is induced by, but not necessarily be, the space of quadratic functions.

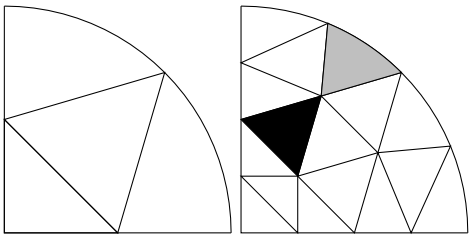

FIGURE 1: Isoparametric triangularization for Example 1. Left: $n = 1$; Right: $n = 2$.

With the finite dimensional space V_n constructed, we obtain an approximate solution u_n by solving (23) with V_n.

Isoparametric Numerical Integration. The numerical solution u_n in (23) requires the evaluation of the integrals $b(w,v) = \sum_{T\in\mathcal{T}_n}\int_T a(x)\nabla w(x)\cdot\nabla v(x)dx$ and $L(v) = \sum_{T\in\mathcal{T}_n}\int_T f(x)v(x)dx$. For the evaluation of these integrals we apply quadrature approximation, which approximates the integral in the form of $\int_T \phi(x)dx$ by $\sum_{l=1}^M w_{l,T}\phi(b_{l,T})$ for some pre-specified weights $w_{l,T}$ and points $b_{l,T}$ for a positive integer M and $1\le l\le M$. The precise choice of $w_{l,T}$ and $b_{l,T}$ are given in Appendix C. We point out that the choice of $w_{l,T}$ and $b_{l,T}$ depends on the isoparametric triangularization only, and is independent with the integrand $\phi(\cdot)$. By setting the function ϕ to be $a(x)\nabla w(x)\cdot\nabla v(x)$, and $f(x)v(x)$, we approximate the bilinear form $b(w,v)$ and linear form $L(v)$ with their numerical approximation, denoted by $\tilde{b}(w,v)$ and $\tilde{L}(v)$, respectively. Based on the numerical integration, we define $\tilde{u}_n$ such that

$$\tilde{b}_n(\tilde{u}_n, v) = \tilde{L}(v), \text{ for all } v\in V_n. \tag{25}$$

Error analysis for isoparametric finite element method. In what follows, we present an upper bound of the convergence rate of $\|\tilde{u}_n - u\|_{H^1(U)}$, where u is the solution to (22) and $\tilde{u}_n$ is the solution to (25). Define the minimum and maximum of $a(x)$ as $a_{\min} = \min_{x\in\bar{U}} a(x)$ and $a_{\max} = \max_{x\in\bar{U}} a(x)$. We make the following assumptions on the random coefficients $a(x)$ and $f(x)$.

A1. $a_{\min} > 0$ almost surely and $\mathbb{E}(1/a_{\min}^p) < \infty$, for all $p\in(0,\infty)$.

A2. a is almost surely continuously twice differentiable and $\mathbb{E}(\|a\|^p_{C^2(\bar{U})}) < \infty$ for all $p\in(0,\infty)$.

A3. $f\in H^2(U)$ almost surely and $\mathbb{E}(\|f\|^p_{H^2(U)}) < \infty$ for all $p\in(0,\infty)$.

A4. There exist non-negative constants p' and κ_q such that for all $w_1, w_2\in H_0^1(U)$,

$$|\mathcal{Q}(w_1) - \mathcal{Q}(w_2)| \le \kappa_q \max\{\|w_1\|^{p'}_{H^1(U)}, \|w_2\|^{p'}_{H^1(U)}\}\|w_1 - w_2\|_{H^1(U)}.$$

With the assumptions A1-A4, we are able to construct an unbiased estimator for $w_{\mathcal{Q}} = \mathbb{E}[\mathcal{Q}(u)]$ with both finite variance and finite expected computational time.

We start with the existence and the uniqueness of the solution. Notice that $a(x)$ is bounded below by positive random variables $a_{\min}$ and above by $a_{\max}$. According to [2], Lemma 2.1., (22) has a unique solution $u \in H_0^1(U)$ almost surely satisfying

$$\|u\|_{H^1(U)} \leq \kappa \frac{\|f\|_{L^2(U)}}{a_{\min}}. \tag{26}$$

The next theorem establishes the convergence rate of the approximate solution $\tilde{u}_n$ to the exact solution u.

Theorem 1. *Let $\tilde{u}_n$ be the solution to* (25)*. For* $\dim(U) \leq 3$ *with a* 3*-time differentiable boundary ∂U, if $a(x) \in C^2(\bar{U})$ and $f(x) \in H^2(U)$, then we have*

$$\|u - \tilde{u}_n\|_{H^1(U)} = O\Big(\frac{\max(\|a\|_{C^2(\bar{U})}, 1)^{12}}{\min(a_{\min}, 1)^{11}} \|f\|_{H^2(U)} 2^{-2n}\Big). \tag{27}$$

The proof of Theorem 1 is given in Appendix A.

3.2. Construction of the unbiased estimator

In this section, we apply the results obtained in Section 3.1 to construct an unbiased estimator with both finite variance and finite expected computational cost through (9). We start with providing an upper bound of $\mathbb{E}[\mathcal{Q}(u) - \mathcal{Q}(\tilde{u}_n)]^2$.

Proposition 1. *Under assumptions A1-A4, we have*

$$\mathbb{E}[\mathcal{Q}(u) - \mathcal{Q}(\tilde{u}_n)]^2 = O(\kappa_q 2^{-4n}), \tag{28}$$

where u is the solution to (22) *and $\tilde{u}_n$ be the solution to* (25)*, and κ_q the Lipschitz constant appeared in condition A4.*

The proof is a direct application of (26), Theorem 1 and A4 and therefore is omitted. We proceed to the construction of the unbiased estimator $\tilde{Z}$ via (9). Choose $\mathbb{P}(N = n) = p_n \propto 2^{-\frac{4+d}{2}n}$. For each n, let $\tilde{u}_{n-1}$ and $\tilde{u}_n$ be defined as in (25) with respect to the same a and f. Notice that the computation of $\tilde{u}_n$ requires the values of a and f only on the vertices of $\mathcal{T}_n$. Then, Z_{n-1} and Z_n are given by $Z_{n-1} = \mathcal{Q}(\tilde{u}_{n-1})$ and $Z_n = \mathcal{Q}(\tilde{u}_n)$. With this coupling, according to Proposition 1, we have that $\mathbb{E}(Z_n - Z_{n-1})^2 \leq 2\mathbb{E}[\mathcal{Q}(\tilde{u}_n) - \mathcal{Q}(u)]^2 + 2\mathbb{E}[\mathcal{Q}(\tilde{u}_{n-1}^{(2)}) - \mathcal{Q}(u)]^2 = O(2^{-4n})$. According to equation (10), for $d = \dim(U) \leq 3$, we have $\mathbb{E}(\tilde{Z}^2) \leq \sum_{n=1}^{\infty} 2^{-4n}/2^{-(4+d)n/2} < \infty$. Furthermore, (25)

requires solving $O(2^{dn})$ sparse linear equations. The computational cost of obtaining u_n is $O(n2^{dn})$. According to (13), the expected cost of generating a single copy of $\tilde{Z}$ is

$$\mathbb{E}(C) = \sum_{n=1}^{\infty} p_n c_n \leq \sum_{i=1}^{\infty} n2^{dn} \cdot 2^{-(4+d)n/2} < \infty.$$

This guarantees that the unbiased estimator $\tilde{Z}$ has a finite variance and can be generated in finite expected time.

Sampling of the random coefficients. In some cases, we also need to consider the computational complexity for simulating a and f in addition to the computational cost of solving the PDE. For example, if $\log a(x)$ is modeled as a Gaussian random field, then the computational complexity for generating $\log a(x)$ over $O(2^{dn})$ grid points is $O(2^{3dn})$ if the Cholesky decomposition is adopted. This computational complexity is of a higher order than that of solving an isoparametric FEM with a grid size 2^{-n}, and the corresponding unbiased estimator may not have a finite variance and finite computational cost at the same time.

If the random fields a and f can be approximated by $\{a_k\}$ and $\{f_k\}$ with a relatively low computational cost, then we can still achieve a similar error bound for the resulting numerical solver. In the next example, we show a situation where we can construct such an approximation for $a(\cdot)$.

Example 2. Assume that $\log a(x) = g(x)$, and $g(x)$ has the following expansion. For all $x \in U$, $g(x) = \sum_{l=0}^{\infty} \lambda_l W_l \phi_l(x)$, where $W_1, W_2, ...$ are i.i.d. random variables following the standard normal distribution, $\{\lambda_l\}$ is a sequence of numbers that tend to 0 as $l \to \infty$ and $\{\phi_l\}$ is a sequence of functions over U. To approximate the Gaussian random field $g(x)$, we could use the truncated field $g_k(x) = \sum_{l=0}^{k} \lambda_l W_l \phi_l(x)$.The computational cost for simulating $g_k(x)$ over $O(2^{dn})$ grid points is of the order $O(k \times 2^{dn})$. We can see that if $k = k_n$ grows in a speed no faster than $O(n2^{dn})$, then the computational complexity for generating $g_{k_n}(x)$ is much smaller than the cost for simulating $g(x)$ exactly. The approximation accuracy of g_k can be obtained via standard analysis of $g(x) - g_k(x)$ with additional assumptions on the decaying speed of $\lambda_l \|\phi_l\|_{C^2(\bar{U})}$. For more detailed analysis, see, for example, [1].

We omit details of the precise requirement of λ_l and $\phi_l(x)$ and present the following

results under generic assumptions on a_n.

Theorem 2. *Define*

$$\tilde{W}_n = 2^{2n}\|a - a_n\|_{C^2(\bar{U})}.$$

We make the following additional assumptions on the sequence $\{a_n\}$.

B1. $\max_n \mathbb{E}\min_{x\in U}(a_n(x),1)^{-p} < \infty$ *for all* $p > 0$.

B2. $\max_n \mathbb{E}\|a_n\|^p_{C^2(\bar{U})} < \infty$ *for all* $p > 0$.

B3. There exists a constant $\delta > 0$ *such that* $\max_n \mathbb{E}\tilde{W}_n^{2+\delta} < \infty$.

B4. Simulating $a_n(\cdot)$ *at the nodes of* $\mathcal{T}_n$ *requires a computational cost of the order* $O(n2^{dn})$.

Let the solution $\bar{u}_n$ *be the solution to* (25) *with* $a(\cdot)$ *replaced by* $a_n(\cdot)$ *in the bilinear form* $\tilde{b}_n$. *Furthermore, let* $Z_n = \mathcal{Q}(\bar{u}_n)$ *and* $Z_{n-1} = \mathcal{Q}(\bar{u}_{n-1})$ *be constructed with the same sample path* ω. *Then, the unbiased estimator* $\tilde{Z}$ *constructed via* (9) *has a finite variance and a finite computational cost.*

Similar to the simulation of $a(\cdot)$, we could approximate the random field $f(\cdot)$ as well. The analysis is similar and we omit the repetitive details.

4. Simulation Study

4.1. An illustrating example

We start with a simple example for which closed form solution is available and therefore we are able to check the accuracy of the simulation. Let $U = B(0,1)$, $f(x) = 2e^{W_1+W_2x_1+W_3x_2}(2 + W_2x_1 + W_3x_2)$ and $a(x) = e^{W_2x_1+W_3x_2}$, where W_1, W_2, W_3 are independent and identically distributed standard normal random variable. In this example, the solution to (1) is

$$u(x_1, x_2) = e^{W_1}(1 - x_1^2 - x_2^2). \tag{29}$$

We are interested in the output functional $\mathcal{Q}(u) = |u|^2_{H^1(U)}$ whose expectation is in a closed form.

$$\mathbb{E}|u|^2_{H^1(U)} = \mathbb{E}[2\pi e^{2W_1}] = 2\pi e^2 \approx 46.4268.$$

Let $p_n = 0.875 \times 0.125^n$ and $Z_n = \mathcal{Q}(\tilde{u}_n)$ for $n > 0$. Here we define $Z_0 = 0$. Thus, the estimator according to (9) is

$$\tilde{Z} = \frac{Z_N - Z_{N-1}}{p_N}. \tag{30}$$

We perform Monte Carlo simulation with $M = 300000$ replications. The averaged estimator is 46.5572 with the standard deviation 0.8212. Figure 2 shows the histogram of samples of $\tilde{Z}$ and $\log \tilde{Z}$.

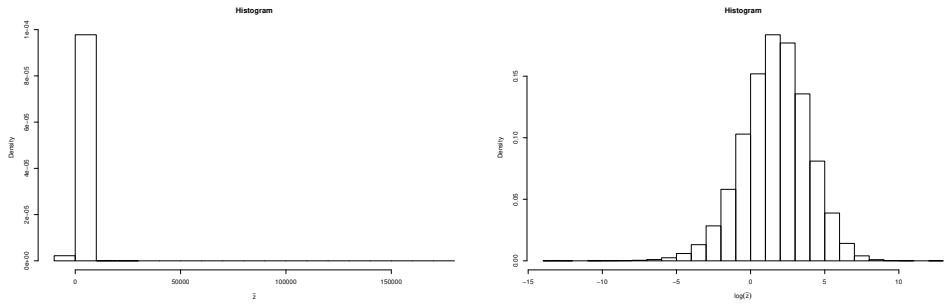

FIGURE 2: Histogram of Monte Carlo sample of $\tilde{Z}$ and $\log \tilde{Z}$ that are defined in Section 4.1.

4.2. Log-normal random field

In this example, we let $U = B(0,1)$, $f(x) = 1$ for all $x \in B(0,1)$ and we consider a more complicated random field $a(x)$. In particular, we let

$$\log a(x_1, x_2) = \sum_{m=1}^{\infty} \frac{1}{2^m}(W_{2m-1}x_1^m + W_{2m}x_2^m), \tag{31}$$

where $W_1, W_2, ...$ are independent and identically distributed standard normal random variable. It is not hard to verify that $a(x_1, x_2)$ satisfies Assumptions A1 and A2. We further approximate the field a by

$$a_n(x_1, x_2) = \exp\{\sum_{m=1}^{3n} \frac{1}{2^m}(W_{2m-1}x_1^m + W_{2m}x_2^m)\}, \tag{32}$$

and compute the finite element solution based on this approximation. We let $Z_n = \mathcal{Q}(\bar{u}_n)$ as discussed on page 16 and take the same estimator (30) and functional $\mathcal{Q}$ as the previous example. We perform Monte Carlo simulation for M = 300000 replications. The averaged estimator for the expectation $\mathbb{E}Q(u)$ is 0.4608 and the standard deviation is 0.0004 for the averaged estimator. Figure 3 shows the histogram of the Monte Carlo sample.

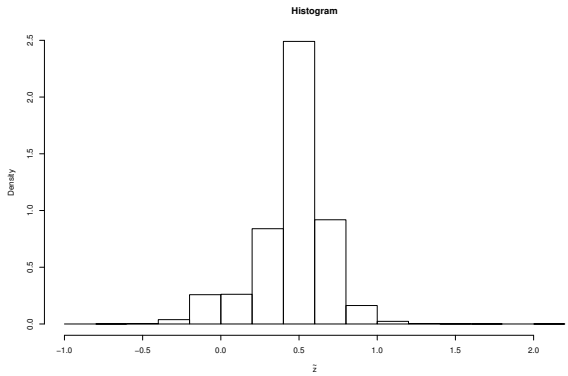

FIGURE 3: Histogram of Monte Carlo sample of $\tilde{Z}$ when log a has a Gaussian covariance.

Appendix A. Proof of the Theorems

In this section, we provide technical proofs of Theorem 1 and Theorem 2. Throughout the proof we will use κ as a generic notation to denote large and not-so-important constants whose value may vary from place to place. Similarly, we use ε as a generic notation for small positive constants.

Before we start the main proof, we first present a proposition on the higher oder regularity of the solution u, whose proof is given after the proofs of theorems.

Proposition 2. *For* $\dim(U) \leq 3$ *with a* $(k+1)$*-time differentiable boundary* ∂U*, if* $a(x) \in C^k(\bar{U})$ *and* $f(x) \in H^{k-1}(U)$ *for some positive integer* k*, then we have*

$$\|u\|_{H^{k+1}(U)} \leq \kappa \frac{\max(\|a\|_{C^k(\bar{U})}, 1)^{\frac{k^2}{2}+\frac{9}{2}k-1}}{\min(a_{\min}, 1)^{\frac{k^2}{2}+\frac{7}{2}k}} \Big(\|f\|_{H^{k-1}(U)} + \|u\|_{L^2(U)} \Big).$$

Proof of Theorem 1. We start with a useful lemma, which is essentially Theorem 43.1 in [3] with the constant $C = \kappa \frac{\|a\|_{C(\bar{U})}}{\min(a_{\min},1)}$ being explicit. We omit the details of the proof of this lemma.

Lemma 1.

$$\|u - \tilde{u}_n\|_{H^1(U)} \leq \kappa \frac{\|a\|_{C(\bar{U})}}{\min(a_{\min}, 1)} 2^{-2n} \times \left\{ \|u\|_{H^3(U)} + \|a\|_{C^2(\bar{U})} \|u\|_{H^3(U)} + \|f\|_{H^2(U)} \right\}. \tag{33}$$

Combining the above display with Proposition 2 with $k = 2$, we have

$$\|u - \tilde{u}_n\|_{H^1(U)} = O\left(2^{-2n} \kappa(a,2) \frac{\|a\|^2_{C^2(\bar{U})}}{\min(a_{\min}, 1)} (\|f\|_{H^2(U)} + \|u\|_{L^2(U)}) \right), \tag{34}$$

where $\kappa(a,k) = \frac{\max(\|a\|_{C^k(\bar{U})},1)^{\frac{k^2}{2}+\frac{9}{2}k-1}}{\min(a_{\min},1)^{\frac{k^2}{2}+\frac{7}{2}k}}$. That is,

$$\|u-\tilde{u}_n\|_{H^1(U)} = O\left(2^{-2n}\frac{\max(\|a\|_{C^2(\bar{U})},1)^{12}}{\min(a_{\min},1)^{10}}(\|f\|_{H^2(U)}+\|u\|_{L^2(U)})\right). \tag{35}$$

Thanks to (26), the above display can be further bounded by

$$\|u\|_{L^2(U)} \leq \kappa\frac{\|f\|_{L^2(U)}}{\min(a_{\min},1)}.$$

We complete the proof by combining the above expression and (35). □

Proof of Theorem 2. We start with the inequality,

$$\|\bar{u}_n - u\|_{H^1(U)} \leq \|\bar{u}_n - \tilde{u}_n\|_{H^1(U)} + \|\tilde{u}_n - u\|_{H^1(U)}. \tag{36}$$

The second term on the right-hand side of the above inequality is already bounded from above by Theorem 1. That is,

$$\|\bar{u}_n - u\|_{H^1(U)} \leq \|u_n - \tilde{u}_n\|_{H^1(U)} + O(2^{-2n}\frac{\max(\|a\|_{C^2(\bar{U})},1)^{12}}{\min(a_{\min},1)^{11}}\|f\|_{H^2(U)}). \tag{37}$$

We proceed to an upper bound of the first term. Let $\bar{b}_n$ be the bilinear form with a being replaced by a_n in the bilinear form $\tilde{b}_n$. Note that $\bar{u}_n$ is obtained by replacing $\tilde{b}_n$ by $\bar{b}_n$ in (25), we have

$$\bar{b}_n(\bar{u}_n, w) = \tilde{L}_n(w) = \tilde{b}_n(\tilde{u}_n, w) \tag{38}$$

for all $w \in V_n$. Subtracting $\tilde{b}_n(\bar{u}_n, w)$ on both sides, we arrive at

$$(\bar{b}_n - \tilde{b}_n)(\bar{u}_n, w) = \tilde{b}_n(\tilde{u}_n - \bar{u}_n, w), \tag{39}$$

where we write $(\bar{b}_n - \tilde{b}_n)(v,w) = \bar{b}_n(v,w) - \tilde{b}_n(v,w)$. Set $w = \tilde{u}_n - \bar{u}_n$ in the above display, we arrive at

$$(\bar{b}_n - \tilde{b}_n)(\tilde{u}_n, \tilde{u}_n - \bar{u}_n) = \tilde{b}_n(\tilde{u}_n - \bar{u}_n, \tilde{u}_n - \bar{u}_n). \tag{40}$$

According to the same arguments as those on [3, page 258-260], the right-hand side of the above display is bounded from below by

$$\tilde{b}_n(\tilde{u}_n - \bar{u}_n, \tilde{u}_n - \bar{u}_n) \geq \varepsilon a_{\min}\|\tilde{u}_n - \bar{u}_n\|^2_{H^1(U)}. \tag{41}$$

On the other hand, we have

$$|(\bar{b}_n - \tilde{b}_n)(\bar{u}_n, \tilde{u}_n - \bar{u}_n)| \leq \|a - a_n\|_{C(\bar{U})}\|\bar{u}_n - \tilde{u}_n\|_{H^1(U)}\|\bar{u}_n\|_{H^1(U)}. \tag{42}$$

Combining (40), (41) and (42), we arrive at

$$\|\tilde{u}_n - \bar{u}_n\|_{H^1(U)} \leq \kappa \frac{\|a - a_n\|_{C(\bar{U})} \|\bar{u}_n\|_{H^1(U)}}{a_{\min}} \leq \kappa 2^{-2n} \frac{\|\bar{u}_n\|_{H^1(U)}}{a_{\min}} \tilde{W}_n. \tag{43}$$

Because a_n satisfies Assumptions A1-A2, we can apply Theorem 1 to the solution $\bar{u}_n$ and arrive at

$$\|\bar{u}_n\|_{H^1(U)} = O\Big(\frac{\max(\|a_n\|_{C^2(\bar{U})}, 1)^{12}}{\min(\min_{x\in U}(a_n(x)), 1)^{11}} \|f\|_{H^2(U)}\Big). \tag{44}$$

Combining (37), (43) and (44), we arrive at

$$\|\bar{u}_n - u\|_{H^1(\bar{U})} = O\left(\{\frac{\max(\|a\|_{C^2(\bar{U})}, 1)^{12}}{\min(a_{\min}, 1)^{11}} + \frac{\max(\|a_n\|_{C^2(\bar{U})}, 1)^{12}}{\min(\min_{x\in U}(a_n(x)), 1)^{11} \min(a_{\min}, 1)} \tilde{W}_n\} \|f\|_{H^2}(U) 2^{-2n}\right). \tag{45}$$

The rest of the proof is similar to the analysis under Proposition 1. We omit the details. □

For the rest of the section, we provide the proof for Proposition 2. Proposition 2 is similar to Theorem 5 in Chapter 6.3 of [7] but we provide explicitly the dependence of constants on a and f.

Proof of Proposition 2. We prove Proposition 2 by proving the following result for the weak solution $w \in H_0^1(U)$ to a more general PDE,

$$\begin{cases} -\nabla \cdot (A \nabla w) &= f \text{ in } U \\ w &= 0 \text{ on } \partial U, \end{cases} \tag{46}$$

where $A(x) = (A_{ij}(x))_{1\leq i,j\leq d}$ is a symmetric positive definite matrix function in the sense that there exist $A_{\min} > 0$ satisfying

$$\xi^T A(x) \xi \geq A_{\min} |\xi|^2 \tag{47}$$

for all $x \in \bar{U}$ and $\xi \in R^d$. Assume that $A_{ij}(x) \in C^k(\bar{U})$ for all $i, j = 1, ..., d$. Then, it is sufficient to show that

$$\|w\|_{H^{k+1}(U)} \leq \kappa_r(A, k) \Big(\|f\|_{H^{k-1}(U)} + \|w\|_{L^2(U)}\Big), \tag{48}$$

where $\kappa_r(A, k) = \kappa \frac{\max(\|A\|_{C^k(\bar{U})}, 1)^{\frac{k^2}{2} + \frac{9}{2}k - 1}}{\min(A_{\min}, 1)^{\frac{k^2}{2} + \frac{7}{2}k}}$, and $\|A\|_{C^k(\bar{U})} = \max_{1\leq i,j\leq d} \|A_{ij}\|_{C^k(\bar{U})}$.

Let $B^0(0,r)$ denote the open ball $\{x : |x| < r\}$ and $R^d_+ = \{x \in R^d : x_d > 0\}$. We will first prove that if $U = B^0(0,r) \cap R^d_+$ and $V = B^0(0,t) \cap R^d_+$, then for all t and r such that and $0 < t < r$,

$$\|w\|_{H^{m+2}(V)} \leq \kappa_{r,t,m+1} \frac{\max(\|A\|_{C^k(\bar{U})}, 1)^{\frac{(m+1)^2}{2}+\frac{9}{2}(m+1)-1}}{\min(A_{\min}, 1)^{\frac{(m+1)^2}{2}+\frac{7}{2}(m+1)}} \Big(\|f\|_{H^m(U)} + \|w\|_{L^2(U)} \Big), \tag{49}$$

where $\kappa_{r,t,m+1}$ is a constant depending only on r, t, and $m+1$. The following lemma establish (49) for $m = 0$.

Lemma 2. (Boundary H^2-regularity.) *Assume ∂U is twice differentiable and $A(x)$ satisfies* (47)*. Assume that $A_{ij}(x) \in C^1(\bar{U})$ for all $i, j = 1, ..., d$. Suppose furthermore $w \in H^1_0(U)$ is a weak solution to the elliptic PDE with boundary condition* (46)*. Then $w \in H^2(U)$ and*

$$\|w\|_{H^2(U)} \leq \kappa \frac{\max(\|A\|_{C^1(\bar{U})}, 1)^4}{\min(A_{\min}, 1)^4} \Big(\|f\|_{L^2(U)} + \|w\|_{L^2(U)} \Big).$$

We establish (49) by induction. Suppose for some m

$$\|w\|_{H^{m+1}(W)} \leq \kappa_{t,s,m} \frac{\max(\|A\|_{C^k(\bar{U})}, 1)^{\frac{m^2}{2}+\frac{9}{2}m-1}}{\min(A_{\min}, 1)^{\frac{m^2}{2}+\frac{7}{2}m}} (\|f\|_{H^{m-1}(U)} + \|w\|_{L^2(U)}), \tag{50}$$

where

$$W = B^0(0,s) \cap R^d_+, \text{ and } s = \frac{t+1}{2}. \tag{51}$$

Since w is a weak solution to (46), it satisfies the integration equation

$$\int_D \nabla w(x)^T A(x) \nabla v(x) dx = \int_D f(x) v(x) dx, \text{ for all } v \in H^1_0(U). \tag{52}$$

Let $\alpha = (\alpha_1, ..., \alpha_d)$ be a multiple index with such that $\alpha_d = 0$ and $|\alpha| = m$. We consider the multiple weak derivative $\bar{w} = D^\alpha w$ and investigate the PDE that $\bar{w}$ satisfies. For any $\bar{v} \in C^\infty_c(W)$, where $C^\infty_c(W)$ is the space of infinitely differentiable functions that have compact support in W, we plug $v = (-1)^{|\alpha|} D^\alpha \bar{v}$ into (52). With some calculations, we have

$$\int_W (\nabla \bar{w}(x))^T A(x) \nabla \bar{v}(x) = \int_W \bar{f}(x) \bar{v}(x) dx,$$

where

$$\bar{f} = D^\alpha f - \sum_{\beta \leq \alpha, \beta \neq \alpha} \binom{\alpha}{\beta} \Big[-\nabla \cdot (D^{\alpha-\beta} A \nabla D^\beta w) \Big]. \tag{53}$$

Consequently, $\bar{w}$ is a weak solution to the PDE

$$-\nabla \cdot (A\nabla \bar{w}) = \bar{f} \quad \text{for } x \text{ in } W. \tag{54}$$

Furthermore, we have the boundary condition $\bar{w}(x) = 0$ for $x \in \partial W \cap \{x_d = 0\}$. By the induction assumption (50) and (53), we have

$$\|\bar{f}\|_{L^2(W)} \leq \|f\|_{H^m(U)} + \kappa_{t,s,m} \frac{\max(\|A\|_{C^k(\bar{U})}, 1)^{\frac{m^2}{2}+\frac{9}{2}m-1}}{\min(A_{\min}, 1)^{\frac{m^2}{2}+\frac{7}{2}m}} \|A\|_{C^{m+1}(\bar{U})} \Big(\|f\|_{H^{m-1}(U)} + \|w\|_{L^2(U)} \Big). \tag{55}$$

According to the definition of $\bar{w}$, we have

$$\|\bar{w}\|_{L^2(W)} \leq \|w\|_{H^m(W)}. \tag{56}$$

Applying Lemma 2 to $\bar{w}$ with (55) and (56), we have

$$\begin{aligned} &\|D^\alpha w\|_{H^2(V)} \\ \leq &\kappa_{t,s,m} \kappa \frac{\max(\|A\|_{C^1(\bar{U})}, 1)^4}{\min(A_{\min}, 1)^4} \frac{\max(\|A\|_{C^k(\bar{U})}, 1)^{\frac{m^2}{2}+\frac{9}{2}m-1}}{\min(A_{\min}, 1)^{\frac{m^2}{2}+\frac{7}{2}m}} \|A\|_{C^{m+1}(\bar{U})} \Big(\|f\|_{H^m(U)} + \|w\|_{L^2(U)} \Big). \end{aligned} \tag{57}$$

Because α is an arbitrary multi-index such that $\alpha_d = 0$, and $|\alpha| = m$, (57) implies that $D^\beta w \in L^2(W)$ for any multiple index β such that $|\beta| \leq m+2$ and $\beta_d = 0, 1, 2$. We now extend this result to multiple index β whose last component is greater than 2. Suppose for all β such that $|\beta| \leq m+2$ and $\beta_d \leq j$, we have

$$\|D^\beta w\|_{H^2(V)} \leq \kappa_r^{(j)} \Big(\|f\|_{H^m(U)} + \|w\|_{L^2(U)} \Big), \tag{58}$$

where $\kappa_r^{(j)}$ is a constant depending on A, m and j that we are going to determine later. We establish the relationship between $\kappa_r^{(j)}$ and $\kappa_r^{(j+1)}$. For any γ that is a multiple index such that $|\gamma| = m+2$ and $\gamma_d = j+1$, we use (58) to develop an upper bound for $\|D^\gamma w\|_{H^2(V)}$. In particular, let $\beta = (\gamma_1, .., \gamma_{d-1}, j-1)$. According to the remark (ii) after Theorem 1 of Chapter 6.3 in [7], we have that

$$-\nabla \cdot (A\nabla(D^\beta w)) = f^\dagger \text{ in } W \text{ a.e}, \tag{59}$$

where

$$f^\dagger = D^\beta f - \sum_{\delta \leq \beta, \delta \neq \beta} \binom{\beta}{\delta} \Big[-\nabla \cdot (D^{\beta-\delta} A \nabla D^\delta w) \Big]. \tag{60}$$

Notice that

$$-\nabla\cdot(A\nabla(D^\beta w)) = -A_{dd}D^\gamma w + \text{ sum of terms involves at most } j \text{ times weak derivatives of } w \text{ with respect to } x_d \text{ and at most } m+2 \text{ times derivatives in total.}$$

According to (58), (59), (60), and the above display, we have

$$\|D^\gamma w\|_{L^2(U)} \le \kappa\frac{1}{\min(A_{\min},1)}\Big\{\|A\|_{C^{m+1}(\bar U)}\kappa_r^{(j)}\Big(\|f\|_{H^m(U)}+\|w\|_{L^2(U)}\Big)+\|f\|_{H^m(U)}\Big\}.$$

Therefore,

$$\|D^\gamma w\|_{L^2(U)} \le \kappa_r^{(j+1)}\Big(\|f\|_{H^m(U)}+\|w\|_{L^2(U)}\Big),$$

where

$$\kappa_r^{(j+1)} = \kappa_r^{(j)}\frac{\max(\|A\|_{C^{m+1}(\bar U)},1)}{\min(A_{\min},1)}. \tag{61}$$

The above expression provides a relationship for $\kappa_r^{(j+1)}$ and $\kappa_r^{(j)}$. According to (57),

$$\kappa_r^{(2)} = \kappa_{t,s,m}\kappa\frac{\max(\|A\|_{C^1(\bar U)},1)^4}{\min(A_{\min},1)^4}\frac{\max(\|A\|_{C^k(\bar U)},1)^{\frac{m^2}{2}+\frac{9}{2}m-1}}{\min(A_{\min},1)^{\frac{m^2}{2}+\frac{7}{2}m}}\max(\|A\|_{C^{m+1}(\bar U)},1).$$

Using (61) and the above initial value for the iteration, we have

$$\begin{aligned}&\kappa_r^{(m+2)}\\ =&\kappa_{t,s,m}\kappa\frac{\max(\|A\|_{C^1(\bar U)},1)^4}{\min(A_{\min},1)^4}\frac{\max(\|A\|_{C^k(\bar U)},1)^{\frac{m^2}{2}+\frac{9}{2}m-1}}{\min(A_{\min},1)^{\frac{m^2}{2}+\frac{7}{2}m}}\\ &\times\max(\|A\|_{C^{m+1}(\bar U)},1)\Big\{\frac{\max(\|A\|_{C^{m+1}(\bar U)},1)}{\min(A_{\min},1)}\Big\}^m.\end{aligned}$$

Consequently,

$$\|w\|_{H^{m+2}(V)} \le \kappa_{t,s,m}\kappa\frac{\max(\|A\|_{C^k(\bar U)},1)^{\frac{m^2}{2}+\frac{11}{2}m+4}}{\min(A_{\min},1)^{\frac{m^2}{2}+\frac{9}{2}m+4}}\Big(\|f\|_{H^m(U)}+\|w\|_{L^2(V)}\Big).$$

Using induction, we complete the proof of (48) for the case where U is a half ball.

Now we extend the result to the case that U has a C^{k+1} boundary ∂U. We first prove the theorem locally for any point $x^0\in\partial U$. Because ∂U is $(k+1)$-time differentiable, with possibly relabeling, the coordinates of x there exist a function $\gamma: R^{d-1}\to R$ and $r>0$ such that,

$$B(x^0,r)\cap U = \{x\in B(x^0,r): x_d > \gamma(x_1,...,x_{d-1})\}.$$

Let $\Phi = (\Phi_1, ..., \Phi_d)^T : R^d \to R^d$ be a function such that

$$\Phi_i(x) = x_i \text{ for } i = 1, ..., d-1 \text{ and } \Phi_d(x) = x_d - \gamma(x_1, ..., x_{d-1}).$$

Let $y = \Phi(x)$ and choose $s > 0$ sufficiently small such that

$$U^* = B^0(0, s) \cap \{y_d > 0\} \subset \Phi(U \cap B(x^0, r)).$$

Furthermore, we let $V^* = B^0(0, \frac{s}{2}) \cap \{y_d > 0\}$ and set

$$w^*(y) = w(x) = w(\Phi^{-1}(y)).$$

With some calculation, we have that w^* is a weak solution to the PDE

$$-\nabla \cdot \Big(A^*(y)\nabla w^*(y)\Big) = f^*(y),$$

where $A^*(y) = J(y)A(\Phi^{-1}(y))J^T(y)$ and $J(y)$ is the Jacobian matrix for Φ with $J_{ij}(y) = \frac{\partial \Phi_i(x)}{\partial x_j}|_{x=\Phi^{-1}(y)}$, and $f^*(y) = f(\Phi^{-1}(y))$. In addition, $w^* \in H^1(U^*)$ and $w^*(y) = 0$ for $y \in \partial U^* \cap \{y_d = 0\}$. It is easy to check A^* is symmetric and $A^*_{ij} \in C^k(\bar{U})$ for all $1 \le i, j \le d$. Furthermore, according to the definition of J and Φ, all the eigenvalues of $J(y)$ are 1 and thus $\zeta^T A^*(y)\xi \ge A_{\min}|J^T(y)\xi|^2 \ge \varepsilon A_{\min}|\xi|^2$ for all $\xi \in R^d$. By substituting U, V, A, f with U^*, V^*, A^* and f^* in (49) we have

$$\|w^*\|_{H^2(V^*)} \le \kappa_r(A, k)\Big(\|w^*\|_{L^2(U^*)} + \|f^*\|_{H^{k-1}(U^*)}\Big).$$

According to the definitions of w^* and f^*, the above display implies

$$\|w\|_{H^2(\Phi^{-1}(V^*))} \le \kappa_r(A, k)\Big(\|w\|_{L^2(U)} + \|f\|_{H^{k-1}(U)}\Big).$$

Because U is bounded, ∂U is compact and thus can be covered by finitely many sets $\Phi^{-1}(V_1^*), .., \Phi^{-1}(V_K^*)$ that are constructed similarly as $\Phi^{-1}(V^*)$. We finish the proof by combining the result for points around ∂U and the following Lemma 3 for interior points.

Lemma 3. (Higher order interior regularity.) *Under the setting of Lemma 2, we assume that ∂U is C^{k+1}, $A_{ij}(x) \in C^k(U)$ for all $i, j = 1, ..., d$, and $f \in H^{k-1}(U)$, and that $w \in H^1(U)$ is one of the weak solutions to the PDE* (46) *without boundary condition. Then, $w \in H^{k+1}_{loc}(U)$. For each open set $V \subsetneqq U$*

$$\|w\|_{H^{k+1}(V)} \le \kappa_i(A, k)\Big(\|f\|_{H^{k-1}(U)} + \|w\|_{L^2(U)}\Big),$$

where $\kappa_i(A, k) = \frac{\max(\|A\|_{C^k(\bar{U})}, 1)^{3k-1}}{\min(A_{\min}, 1)^{2k}}\kappa$, and κ is a constant depending on V.

□

Appendix B. Proof of supporting lemmas

In this section, we provide the proofs for lemmas that are necessary for the proof of Proposition 2. We start with a useful lemma showing $w \in H^2_{loc}(U)$ which will be used in the proof of Lemma 2

Lemma 4. (Interior H^2-regularity.) *Under the setting of Lemma 2, we further assume that $A_{ij}(x) \in C^1(\bar{U})$ for all $i, j = 1, ..., d$, and $f \in L^2(U)$, and that $w \in H^1(U)$ is one of the weak solutions to the PDE (46) without boundary condition. Then, $w \in H^2_{loc}(U)$. For each open subset $V \subsetneqq U$, there exist κ depending on V such that*

$$\|w\|_{H^2(V)} \leq \kappa \frac{\max(\|A\|_{C^1(U)}, 1)^2}{\min(A_{\min}, 1)^2} \Big(\|f\|_{L^2(U)} + \|w\|_{L^2(U)} \Big),$$

where we define the norm $\|A\|_{C^1(\bar{U})} = \max_{1 \leq i,j \leq d} \|A_{ij}\|_{C^1(\bar{U})}$.

Proof of Lemma 4. Let h be a real number whose absolute value is sufficiently small, we define the difference quotient operator

$$D_k^h w(x) = \frac{w(x + he_k) - w(x)}{h},$$

where e_k is the kth unit vector in R^d. According to Theorem 3 in Chapter 5.8 of [7], if there exist a positive constant κ such that $\|D_k^h w\|_{L^2(U)} \leq \kappa$ for all h, then $\frac{\partial w}{\partial x_k} \in L^2(U)$ and $\|\frac{\partial w}{\partial x_k}\|_{L^2(U)} \leq \kappa$. We use this theorem and seek for an upper bound of

$$\int_V |U_k^h \nabla w|^2 dx, \tag{62}$$

for $k = 1, ..., d$ for the rest of the proof.

We derive a bound of (62) by plugging an appropriate v in (52). Let W be an open set such that $V \subsetneqq W \subsetneqq U$. We select a smooth function ζ such that

$$\zeta = 1 \text{ on } V, \qquad \zeta = 0 \text{ on } W^c, \qquad \text{and } 0 \leq \zeta \leq 1.$$

We plug

$$v = -D_k^{-h}(\zeta^2 D_k^h w)$$

into (52), and have

$$-\int_D \nabla w^T A \nabla [D_k^{-h}(\zeta^2 D_k^h w)] dx = -\int_D f D_k^{-h}(\zeta^2 D_k^h w) dx. \tag{63}$$

We give a lower bound of the left-hand side of (63) and an upper bound of the right-hand. We use two basic formulas that are similar to integration by part and derivative of product respectively. For any functions $w_1, w_2 \in L^2(U)$, such that $w_2(x) = 0$ if $dist(x, \partial U) < h$, we have

$$\int_D w_1 D_k^{-h} w_2 dx = -\int_D D_k^h w_1 w_2 dx \text{ and } D_k^h(w_1 w_2) = w_1^h D_k^h w_2 + w_2 D_k^h w_1,$$

where we define $w_1^h(x) = w_1(x + he_k)$. Similarly, we define the matrix function $A^h = A(x + he_k)$. Applying the above formulas to the left hand side of (63), we have

$$\begin{aligned}
& -\int_D \nabla w^T A \nabla [D_k^{-h}(\zeta D_k^h w)] dx \\
= & \int_D D_k^h(\nabla w^T A)\nabla(\zeta^2 D_k^h w) dx \\
= & \int_D D_k^h(\nabla w^T) A^h \nabla(\zeta^2 D_k^h w) + \nabla w^T D_k^h A \nabla(\zeta^2 D_k^h w) dx \\
= & \underbrace{\int_D \zeta^2 D_k^h \nabla w^T A^h D_k^h \nabla w dx}_{J_1} \\
& + \underbrace{\int_D 2\zeta(D_k^h \nabla w^T A^h \nabla \zeta) D_k^h w + 2\zeta(\nabla w^T D_k^h A \nabla \zeta) D_k^h w + \zeta^2 \nabla w^T D_k^h A D_k^h \nabla w dx}_{J_2}.
\end{aligned}$$

J_1 in the above expression has a lower bound

$$J_1 \geq A_{\min} \int_D \zeta^2 |D_k^h \nabla w|^2 dx$$

due to the positively definitiveness of $A(x)$. $|J_2|$ is bounded above by

$$|J_2| \leq \kappa \|A\|_{C^1(\bar{U})} \Big(\int_D \zeta |D_k^h \nabla w||D_k^h w| + \zeta |\nabla w||D_k^h w| + \zeta |\nabla w||D_k^h \nabla w| dx \Big). \quad (64)$$

The expression (64) can be further bounded by

$$|J_2| \leq \frac{A_{\min}}{2} \int_D \zeta^2 |D_k^h \nabla w|^2 dx + \kappa \|A\|_{C^1(\bar{U})} \times (1 + \frac{\|A\|_{C^1(\bar{U})}}{A_{\min}}) \int_W |\nabla w|^2 + |D_k^h w|^2 dx. \quad (65)$$

thanks to Cauchy-Schwarz inequality. According to Theorem 3 in Chapter 5.8 of [7],

$$\int_W |D_k^h w|^2 dx \leq \kappa \int_W |\nabla w|^2 dx. \quad (66)$$

Therefore, (65) is bounded above by

$$|J_2| \leq \frac{A_{\min}}{2} \int_D \zeta^2 |D_k^h \nabla w|^2 dx + \kappa^2 \|A\|_{C^1(\bar{U})} \times (1 + \frac{\|A\|_{C^1(\bar{U})}}{A_{\min}}) \int_W |\nabla w|^2 dx. \quad (67)$$

Combining (64) and (67), we have

$$\begin{aligned}&\text{LHS of (63)}\\ =&J_1+J_2\\ \geq&J_1-|J_2|\\ \geq&\frac{A_{\min}}{2}\int_D\zeta^2|D_k^h\nabla w|^2dx-\kappa^2\|A\|_{C^1(\bar{U})}\times(1+\frac{\|A\|_{C^1(\bar{U})}}{A_{\min}})\int_W|\nabla w|^2dx.\end{aligned}\tag{68}$$

We proceed to an upper bound of the right hand side of (63). According to (66), we have

$$\begin{aligned}&\int_D|D_k^{-h}(\zeta^2D_k^hw)|^2dx\\ \leq\ &\kappa\int_D|\nabla(\zeta^2D_k^hw)|^2dx\\ \leq\ &\kappa\int_W4|D_k^hw|^2|\nabla\zeta|^2\zeta^2+\zeta^2|D_k^h\nabla w|^2dx\\ \leq\ &\kappa^3\int_W|\nabla w|^2+\zeta^2|D_k^h\nabla w|^2dx.\end{aligned}\tag{69}$$

Apply Cauchy's inequality to the right-hand side of (63), we have

$$\text{RHS of (63)}\leq\int_D|f||D_k^{-h}(\zeta^2D_k^hw)|dx\leq\frac{2\kappa^3}{A_{\min}}\int_D|f|^2dx+\frac{A_{\min}}{4\kappa^3}\int_D|D_k^{-h}(\zeta^2D_k^hw)|^2dx.\tag{70}$$

We combine (69) and (70),

$$\text{RHS of (63)}\leq\frac{A_{\min}}{4}\int_W\zeta^2|D_k^h\nabla w|^2dx+\frac{A_{\min}}{4}\int_W|\nabla w|^2dx+\frac{2\kappa^3}{A_{\min}}\int_W|f|^2dx.\tag{71}$$

Combining (68) and (71), we have

$$\int_D\zeta^2|D_k^h\nabla w|^2dx\leq\frac{8\kappa^3}{A_{\min}^2}\int_W|f|^2dx+\Big[1+4\kappa^2\|A\|_{C^1(\bar{U})}\frac{\|A\|_{C^1(\bar{U})}+A_{\min}}{A_{\min}^2}\Big]\int_W|\nabla w|^2dx.\tag{72}$$

Therefore,

$$\int_D\zeta^2|D_k^h\nabla w|^2dx\leq\kappa\frac{\max(\|A\|_{C^1(\bar{U})},1)^2}{\min(A_{\min},1)^2}\Big(\int_W|f|^2dx+\int_W|\nabla w|^2\Big).\tag{73}$$

Now we give an upper bound of $\int_D|\nabla w|$ by taking $v=\tilde{\zeta}^2w$ in (52), where we choose $\tilde{\zeta}$ to be a smooth function such that $\tilde{\zeta}=1$ on W and $\tilde{\zeta}=0$ on U^c. Using similar arguments as that for (73), we have

$$\int_W|\nabla w|^2dx\leq\kappa\frac{\max(\|A\|_{C^1(\bar{U})},1)^2}{\min(A_{\min},1)^2}\Big(\int_W|f|^2dx+\int_W|\nabla w|^2\Big).\tag{74}$$

(73) and (74) together give

$$\int_D \zeta^2 |D_k^h \nabla w|^2 dx \le \kappa \frac{\max(\|A\|_{C^1(\bar{U})}, 1)^4}{\min(A_{\min}, 1)^4} \int_D |f|^2 + |w|^2 dx. \tag{75}$$

We complete our proof by combining (75) for all $k = 1, ..., d$. □

Proof of Lemma 2. We first consider a special case when U is a half ball

$$U = B^0(0,1) \cap R_+^d.$$

Let $V = B^0(0, \frac{1}{2}) \cap R_+^d$, and select a smooth function ζ such that

$$\zeta = 1 \text{ on } B(0, \frac{1}{2}), \zeta = 0 \text{ on } B(0,1)^c, \text{ and } 0 \le \zeta \le 1.$$

For $k = 1, ..., d-1$, we plug

$$v = -D_k^{-h}(\zeta^2 D_k^h w)$$

into (52). Using the same arguments for deriving (72) as in the proof for Lemma 4, we obtain that

$$\int_V |D_k^h \nabla w|^2 dx \le \kappa \frac{\max(\|A\|_{C^1(\bar{U})}, 1)^2}{\min(A_{\min}, 1)^2} \int_W |f|^2 + |\nabla w|^2 dx.$$

The above display holds for arbitrary h, so we have

$$\sum_{i,j=1, i+j<2d}^{d} \int_V |\frac{\partial^2 w}{\partial x_i \partial x_j}|^2 dx \le \kappa \frac{\max(\|A\|_{C^1(\bar{U})}, 1)^2}{\min(A_{\min}, 1)^2} \int_W |f|^2 + |\nabla w|^2 dx. \tag{76}$$

We proceed to an upper bound for

$$\int_V |\frac{\partial^2 w}{\partial x_d \partial x_d}|^2 dx.$$

According to the remark (ii) after Theorem 1 in Chapter 6.3 of [7], with the interior regularity obtained by Lemma 4, w solves (46) almost everywhere in U. Consequently,

$$A_{dd} \frac{\partial^2 w}{\partial x_d \partial x_d} = - \sum_{i,j=1, i+j<2d}^{d} A_{ij} \frac{\partial^2 w}{\partial x_i \partial x_j} - \sum_{i,j=1}^{d} \frac{\partial A_{ij}}{\partial x_j} \frac{\partial w}{\partial x_i} - f \text{ a.e.}$$

Note that $A_{dd} \ge A_{\min}$, so the above display implies that

$$|\frac{\partial^2 w}{\partial x_d \partial x_d}| \le \kappa \frac{\|A\|_{C^1(\bar{U})}}{A_{\min}} \Big(\sum_{i,j=1, i+j<2d}^{d} |\frac{\partial^2 w}{\partial x_i \partial x_j}| + |\nabla w| + |f| \Big).$$

Combining the above display with (76), we have

$$\|w\|_{H^2(V)} \leq \kappa \frac{\max(\|A\|_{C^1(\bar{U})}, 1)^2}{\min(A_{\min}, 1)^2} \Big(\||\nabla w|\|_{L^2(U)} + \|f\|_{L^2(U)} \Big).$$

According to (74), the above display implies

$$\|w\|_{H^2(V)} \leq \frac{\max(\|A\|_{C^1(\bar{U})}, 1)^4}{\min(A_{\min}, 1)^4} \Big(\|w\|_{L^2(U)} + \|f\|_{L^2(U)} \Big).$$

Similar to the proof for Proposition 2, this result can be extended to the case where U has a twice differentiable boundary. We omit the details. □

Proof of Lemma 3. We use induction to prove Lemma 3. When $k = 1$, Lemma 4 gives

$$\|w\|_{H^2(V)} \leq \kappa_i(A, 1) \Big(\|f\|_{L^2(U)} + \|w\|_{L^2(U)} \Big).$$

Suppose for $k = 1, ..., m$, Lemma 3 holds. We intend to prove that for $k = m + 1$,

$$\|w\|_{H^{m+2}}(V) \leq \kappa_i(A, m + 1) \Big(\|f\|_{H^m(U)} + \|w\|_{L^2(U)} \Big).$$

By induction assumption, we have $w \in H^{m+1}_{loc}(U)$ and for any W such that $V \subsetneq W \subsetneq U$

$$\|w\|_{H^{m+1}(W)} \leq \kappa_i(A, m) \Big(\|f\|_{H^{m-1}(U)} + \|w\|_{L^2(U)} \Big). \tag{77}$$

Denote by $\alpha = (\alpha_1, .., \alpha_d)^T$ a multiple index with $|\alpha| = \alpha_1 + ... + \alpha_d = m$. With similar arguments as for (54), we have that $\bar{w} = D^\alpha w$ is a weak solution to the PDE (54) without boundary condition. Similar to the derivation for (57), $w \in H^{m+2}(V)$ and

$$\|w\|_{H^{m+2}(V)} \leq \kappa_i(A, 1)\kappa_i(A, m) \max(\|A\|_{C^{m+1}(\bar{U})}, 1) \Big(\|f\|_{H^m(U)} + \|w\|_{L^2(U)} \Big).$$

We complete the proof by induction. □

Appendix C. Isoparametric finite element method

In this section, we present the precise definition of the finite element method being used. For more details, see [3] and the references therein.

Finite element triplet. The triplet (T, P, Σ) is called an element if T is a Lipschitz domain in R^d; P is a space of functions over T with a finite dimension M; and Σ is a set of linear forms $\eta_1, ..., \eta_M$ with the following P-unisolvent property: given any real numbers $\alpha_1, ..., \alpha_M$, there exists a unique $p \in P$ such that $\eta_i(p) = \alpha_i$, $1 \leq i \leq M$.

Degree of freedom. By the definition of P-unisolvent, there exists $p_1, ..., p_M \in P$ such that $\eta_i(p_j) = \delta_{ij}$ for $1 \le i, j \le M$. Consequently, for all $p \in P$, the following holds:

$$p = \sum_{i=1}^{M} \eta_i(p) p_i.$$

Then, $\eta_1, ..., \eta_M$ are called the degree of freedom of the finite element and $p_1, ..., p_M$ are called the basis functions of P.

Lagrange element. If there exists $a_1, .., a_M$ such that $\eta_i(p) = p(a_i)$ for all $1 \le i \le M$, then the finite element is called a Lagrange finite element. In other words, if (T, P, Σ) is a Lagrange finite element and $p \in P$, then p is completely determined by its value at the nodes $a_1, ..., a_M$. Throughout this paper, we will only consider Lagrange elements.

Affine-equivalence. Let $(T, P, \{p(a_i); 1 \le i \le M\})$ and $(\hat{T}, \hat{P}, \{\hat{p}(\hat{a}_i); 1 \le i \le M\})$ be two Lagrange finite elements. They are called affine-equivalent if there exists an invertible linear operator $B_T : \hat{T} \to T$ and $b_T \in R^d$ such that (i) $F_T : \hat{T} \to T$, $F_T(\hat{x}) = B_T\hat{x} + b_T$, (ii) $a_i = F_T(\hat{a}_i), 1 \le i \le M$, and (iii) $p_i(x) = \hat{p}_i(F_T^{-1}(x)), 1 \le i \le M$. The mapping F_T is called affine mapping.

Isoparametric equivalent elements. A Lagrange finite element $(T, P, \{p(a_i); 1 \le i \le M\})$ is called isoparametric equivalent to $(\hat{T}, \hat{P}, \{\hat{p}(\hat{a}_i); 1 \le i \le M\})$ if there exists an invertible mapping $F : \hat{x} \in \hat{T} \to F(\hat{x}) = (F_i(\hat{x}))_{i=1}^d$ such that $F_i \in \hat{P}$, $1 \le i \le M$ and (i) $T = F(\hat{T})$, (ii) $P = \{p = \hat{p} \circ F^{-1}; \hat{p} \in \hat{P}\}$, and (iii) $a_i = F(\hat{a}_i)$ for $1 \le i \le M$. In particular, when F is a linear mapping, these two finite elements are affine equivalent.

d-simplex. The set $\{(x_1, .., x_d) : \sum_{i=1}^d x_i = 1, x_i \ge 0, i = 1, .., d\}$ is called a d-simplex. When $d = 1, 2, 3$, the d-simplex is a line segment, triangle, and a tetrahedron respectively.

Isoparametric family and reference element Consider a class of Lagrange finite elements indexed by T, $\mathcal{F} = \{(T, P_T, \{p_T(a_{i,T}); 1 \le i \le M\})\}$. It is called an isoparametric family if there exists a finite element $(\hat{T}, \hat{P}, \{\hat{p}(\hat{a}_i); 1 \le i \le M\})$ such that

all $(T, P_T, \{p_T(a_{i,T}); 1 \le i \le M\}) \in \mathcal{F}$ is isoparametric-equivalent to $(\hat{T}, \hat{P}, \{\hat{p}(\hat{a}_i); 1 \le i \le M\})$. The finite element $(\hat{T}, \hat{P}, \{\hat{p}(\hat{a}_i); 1 \le i \le M\})$ is called the *reference element*. For the simplicity of notation, we will sometime omit the index T in P_T and $a_{i,T}$ and write the element as $(T, P, \{p(a_i); 1 \le i \le M\})$ later.

Choice of the reference element. Throughout this paper, we consider the reference element $\hat{T}$ to be the d-simplex. In addition, the space $\hat{P}$ is chosen to be the space of *quadratic polynomials* over $\hat{T}$. The dimension of $\hat{P}$ is $\dim(\hat{P}) = \frac{d(d-1)}{2} + d$. The degree of freedom is chosen as follows: (i) $\hat{a}_i$ is the vector with the i-th entry being 1 and the other entries being 0, and (ii) $\hat{a}_{ij} = \frac{1}{2}(\hat{a}_i + \hat{a}_j)$ is the mid point of $\hat{a}_i$ and $\hat{a}_j$ for $1 \le i, j \le d$.

Triangularization of a domain. With the reference element specified, one can generate a family of finite elements that are isoparametric-equivalent to the reference element and form a partition of a domain of interest. If a partition is not possible, one may choose to partition the domain approximately. We elaborate on the requirement on the partition.

If a domain is a polygon, one can define a triangularization based on affine-equivalent elements only. However, when the domain U is curved with a smooth boundary, it is not possible to partition $\bar{U}$ into triangles. Indeed, if affine family with a mesh size $\max_{T \in \mathcal{T}_n} \operatorname{diam}(T) = O(2^{-n})$ is used to approximately cover the space U, that is, only straight triangle is in use, then the error rate of the finite element method $\|u_n - u\|_{H^1(U)}$ is known to be at most $O(2^{-3/2n})$, even with quadratic basis functions. See [3, page 268] for more details. In this case, isoparametric triangularization can be used to ensure the convergence rate of $\|u_n - u\|_{H_0^1(U)} = O(2^{-2n})$, when a and f are deterministic functions with sufficient smoothness. The precise definition of an isoparametric triangularization of a domain U is given as follows. Let $\{(T, P_T, \Sigma_T) : T \in \mathcal{T}_n\}$ be a family of finite elements that are isoparametric-equivalent to the reference element $(\hat{T}, \hat{P}, \{\hat{p}(\hat{a}_i), i = 1, ..., M\})$ with $\max_{T \in \mathcal{T}_n} \operatorname{diam}(T) = O(2^{-n})$ satisfying the following requirements.

(1) $\bar{U} = \cup_{T \in \mathcal{T}_n} T$;

(2) For any $T \in \mathcal{T}_n$, the corresponding degree of freedom a_i, a_{ij} is either inside the domain U or on the boundary ∂U for all $1 \le i, j \le M$.

(3) For $T, T' \in \mathcal{T}_n$, $T \neq T'$, $int(T) \cap int(T') = \emptyset$, where $int(T)$ denote the interior of the triangle T;

(4) If $T \neq T'$ but $T \cap T' \neq \emptyset$, then $T \cap T'$ is either a point or a common edge of T and T'.

Here, the edges and vertices of an isoparametric element is the image of the corresponding isoparametric mapping of the edges and vertices of the reference element, respectively.

Remark 1. Among the requirement (1)-(4), (2)-(4) are standard assumptions and can be satisfied by many applications of interest. Assumption (1) requires that the domain U is covered exactly by the isoparametric elements, which can be satisfied when the boundary ∂U is piecewise quadratic. It is also possible, but may require more tedious analysis, to extend our result to the case where ∂U is smooth but not piecewise quadratic. We omit the details for the simplicity of the presentation. For the analysis of such a case when a and f are deterministic, see [3, Chapter VI].

Regular isoparametric family. Define

$$h_T = \text{diam}(T) \text{ and } \rho_T = \sup\{\text{diam}(S) : S \text{ is a ball in } R^d \text{ and } S \subset T\}$$

for each $T \in \mathcal{T}_n$. The the isoparametric family $\{(T, P, \Sigma), T \in \mathcal{T}_n\}$ is called *regular* if it satisfies the following two conditions.

(1) There exists a constant $\sigma > 0$ such that for all n and all $T \subset \mathcal{T}_n$,

$$\rho_T \geq \sigma h_T.$$

(2) For each $T \in \mathcal{T}_n$, let $\tilde{a}_{ij} = \frac{1}{2}(a_i + a_j)$ for all $1 \leq i, j \leq d$ and a_i, a_j being the vertices of T. We assume

$$\|a_{ij} - \tilde{a}_{ij}\| = O(2^{-2n})$$

uniformly for all $T \in \mathcal{T}_n$.

Throughout the paper, we will only consider regular isoparametric family. In addition, we will assume that the inner elements are affine elements and only the boundary elements are other isoparametric elements. That is, for a finite element (T, P, Σ) that is not on the boundary of the domain, T is a triangle (tetrahedron) for $d = 2$ ($d = 3$).

The function space V_n. Based on a regular isoparametric family $\{(T, P, \Sigma), T \in \mathcal{T}_n\}$ defined above, we are able to state the definition of the space V_n as below,

$$V_n = \Big\{ v \in C(\bar{U}) : v|_T \in P_T \text{ for each } T \in \mathcal{T}_n \text{ and } v|_{\partial D} = 0 \Big\}. \tag{78}$$

Isoparametric numerical integral. For isoparametric elements, the numerical integral is done by first performing quadrature approximation over the reference element, and then transforming it to the isoparametric family. We first describe the integral approximation over the reference element $\hat{T} = \{(x_1, .., x_d) : x_i \geq 0, \sum_{i=1}^d x_i \leq 1; 1 \leq i \leq d\}$. Typically, a quadrature scheme for numerical integration is described in the following form. For a function $\hat{\phi} : \hat{T} \to \mathbb{R}$, the integral $\int_{\hat{T}} \hat{\phi}(\hat{x}) d\hat{x}$ is approximated by $\sum_{l=1}^M \hat{w}_l \hat{\phi}(\hat{b}_l)$ for some weights $\hat{w}_l > 0$, points $\hat{b}_l$, $l = 1, ..., M$, and a positive integer M. In order to control the numerical error of the finite element method, we assumes that $\hat{w}_l$'s and $\hat{b}_l$'s are exact for quadratic functions. That is, if $\hat{\phi}$ is a quadratic function over $\hat{T}$, then $\int_{\hat{T}} \hat{\phi}(\hat{x}) d\hat{x} = \sum_{l=1}^M \hat{w}_l \hat{\phi}(\hat{b}_l)$. The choice of such a quadrature scheme is not unique. For example, a popular choice for $d = 2$ is $M = 3$, $b_1 = (0.5, 0)$, $b_2 = (0, 0.5)$, $b_3 = (0.5, 0.5)$, $w_1 = w_2 = w_3 = \frac{1}{6}$. We proceed to the numerical integration over an isoparametric element T with an isoparametric mapping F_T. The standard approximation for the integral in the form $\int_T \phi(x) dx$ is based on the change of variable, where the weights are defined as $w_{l,T} = \hat{w}_l J(F_T)(\hat{b}_{l,T})$, $b_{l,T} = F_T(\hat{b}_l)$, and $J(F_T)$ denotes the Jacobian of the mapping F_T.

Acknowledgements

The authors would like to thank Dr. Hehu Xie for the helpful discussions. Xiaoou Li is partially supported by National Science Foundation (DMS-1712657). Jingchen Liu is partially supported by National Science Foundation (SES-1323977, IIS-1633360, SES-1826540), and Army Research Office (W911NF-15-1-0159).

References

[1] Charrier, J. (2012). Strong and weak error estimates for elliptic partial differential equations with random coefficients. *SIAM Journal on Numerical Analysis* **50,** 216–246.

[2] Charrier, J., Scheichl, R. and Teckentrup, A. L. (2013). Finite element error analysis of elliptic pdes with random coefficients and its application to multilevel monte carlo methods. *SIAM Journal on Numerical Analysis* **51,** 322–352.

[3] Ciarlet, P. (1991). Basic error estimates for elliptic problems. In *Handbook of numerical analysis*. vol. 2 of *Finite Element Methods (Part 1)*. Elsevier pp. 17–351.

[4] Cliffe, K., Giles, M., Scheichl, R. and Teckentrup, A. L. (2011). Multilevel monte carlo methods and applications to elliptic pdes with random coefficients. *Computing and Visualization in Science* **14,** 3–15.

[5] De Marsily, G., Delay, F., Gonçalvès, J., Renard, P., Teles, V. and Violette, S. (2005). Dealing with spatial heterogeneity. *Hydrogeology Journal* **13,** 161–183.

[6] Delhomme, J. (1979). Spatial variability and uncertainty in groundwater flow parameters: A geostatistical approach. *Water Resources Research* **15,** 269–280.

[7] Evans, L. C. (1998). *Partial differential equations*. Providence, Rhode Land: American Mathematical Society.

[8] Giles, M. B. (2008). Multilevel monte carlo path simulation. *Operations Research* **56,** 607–617.

[9] Graham, I. G., Kuo, F. Y., Nuyens, D., Scheichl, R. and Sloan, I. H. (2011). Quasi-monte carlo methods for elliptic pdes with random coefficients and applications. *Journal of Computational Physics* **230,** 3668–3694.

[10] Knabner, P. and Angermann, L. (2003). *Numerical methods for elliptic and parabolic partial differential equations*. Springer.

[11] Ostoja-Starzewski, M. (2007). *Microstructural randomness and scaling in mechanics of materials*. CRC Press.

[12] Rhee, C.-h. and Glynn, P. W. (2012). A new approach to unbiased estimation for SDE's. In *Proceedings of the Winter Simulation Conference*. Winter Simulation Conference. p. 17.

[13] Rhee, C.-h. and Glynn, P. W. (2013). Unbiased estimation with square root convergence for SDE models.

[14] Sobczyk, K. and Kirkner, D. J. (2001). *Stochastic modeling of microstructures*. BIRKHÄUSER.

[15] Teckentrup, A., Scheichl, R., Giles, M. and Ullmann, E. (2013). Further analysis of multilevel monte carlo methods for elliptic pdes with random coefficients. *Numerische Mathematik* **125,** 569–600.