Adv. Appl. Prob. **50,** 1007–1031 (2018)
*doi:*10.1017/apr.2018.49

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

XIAOOU LI,* *University of Minnesota*
JINGCHEN LIU ** AND
SHUN XU,** *Columbia University*

Abstract

Partial differential equations are powerful tools for used to characterizing various physical 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 a multilevel Monte Carlo method (Giles (2008)) and a randomization scheme proposed by Rhee and Glynn (2012), (2013). Furthermore, to obtain an estimator with both finite variance and finite expected computational cost, we employ higher-order approximations.

Keywords: Unbiased sampling; partial differential equations with random coefficients; Monte Carlo method

2010 Mathematics Subject Classification: Primary 65C05
Secondary 35R60; 82B80

1. Introduction

Elliptic partial differential equations are classic equations used to describe various static physical systems. In practice such systems are not usually described precisely; for instance, imprecision could be due to microscopic heterogeneity or 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 $u\colon U \to R$:

$$-\nabla \cdot (a(x)\nabla u(x)) = f(x) \quad \text{for } x \in U. \tag{1.1}$$

Here $f(x)$ 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, we solve u subject to certain boundary conditions that are necessary for the uniqueness of the solution. This will be discussed in the sequel. 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 $a(x, \omega)$, $f(x, \omega)$,

Received 17 June 2016; revision received 29 November 2017.
* Postal address: School of Statistics, University of Minnesota, 224 Church Street SE, Minneapolis, MN 55455, USA. Email address: xiaoou@stat.columbia.edu
** Postal address: Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, 10027, USA.

and $u(x, \omega)$, and write $a(x)$, $f(x)$, and $u(x)$ instead that satisfy (1.1) and boundary conditions almost surely. Throughout this paper, we consider $d \leq 3$ to be sufficient for most physical 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 simulation. 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

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

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 physical systems; see, for instance, [5] and [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 such an estimator is finite. The analysis strategy is a combination of a multilevel Monte Carlo method and a randomization scheme. The multilevel Monte Carlo method is a recent advancement in the simulation and approximation of continuous processes [4], [8], [9]. The randomization scheme was developed by Rhee and Glynn [12], [13]. Under the current setting, a direct application of these two methods leads to either an estimator with infinite variance or an 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 the mesh size [2], [15]. The technical contribution of this paper is to employ the finite element method with quadratic isoparametric elements to solve partial differential equations (PDEs) under certain smoothness conditions on $a(x)$ and $f(x)$, and to perform careful analysis of the numerical solver for (1.1).

1.1. Physics applications

Equation (1.1) 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 specific physical meanings. When the elliptic PDE is used to describe the steady-state distribution of heat (as temperature), $u(x)$ represents the temperature at x and the coefficient $a(x)$ represents 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, $u(x)$ is the hydraulic head (water level elevation) and $a(x)$ is the hydraulic conductivity (or permeability). The physical laws for the above three problems to derive the same type of elliptic PDE are respectively called Fourier's law, Gauss's law, and Darcy's law. In classical continuum mechanics, (1.1) 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 physical interpretation in the context of a material deformation application. In the model of classical continuum mechanics the domain U is a smooth manifold denoting the physical

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

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; cf. [11] and [14]). 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 account for the heterogeneities and 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. In Section 3 we present the construction of the unbiased Monte Carlo estimator for $w_{\mathcal{Q}}$ and a rigorous complexity analysis. In Section 4 we include numerical implementations. Technical proofs and a detailed definition of finite element methods are included in the appendices.

2. Preliminary analysis

Throughout this paper, we consider (1.1) living on a bounded domain $U \subset R^d$ with a twice differentiable boundary denoted by ∂U. To ensure the uniqueness of the solution, we consider the Dirichlet boundary condition

$$u(x) = 0 \quad \text{for } x \in \partial U. \tag{2.1}$$

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

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

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 a Monte Carlo simulation. 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

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

by a Monte Carlo simulation. The notation $\bar{U}$ is used to denote the closure of domain U and $C(\bar{U})$ is used to denote 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

$$\mathbb{E}(\hat{Z} - w_{\mathcal{Q}})^2 = \mathrm{var}(\hat{Z}) + \{\mathbb{E}[\hat{Z}] - w_{\mathcal{Q}}\}^2$$

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 a multilevel Monte Carlo method. To start with, we present the basics of the multilevel Monte Carlo method and the randomization scheme.

2.1. Multilevel Monte Carlo

Consider a biased estimator of $w_{\mathcal{Q}}$ denoted by Z_n. In the current context, Z_n is the estimator corresponding to some numerical solution based on a 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

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

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, $\mathbb{E}(Z_n) \to w_{\mathcal{Q}}$. The multilevel Monte Carlo method is based on the following telescope sum:

$$w_{\mathcal{Q}} = \mathbb{E}[Z_0] + \sum_{i=0}^{\infty} \mathbb{E}[Z_{i+1} - Z_i]. \tag{2.2}$$

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 to ∞. The coupling is commonly done by constructing Z_{i+1} and Z_i with the same sample path ω (that is, the same $a(\cdot, \omega)$ and $f(\cdot, \omega)$). The specific choices of our Y_i and Z_i in this paper are given in Section 3.2. Let

$$\Delta_i = \mathbb{E}[Z_{i+1} - Z_i]$$

be estimated by

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

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

$$\hat{Z} = \sum_{i=1}^{I} \hat{\Delta}_i, \tag{2.3}$$

where I is a large integer truncating the infinite sum (2.2).

2.2. An unbiased estimator via a randomization scheme

In the construction of the multilevel Monte Carlo estimator (2.3), 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 was originally proposed by Rhee and Glynn [12], [13].

Let N be a positive, integer-valued random variable that is independent of $\{Z_i\}_{i=1,2,\ldots}$. 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:

$$w_{\mathcal{Q}} = \sum_{i=1}^{\infty} \mathbb{E}[Z_n - Z_{n-1}] = \sum_{n=1}^{\infty} \frac{\mathbb{E}[Z_n - Z_{n-1}; N = n]}{p_n} = \mathbb{E}\left[\frac{Z_N - Z_{N-1}}{p_N}\right].$$

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

$$\tilde{Z} = \frac{Z_N - Z_{N-1}}{p_N}. \tag{2.4}$$

Let $\tilde{Z}_i$, $i = 1, \ldots, M$ 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 $\mathrm{var}(\tilde{Z})/M$ if finite.

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

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

$$\mathbb{E}[\tilde{Z}^2] = \mathbb{E}\left[\frac{(Z_N - Z_{N-1})^2}{p_N^2}\right] = \sum_{n=1}^{\infty} \frac{\mathbb{E}(Z_n - Z_{n-1})^2}{p_n}. \tag{2.5}$$

In order to have a finite second moment, the sequence $\mathbb{E}(Z_n - Z_{n-1})^2/p_n$ needs to tend to 0 no slower than n^{-1}. Thus, we would like to choose the random variable N such that

$$p_n > n\mathbb{E}(Z_n - Z_{n-1})^2 \quad \text{for all sufficiently large } n. \tag{2.6}$$

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 (2.6), we have

$$n^{-1} > p_n > n\mathbb{E}(Z_n - Z_{n-1})^2. \tag{2.7}$$

Note that we have not yet specified a discretization method; thus, (2.7) can typically be met by appropriately indexing the mesh size. For instance, in the context of solving a PDE numerically, we 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 to allow quite some flexibility in choosing p_n. Thus, constraint (2.7) 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

$$C = \sum_{i=1}^{n} p_n c_n. \tag{2.8}$$

In order to have C finite for sufficiently large n,

$$p_n < n^{-1} c_n^{-1}. \tag{2.9}$$

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 (2.7) and (2.9), which suggests that

$$\mathbb{E}(Z_n - Z_{n-1})^2 < n^{-2} c_n^{-1}. \tag{2.10}$$

That is, we must be able to construct a coupling between Z_n and Z_{n-1} such that (2.10) holds. In Section 3 we provide a 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 discussions. Let $U \subset R^d$ be a bounded open set. We define the following spaces of functions:

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

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

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

$$L^p(U) = \left\{ u \colon U \to R \;\middle|\; \int_U |u(x)|^p \, \mathrm{d}x < \infty \right\},$$
$$L^p_{\mathrm{loc}}(U) = \{ u \colon U \to R \mid u \in L^p(K) \text{ for any compact subset } K \subset U \},$$
$$C^\infty_{\mathrm{c}}(U) = \{ u \colon U \to R \mid u \text{ is infinitely differentiable with a compact support that is a subset of } U \}.$$

Definition 2.1. For $u, w \in L^1_{\mathrm{loc}}(U)$ and a multiple index α, we say that w is the α-weak derivative of u, and write $D^\alpha u = w$ if $\int_U u D^\alpha \phi \, \mathrm{d}x = (-1)^{|\alpha|} \int_U w\phi \, \mathrm{d}x$ for all $\phi \in C^\infty_{\mathrm{c}}(U)$, where $D^\alpha \phi$ in the above expression denote the usual α-partial derivative of ϕ.

If $u \in C^k(\bar{U})$ and $|\alpha| \le k$ 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 ambiguity.

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

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

and

$$\|u\|_{L^p(U)} = \left(\int_U |u|^p \, \mathrm{d}x \right)^{1/p}.$$

We proceed to the definition of the Sobolev spaces $H^k(U)$ and $H^k_{\mathrm{loc}}(U)$:

$$H^k(U) = \{ u \colon U \to R \mid D^\alpha u \in L^2(U) \text{ for all multiple index } \alpha \text{ such that } |\alpha| \le k \},$$
$$H^k_{\mathrm{loc}}(U) = \{ u \colon U \to R \mid u|_V \in H^k(V) \text{ for all } V \subsetneq U \}.$$

For $u \in H^k(U)$, the norm $\|u\|_{H^k(U)}$ and the semi-norm $|u|_{H^k(U)}$ are defined as

$$\|u\|_{H^k(U)} = \left(\sum_{|\alpha| \le k} \|D^\alpha u\|^2_{L^2(U)} \right)^{1/2}$$

and

$$|u|_{H^k(U)} = \left(\sum_{|\alpha| = k} \|D^\alpha u\|^2_{L^2(U)} \right)^{1/2}. \tag{2.11}$$

We define the space $H^1_0(U)$ as

$$H^1_0(U) = \{ u \in H^1(U) \colon u(x) = 0 \text{ for } x \in \partial U \}.$$

On the space $H^1_0(U)$, the norm $\| \cdot \|_{H^1(U)}$ and the semi-norm $| \cdot |_{H^1(U)}$ are equivalent.

2.4. Finite element method for PDEs

We briefly describe the finite element method for PDEs. The weak solution $u \in H^1_0(U)$ to (1.1) under the Dirichlet boundary condition (2.1) is defined through the variational form

$$b(u, v) = L(v) \quad \text{for all } v \in H^1_0(U), \tag{2.12}$$

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

where we define the bilinear and linear forms as

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

and use the dot notation for 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 (1.1). 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 $V_n = \mathrm{span}\{\phi_1, \ldots, \phi_{L_n}\}$, where L_n is the dimension of V_n. The approximate solution $u_n \in V_n$ is defined through the set of equations

$$b(u_n, v) = L(v) \quad \text{for all } v \in V_n. \tag{2.13}$$

Both sides of the equations are linear in v. Then (2.13) is equivalent to $b(u_n, \phi_i) = L(\phi_i)$ for $i = 1, \ldots, L_n$. We further write $u_n = \sum_{i=1}^{L_n} d_i\phi_i$ as a linear combination of the basis functions. Then (2.13) is equivalent to solving the linear equations

$$\sum_{j=1}^{L_n} d_j b(\phi_j, \phi_i) = L(\phi_i) \quad \text{for } i = 1, \ldots, L_n. \tag{2.14}$$

The basis functions $\phi_1, \ldots, \phi_{L_n}$ are often chosen such that (2.14) is a sparse linear system. That is, the order of the number of nonzero $b(\phi_j, \phi_i)$ 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 [10, Chapter 5] for more details.

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

In this paper our choice of 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 Subsection 3.1.2, and the precise definition will be delayed to Appendix C. The advantage of using quadratic isoparametric elements over the linear elements is twofold. First, the quadratic approximation provides a 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 reader to [3] and the references therein.

3. Main results

In this section we present the construction of $\tilde{Z}$ and its complexity analysis. We use a 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 a usual finite element method with linear basis functions, with which we show that (2.7) and (2.9) cannot be satisfied simultaneously. Thus, $\tilde{Z}$ either has infinite variance or has an 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.

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

3.1. Finite element method

3.1.1. *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 the 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 are provided in Appendix C.

Once a set of basis functions has been chosen, the coefficients d_i are solved according to the linear equations (2.14) 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 note 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 (2.10) for linear basis functions. The dimension of V_n is of the 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 (2.11).

According to Proposition 4.2 of [2], under the conditions that $\mathbb{E}[1/\min_{x\in U} a^p(x)] < \infty$, $\mathbb{E}[\|a\|^p_{C^1(\bar{U})}] < \infty$, 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. Condition (2.10) 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 (2.10) for higher dimensions, we need a faster convergence rate of the PDE numerical solver.

3.1.2. *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 (see, 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 (C.1) below with a mesh size of $O(2^{-n})$. 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 to 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 $V_n = \{v \in C(\bar{U})\colon v|_{\partial U} = 0 \text{ and } v|_T \in P_T \text{ for } T \in \mathcal{T}_n\}$.

Step 1 is usually done by certain mesh generating algorithms 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 3.1. Let $U = B(0,1) = \{(x,y)\colon x^2 + y^2 < 1\}$. 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 diagram of Figure 1 shows a possible choice of the partition when $n = 1$. The right diagram 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. the black region in Figure 1)

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

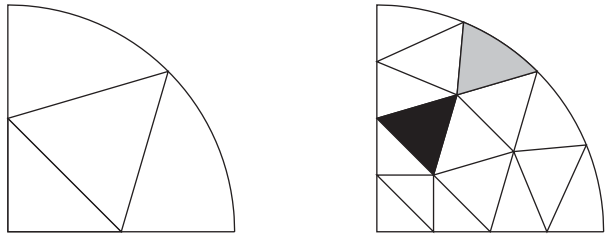

FIGURE 1: Isoparametric triangularization for Example 3.1 when $n = 1$ (*left*) and $n = 2$ (*right*).

then T is a triangle; if an edge of T lies on the boundary (e.g. the 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 is not necessarily, the space of quadratic functions.

With the finite-dimensional space V_n constructed, we obtain an approximate solution u_n by solving (2.13) with V_n.

3.1.3. *Isoparametric numerical integration.* The numerical solution u_n in (2.13) requires the evaluation of the integrals

$$b(w, v) = \sum_{T \in \mathcal{T}_n} \int_T a(x) \nabla w(x) \cdot \nabla v(x) \, \mathrm{d}x \quad \text{and} \quad L(v) = \sum_{T \in \mathcal{T}_n} \int_T f(x) v(x) \, \mathrm{d}x.$$

For the evaluation of these integrals, we apply a quadrature approximation which approximates the integral in the form of $\int_T \phi(x) \, \mathrm{d}x$ by $\sum_{l=1}^{M} w_{l,T} \phi(b_{l,T})$ for some prespecified weights $w_{l,T}$ and points $b_{l,T}$ for a positive integer M and $1 \leq l \leq M$. The precise choices of $w_{l,T}$ and $b_{l,T}$ are given in Appendix C. We point out that the choices of $w_{l,T}$ and $b_{l,T}$ depend on the isoparametric triangularization only, and are independent of 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 the linear form $L(v)$ by their numerical approximations, denoted by $\tilde{b}(w, v)$ and $\tilde{L}(v)$, respectively. Based on the numerical integration, we define $\tilde{u}_n$ such that

$$\tilde{b}_n(\tilde{u}_n, v) = \tilde{L}(v) \quad \text{for all } v \in V_n. \tag{3.1}$$

3.1.4. *Error analysis for the 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 (2.12) and $\tilde{u}_n$ is the solution to (3.1).

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 $\mathbb{E}[\|a\|_{C^2(\bar{U})}^p] < \infty$ for all $p \in (0, \infty)$.

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

(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 nonnegative constants p' and κ_q such that, for all $w_1, w_2 \in H^1_0(U)$,

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

With 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 a finite expected computational time.

We start with the existence and the uniqueness of the solution. Note that $a(x)$ is bounded below by positive random variables $a_{\min}$ and above by $a_{\max}$. According to [2, Lemma 2.1], (2.12) has a unique solution $u \in H^1_0(U)$ almost surely satisfying

$$\|u\|_{H^1(U)} \le \kappa \frac{\|f\|_{L^2(U)}}{a_{\min}}. \tag{3.2}$$

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

Theorem 3.1. *Let $\tilde{u}_n$ be the solution to (3.1). For $\dim(U) \le 3$ with a three-time differentiable boundary ∂U, if $a(x) \in C^2(\bar{U})$ and $f(x) \in H^2(U)$, then we have*

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

Proof. See Appendix A. □

3.2. Construction of the unbiased estimator

In this section we apply the results obtained in Subsection 3.1.2 to construct an unbiased estimator with both finite variance and a finite expected computational cost through (2.4). We start with providing an upper bound of $\mathbb{E}[\mathcal{Q}(u) - \mathcal{Q}(\tilde{u}_n)]^2$.

Proposition 3.1. *Under assumptions (A1)–(A4), we have*

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

where u is the solution to (2.12), $\tilde{u}_n$ is the solution to (3.1), and κ_q is the Lipschitz constant that appeared in (A4).

The proof is a direct application of (3.2), Theorem 3.1, and (A4) and is therefore omitted. We proceed to the construction of the unbiased estimator $\tilde{Z}$ via (2.4). Choose $\mathbb{P}(N = n) = p_n \propto 2^{-(4+d)n/2}$. For each n, let $\tilde{u}_{n-1}$ and $\tilde{u}_n$ be defined as in (3.1) with respect to the same a and f. Note 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 3.1, we have $\mathbb{E}(Z_n - Z_{n-1})^2 \le 2\mathbb{E}[\mathcal{Q}(\tilde{u}_n) - \mathcal{Q}(u)]^2 + 2\mathbb{E}[\mathcal{Q}(\tilde{u}^{(2)}_{n-1}) - \mathcal{Q}(u)]^2 = O(2^{-4n})$. According to (2.5), for $d = \dim(U) \le 3$, we have $\mathbb{E}[\tilde{Z}^2] \le \sum_{n=1}^{\infty} 2^{-4n}/2^{-(4+d)n/2} < \infty$. Furthermore, (3.1) requires solving $O(2^{dn})$ sparse linear equations. The computational cost of obtaining u_n is $O(n2^{dn})$. According to (2.8), the expected cost of generating a single copy of $\tilde{Z}$ is

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

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

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

3.2.1. *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 finite element method with a grid size of 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, we can still achieve a similar error bound for the resulting numerical solver. In Example 3.2, we show a situation where we can construct such an approximation for $a(\cdot)$.

Example 3.2. Assume that $\log a(x) = g(x)$ and that $g(x)$ has the following expansion. For all $x \in U$, $g(x) = \sum_{l=0}^{\infty} \lambda_l W_l \phi_l(x)$, where $W_1, W_2, \ldots$ are independent and identically distributed (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 at 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 a 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 3.2. *Define*

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

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

- $\max_n \mathbb{E} \min_{x \in U} (a_n(x), 1)^{-p} < \infty$ *for all* $p > 0$.
- $\max_n \mathbb{E} \|a_n\|^p_{C^2(\bar{U})} < \infty$ *for all* $p > 0$.
- *There exists a constant* $\delta > 0$ *such that* $\max_n \mathbb{E} \tilde{W}_n^{2+\delta} < \infty$.
- *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 (3.1) 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 (2.4) has a finite variance and a finite computational cost.*

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

4. Simulation study

4.1. An illustrating example

We start with a simple example for which a closed-form solution is available and, therefore, we are able to check the accuracy of the simulation. Let $U = B(0, 1)$, $f(x) = 2e^{W_1 + W_2 x_1 + W_3 x_2} \times (2 + W_2 x_1 + W_3 x_2)$ and $a(x) = e^{W_2 x_1 + W_3 x_2}$, where W_1, W_2, and W_3 are i.i.d. standard normal

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

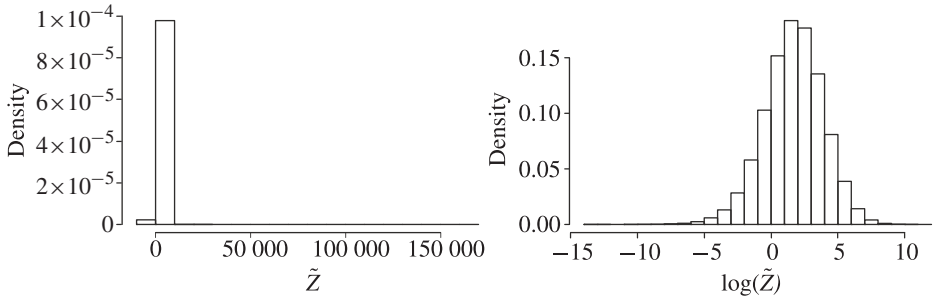

FIGURE 2: Histograms of Monte Carlo samples of $\tilde{Z}$ and $\log \tilde{Z}$ that are defined in Subsection 4.1.

random variables. In this example, the solution to (1.1) is

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

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

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

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 (2.4) is

$$\tilde{Z} = \frac{Z_N - Z_{N-1}}{p_N}. \tag{4.1}$$

We perform Monte Carlo simulation with $M = 300\,000$ replications. The averaged estimator is 46.5572 with the standard deviation 0.8212. In Figure 2 we present the histograms of samples of $\tilde{Z}$ and $\log \tilde{Z}$.

4.2. Log-normal random field

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

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

where $W_1, W_2, \ldots$ are i.i.d. standard normal random variables. It is not hard to verify that $a(x_1, x_2)$ satisfies assumptions (A1) and (A2). We further approximate the field a by

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

and compute the finite element solution based on this approximation. We let $Z_n = \mathcal{Q}(\bar{u}_n)$ as discussed in Theorem 3.2 and take the same estimator (4.1) and functional $\mathcal{Q}$ as the previous example. We perform Monte Carlo simulation for M = 300 000 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. In Figure 3 we present the histogram of the Monte Carlo sample.

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

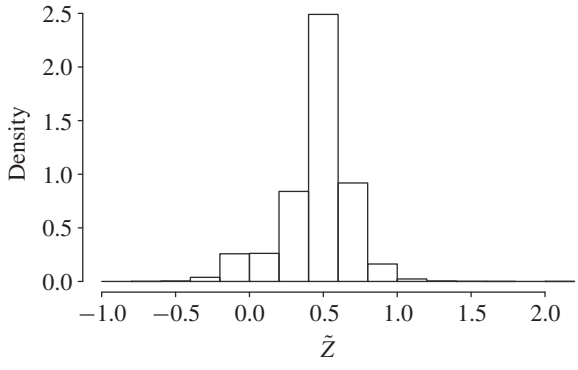

FIGURE 3: Histogram of the Monte Carlo sample of $\tilde{Z}$ when $\log a$ has a Gaussian covariance.

Appendix A. Proofs of the theorems

In this section we provide technical proofs of Theorems 3.1 and 3.2. Throughout the proofs 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-order regularity of the solution u, whose proof is given after the proofs of Theorems 3.1 and 3.2.

Proposition A.1. *For* $\dim(U) \le 3$ *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*

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

A.1. Proof of Theorem 3.1

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

Lemma A.1. *It holds that*

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

Combining the above display with Proposition A.1 for $k = 2$, we have

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

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

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

Thanks to (3.2), the above display can be further bounded by

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

We complete the proof by combining the above expression with (A.1).

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

A.2. Proof of Theorem 3.2

We start with the inequality,

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

The second term on the right-hand side of the above inequality is already bounded from above by Theorem 3.1. That is,

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

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 $\tilde{b}_n$. Noting that $\bar{u}_n$ is obtained by replacing $\tilde{b}_n$ by $\bar{b}_n$ in (3.1), we have

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

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

$$(\bar{b}_n - \tilde{b}_n)(\bar{u}_n, w) = \tilde{b}_n(\tilde{u}_n - \bar{u}_n, w), \tag{A.3}$$

where we write $(\bar{b}_n - \tilde{b}_n)(v, w) = \bar{b}_n(v, w) - \tilde{b}_n(v, w)$. Setting $w = \tilde{u}_n - \bar{u}_n$ in (A.3), we arrive at

$$(\bar{b}_n - \tilde{b}_n)(\bar{u}_n, \tilde{u}_n - \bar{u}_n) = \tilde{b}_n(\tilde{u}_n - \bar{u}_n, \tilde{u}_n - \bar{u}_n). \tag{A.4}$$

According to the same arguments as those given in [3, pp. 258–260], the right-hand side of (A.4) is bounded from below by

$$\tilde{b}_n(\tilde{u}_n - \bar{u}_n, \tilde{u}_n - \bar{u}_n) \ge \varepsilon a_{\min}\|\tilde{u}_n - \bar{u}_n\|^2_{H^1(U)}. \tag{A.5}$$

On the other hand, we have

$$|(\bar{b}_n - \tilde{b}_n)(\bar{u}_n, \tilde{u}_n - \bar{u}_n)| \le \|a - a_n\|_{C(\bar{U})}\|\bar{u}_n - \tilde{u}_n\|_{H^1(U)}\|\bar{u}_n\|_{H^1(U)}. \tag{A.6}$$

Combining (A.4), (A.5), and (A.6), we arrive at

$$\|\tilde{u}_n - \bar{u}_n\|_{H^1(U)} \le \kappa\frac{\|a - a_n\|_{C(\bar{U})}\|\bar{u}_n\|_{H^1(U)}}{a_{\min}} \le \kappa 2^{-2n}\frac{\|\bar{u}_n\|_{H^1(U)}}{a_{\min}}\tilde{W}_n. \tag{A.7}$$

Because a_n satisfies assumptions (A1) and (A2), we can apply Theorem 3.1 to the solution $\bar{u}_n$ and arrive at

$$\|\bar{u}_n\|_{H^1(U)} = O\left(\frac{\max(\|a_n\|_{C^2(\bar{U})}, 1)^{12}}{\min(\min_{x\in U}(a_n(x)), 1)^{11}}\|f\|_{H^2(U)}\right). \tag{A.8}$$

Combining (A.2), (A.7), and (A.8), we arrive at

$$\|\bar{u}_n - u\|_{H^1(\bar{U})} = O\left(\left\{\frac{\max(\|a\|_{C^2(\bar{U})}, 1)^{12}}{\min(a_{\min}, 1)^{11}} + \frac{\max(\|a_n\|_{C^2(\bar{U})}, 1)^{12}}{\min(\min_{x\in U}(a_n(x)), 1)^{11}\min(a_{\min}, 1)}\tilde{W}_n\right\} \times \|f\|_{H^2}(U)2^{-2n}\right). \tag{A.9}$$

The rest of the proof is similar to the analysis under Proposition 3.1; we omit the details.

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

A.3. Proof of Proposition A.1

Proposition A.1 is similar to Theorem 5 of [7, Chapter 6.3], but we explicitly provide the dependence of constants on a and f.

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

$$-\nabla \cdot (A\nabla w) = f \quad \text{in } U, \qquad w = 0 \quad \text{on } \partial U. \tag{A.10}$$

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

$$\xi^\top A(x)\xi \ge A_{\min}|\xi|^2 \tag{A.11}$$

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

$$\|w\|_{H^{k+1}(U)} \le \kappa_r(A,k)(\|f\|_{H^{k-1}(U)} + \|w\|_{L^2(U)}), \tag{A.12}$$

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

Let $B^0(0,r)$ denote the open ball $\{x : |x| < r\}$ and let $R_+^d = \{x \in R^d : x_d > 0\}$. 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 $0 < t < r$,

$$\|w\|_{H^{m+2}(V)} \le \kappa_{r,t,m+1}\frac{\max(\|A\|_{C^k(\bar{U})}, 1)^{(m+1)^2/2+9(m+1)/2-1}}{\min(A_{\min}, 1)^{(m+1)^2/2+7(m+1)/2}}(\|f\|_{H^m(U)} + \|w\|_{L^2(U)}), \tag{A.13}$$

where $\kappa_{r,t,m+1}$ is a constant depending only on r, t, and $m+1$. The following lemma establishes (A.13) for $m = 0$.

Lemma A.2. (Boundary H^2-regularity.) *Assume that ∂U is twice differentiable and that $A(x)$ satisfies (A.11). Assume that $A_{ij}(x) \in C^1(\bar{U})$ for all $i, j = 1, \dots, d$. Suppose further that $w \in H_0^1(U)$ is a weak solution to the elliptic PDE with the boundary condition (A.10). Then $w \in H^2(U)$ and*

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

We establish (A.13) by induction. Suppose that, for some m,

$$\|w\|_{H^{m+1}(W)} \le \kappa_{t,s,m}\frac{\max(\|A\|_{C^k(\bar{U})}, 1)^{m^2/2+9m/2-1}}{\min(A_{\min}, 1)^{m^2/2+7m/2}}(\|f\|_{H^{m-1}(U)} + \|w\|_{L^2(U)}), \tag{A.14}$$

where

$$W = B^0(0,s) \cap R_+^d \quad \text{and} \quad s = \frac{t+1}{2}.$$

Since w is a weak solution to (A.10), it satisfies the integration equation

$$\int_D \nabla w(x)^\top A(x)\nabla v(x)\,\mathrm{d}x = \int_D f(x)v(x)\,\mathrm{d}x \quad \text{for all } v \in H_0^1(U). \tag{A.15}$$

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be a multiple index with $\alpha_d = 0$ and $|\alpha| = m$. We consider the multiple weak derivative $\bar{w} = D^\alpha w$ and investigate the PDE that $\bar{w}$ satisfies. For any $\bar{v} \in C_c^\infty(W)$, where $C_c^\infty(W)$ is the space of infinitely differentiable functions that have compact support in W, we substitute $v = (-1)^{|\alpha|} D^\alpha \bar{v}$ into (A.15). With some calculations, we have

$$\int_W (\nabla \bar{w}(x))^\top A(x) \nabla \bar{v}(x) = \int_W \bar{f}(x) \bar{v}(x)\, \mathrm{d}x,$$

where

$$\bar{f} = D^\alpha f - \sum_{\beta \le \alpha,\, \beta \neq \alpha} \binom{\alpha}{\beta} [-\nabla \cdot (D^{\alpha-\beta} A \nabla D^\beta w)]. \tag{A.16}$$

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

$$-\nabla \cdot (A \nabla \bar{w}) = \bar{f} \quad \text{for } x \text{ in } W. \tag{A.17}$$

Furthermore, we have the boundary condition $\bar{w}(x) = 0$ for $x \in \partial W \cap \{x_d = 0\}$. By the induction assumption (A.14) and (A.16), we have

$$\begin{aligned} \|\bar{f}\|_{L^2(W)} &\le \|f\|_{H^m(U)} \\ &\quad + \kappa_{t,s,m} \frac{\max(\|A\|_{C^k(\bar{U})}, 1)^{m^2/2+9m/2-1}}{\min(A_{\min}, 1)^{m^2/2+7m/2}} \|A\|_{C^{m+1}(\bar{U})} (\|f\|_{H^{m-1}(U)} + \|w\|_{L^2(U)}). \end{aligned} \tag{A.18}$$

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

$$\|\bar{w}\|_{L^2(W)} \le \|w\|_{H^m(W)}. \tag{A.19}$$

Applying Lemma A.2 to $\bar{w}$ with (A.18) and (A.19), we have

$$\begin{aligned} \|D^\alpha w\|_{H^2(V)} \le \kappa_{t,s,m} \kappa \frac{\max(\|A\|_{C^1(\bar{U})}, 1)^4}{\min(A_{\min}, 1)^4} \frac{\max(\|A\|_{C^k(\bar{U})}, 1)^{m^2/2+9m/2-1}}{\min(A_{\min}, 1)^{m^2/2+7m/2}} \\ \times \|A\|_{C^{m+1}(\bar{U})} (\|f\|_{H^m(U)} + \|w\|_{L^2(U)}). \end{aligned} \tag{A.20}$$

Because α is an arbitrary multi-index such that $\alpha_d = 0$ and $|\alpha| = m$, (A.20) implies that $D^\beta w \in L^2(W)$ for any multiple index β such that $|\beta| \le m + 2$ and $\beta_d = 0, 1, 2$. We now extend this result to the multiple index β, whose last component is greater than 2. Suppose that, for all β such that $|\beta| \le m + 2$ and $\beta_d \le j$, we have

$$\|D^\beta w\|_{H^2(V)} \le \kappa_r^{(j)} (\|f\|_{H^m(U)} + \|w\|_{L^2(U)}), \tag{A.21}$$

where $\kappa_r^{(j)}$ is a constant depending on A, m, and j to be determined 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 (A.21) to develop an upper bound for $\|D^\gamma w\|_{H^2(V)}$. In particular, let $\beta = (\gamma_1, \ldots, \gamma_{d-1}, j - 1)$. According to remark (ii) after Theorem 1 of [7, Chapter 6.3], we have

$$-\nabla \cdot (A \nabla (D^\beta w)) = f^\dagger \quad \text{in } W \text{ almost everywhere}, \tag{A.22}$$

where

$$f^\dagger = D^\beta f - \sum_{\delta \le \beta,\, \delta \neq \beta} \binom{\beta}{\delta} [-\nabla \cdot (D^{\beta-\delta} A \nabla D^\delta w)]. \tag{A.23}$$

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

Note 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 (A.21), (A.22), (A.23), and the above display, we have

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

Therefore,

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

where

$$\kappa_r^{(j+1)} = \kappa_r^{(j)} \frac{\max(\|A\|_{C^{m+1}(\bar U)},1)}{\min(A_{\min},1)}. \tag{A.24}$$

The above expression provides a relationship between $\kappa_r^{(j+1)}$ and $\kappa_r^{(j)}$. According to (A.20),

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

Using (A.24) and the above initial value for the iteration, we have

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

Consequently,

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

Using induction, we complete the proof of (A.12) 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 possible relabeling of the x coordinates, there exist a function $\gamma\colon R^{d-1} \to R$ and $r > 0$ such that

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

Let $\Phi = (\Phi_1,\dots,\Phi_d)^\top\colon R^d \to R^d$ be a function such that

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

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

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

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

Some calculation shows that w^* is a weak solution to the PDE

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

where $A^*(y) = J(y)A(\Phi^{-1}(y))J^\top(y)$ and $J(y)$ is the Jacobian matrix for Φ with $J_{ij}(y) = \partial\Phi_i(x)/\partial x_j|_{x=\Phi^{-1}(y)}$, 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 that A^* is symmetric and that $A^*_{ij} \in C^k(\bar{U})$ for all $1 \le i, j \le d$. Furthermore, according to the definitions of J and Φ, all the eigenvalues of $J(y)$ are 1 and, thus, $\zeta^\top A^*(y)\xi \ge A_{\min}|J^\top(y)\xi|^2 \ge \varepsilon A_{\min}|\xi|^2$ for all $\xi \in R^d$. By substituting U, V, A, and f with U^*, V^*, A^*, and f^* in (A.13), we have

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

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

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

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

Lemma A.3. (Higher-order interior regularity.) *Under the setting of Lemma A.2, we assume that ∂U is C^{k+1}, $A_{ij}(x) \in C^k(U)$ for all $i, j = 1, \ldots, d$, and $f \in H^{k-1}(U)$, and that $w \in H^1(U)$ is one of the weak solutions to PDE (A.10) without a boundary condition. Then $w \in H^{k+1}_{\mathrm{loc}}(U)$. For each open set $V \subsetneqq U$,*

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

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

This completes the proof of Proposition A.1.

Appendix B. Proof of the supporting lemmas

In this section we provide the proofs for the lemmas that are necessary for the proof of Proposition A.1. We start with a useful lemma showing that $w \in H^2_{\mathrm{loc}}(U)$, which will be used in the proof of Lemma A.2

Lemma B.1. (Interior H^2-regularity.) *Under the setting of Lemma A.2, we further assume that $A_{ij}(x) \in C^1(\bar{U})$ for all $i, j = 1, \ldots, d$, and $f \in L^2(U)$, and that $w \in H^1(U)$ is one of the weak solutions to the PDE (A.10) without a boundary condition. Then $w \in H^2_{\mathrm{loc}}(U)$. For each open subset $V \subsetneqq U$, there exists κ depending on V such that*

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

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

Proof. Let h be a real number whose absolute value is sufficiently small. We define the difference quotient operator as

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

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

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

$$\int_V |U_k^h \nabla w|^2 \,\mathrm{d}x \tag{B.1}$$

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

We derive a bound of (B.1) by substituting an appropriate v into (A.15). Let W be an open set such that $V \subsetneqq W \subsetneqq U$. We select a smooth function ζ such that

$$\zeta = 1 \quad \text{on } V, \qquad \zeta = 0 \quad \text{on } W^{\mathrm{c}}, \qquad 0 \le \zeta \le 1.$$

We substitute

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

into (A.15), to obtain

$$-\int_D \nabla w^\top A \nabla [D_k^{-h}(\zeta^2 D_k^h w)] \,\mathrm{d}x = -\int_D f D_k^{-h}(\zeta^2 D_k^h w) \,\mathrm{d}x. \tag{B.2}$$

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

$$\int_D w_1 D_k^{-h} w_2 \,\mathrm{d}x = -\int_D D_k^h w_1 w_2 \,\mathrm{d}x, \qquad D_k^h(w_1 w_2) = w_1^h D_k^h w_2 + w_2 D_k^h w_1,$$

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 formulae to the left-hand side of (B.2), we have

$$\begin{aligned}
&-\int_D \nabla w^\top A \nabla [D_k^{-h}(\zeta D_k^h w)] \,\mathrm{d}x \\
&\quad = \int_D D_k^h (\nabla w^\top A) \nabla (\zeta^2 D_k^h w) \,\mathrm{d}x \\
&\quad = \int_D D_k^h (\nabla w^\top) A^h \nabla (\zeta^2 D_k^h w) + \nabla w^\top D_k^h A \nabla (\zeta^2 D_k^h w) \,\mathrm{d}x \\
&\quad = \underbrace{\int_D \zeta^2 D_k^h \nabla w^\top A^h D_k^h \nabla w \,\mathrm{d}x}_{J_1} \\
&\qquad + \underbrace{\int_D 2\zeta (D_k^h \nabla w^\top A^h \nabla \zeta) D_k^h w + 2\zeta (\nabla w^\top D_k^h A \nabla \zeta) D_k^h w + \zeta^2 \nabla w^\top D_k^h A D_k^h \nabla w \,\mathrm{d}x}_{J_2}.
\end{aligned}$$

Here J_1 in the above expression has a lower bound of

$$J_1 \ge A_{\min} \int_D \zeta^2 |D_k^h \nabla w|^2 \,\mathrm{d}x$$

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

$$|J_2| \le \kappa \|A\|_{C^1(\bar{U})} \left(\int_D \zeta |D_k^h \nabla w| |D_k^h w| + \zeta |\nabla w| |D_k^h w| + \zeta |\nabla w| |D_k^h \nabla w| \,\mathrm{d}x \right). \tag{B.3}$$

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

Expression (B.3) can be further bounded by

$$|J_2| \le \frac{A_{\min}}{2} \int_D \zeta^2 |D_k^h \nabla w|^2 \,\mathrm{d}x + \kappa \|A\|_{C^1(\bar{U})} \left(1 + \frac{\|A\|_{C^1(\bar{U})}}{A_{\min}}\right) \int_W |\nabla w|^2 + |D_k^h w|^2 \,\mathrm{d}x, \quad \text{(B.4)}$$

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

$$\int_W |D_k^h w|^2 \,\mathrm{d}x \le \kappa \int_W |\nabla w|^2 \,\mathrm{d}x. \quad \text{(B.5)}$$

Therefore, (B.4) is bounded above by

$$|J_2| \le \frac{A_{\min}}{2} \int_D \zeta^2 |D_k^h \nabla w|^2 \,\mathrm{d}x + \kappa^2 \|A\|_{C^1(\bar{U})} \left(1 + \frac{\|A\|_{C^1(\bar{U})}}{A_{\min}}\right) \int_W |\nabla w|^2 \,\mathrm{d}x. \quad \text{(B.6)}$$

Combining (B.3) and (B.6), the left-hand side of (B.2) becomes

$$\begin{aligned} &- \int_D \nabla w^\top A \nabla [D_k^{-h} (\zeta^2 D_k^h w)] \,\mathrm{d}x \\ &\quad = J_1 + J_2 \\ &\quad \ge J_1 - |J_2| \\ &\quad \ge \frac{A_{\min}}{2} \int_D \zeta^2 |D_k^h \nabla w|^2 \,\mathrm{d}x - \kappa^2 \|A\|_{C^1(\bar{U})} \left(1 + \frac{\|A\|_{C^1(\bar{U})}}{A_{\min}}\right) \int_W |\nabla w|^2 \,\mathrm{d}x. \end{aligned} \quad \text{(B.7)}$$

We proceed to an upper bound of the right-hand side of (B.2). According to (B.5), we have

$$\begin{aligned} \int_D |D_k^{-h} (\zeta^2 D_k^h w)|^2 \,\mathrm{d}x &\le \kappa \int_D |\nabla (\zeta^2 D_k^h w)|^2 \,\mathrm{d}x \\ &\le \kappa \int_W 4 |D_k^h w|^2 |\nabla \zeta|^2 \zeta^2 + \zeta^2 |D_k^h \nabla w|^2 \,\mathrm{d}x \\ &\le \kappa^3 \int_W |\nabla w|^2 + \zeta^2 |D_k^h \nabla w|^2 \,\mathrm{d}x. \end{aligned} \quad \text{(B.8)}$$

Applying Cauchy's inequality to the right-hand side of (B.2), we have

$$\begin{aligned} - \int_D f D_k^{-h} (\zeta^2 D_k^h w) \,\mathrm{d}x &\le \int_D |f| |D_k^{-h} (\zeta^2 D_k^h w)| \,\mathrm{d}x \\ &\le \frac{2\kappa^3}{A_{\min}} \int_D |f|^2 \,\mathrm{d}x + \frac{A_{\min}}{4\kappa^3} \int_D |D_k^{-h} (\zeta^2 D_k^h w)|^2 \,\mathrm{d}x. \end{aligned} \quad \text{(B.9)}$$

Combining (B.8) and (B.9), we have

$$- \int_D f D_k^{-h} (\zeta^2 D_k^h w) \,\mathrm{d}x \le \frac{A_{\min}}{4} \int_W \zeta^2 |D_k^h \nabla w|^2 \,\mathrm{d}x + \frac{A_{\min}}{4} \int_W |\nabla w|^2 \,\mathrm{d}x + \frac{2\kappa^3}{A_{\min}} \int_W |f|^2 \,\mathrm{d}x. \quad \text{(B.10)}$$

Combining (B.7) and (B.10), we have

$$\int_D \zeta^2 |D_k^h \nabla w|^2 \,\mathrm{d}x \le \frac{8\kappa^3}{A_{\min}^2} \int_W |f|^2 \,\mathrm{d}x + \left[1 + 4\kappa^2 \|A\|_{C^1(\bar{U})} \frac{\|A\|_{C^1(\bar{U})} + A_{\min}}{A_{\min}^2}\right] \int_W |\nabla w|^2 \,\mathrm{d}x. \quad \text{(B.11)}$$

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

Therefore,

$$\int_D \zeta^2 |D_k^h \nabla w|^2 \, \mathrm{d}x \le \kappa \frac{\max(\|A\|_{C^1(\bar{U})}, 1)^2}{\min(A_{\min}, 1)^2} \left(\int_W |f|^2 \, \mathrm{d}x + \int_W |\nabla w|^2 \right). \tag{B.12}$$

Now we give an upper bound of $\int_D |\nabla w|$ by taking $v = \tilde{\zeta}^2 w$ in (A.15), 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 those for (B.12), we have

$$\int_W |\nabla w|^2 \, \mathrm{d}x \le \kappa \frac{\max(\|A\|_{C^1(\bar{U})}, 1)^2}{\min(A_{\min}, 1)^2} \left(\int_W |f|^2 \, \mathrm{d}x + \int_W |\nabla w|^2 \right). \tag{B.13}$$

Taking (B.12) and (B.13) together gives

$$\int_D \zeta^2 |D_k^h \nabla w|^2 \, \mathrm{d}x \le \kappa \frac{\max(\|A\|_{C^1(\bar{U})}, 1)^4}{\min(A_{\min}, 1)^4} \int_D |f|^2 + |w|^2 \, \mathrm{d}x. \tag{B.14}$$

We complete our proof by combining (B.14) for all $k = 1, \ldots, d$. □

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

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

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

$$\zeta = 1 \quad \text{on } B(0, \tfrac{1}{2}), \qquad \zeta = 0 \quad \text{on } B(0, 1)^{\mathrm{c}}, \qquad 0 \le \zeta \le 1.$$

For $k = 1, \ldots, d-1$, we substitute

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

into (A.15). Using the same arguments for deriving (B.11) as in the proof for Lemma B.1, we obtain

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

The above display holds for arbitrary h, so we have

$$\sum_{i,j=1,\, i+j<2d}^{d} \int_V \left| \frac{\partial^2 w}{\partial x_i \partial x_j} \right|^2 \mathrm{d}x \le \kappa \frac{\max(\|A\|_{C^1(\bar{U})}, 1)^2}{\min(A_{\min}, 1)^2} \int_W |f|^2 + |\nabla w|^2 \, \mathrm{d}x. \tag{B.15}$$

We proceed to an upper bound for

$$\int_V \left| \frac{\partial^2 w}{\partial x_d \partial x_d} \right|^2 \mathrm{d}x.$$

According to remark (ii) after Theorem 1 of [7, Chapter 6.3], with the interior regularity obtained by Lemma B.1, w solves (A.10) almost everywhere (a.e.) in U. Consequently,

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

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

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

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

Combining the above display with (B.15), we have

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

According to (B.13), the above display implies that

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

Similar to the proof for Proposition A.1, this result can be extended to the case where U has a twice differentiable boundary; we omit the details. □

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

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

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

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

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

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

Denote by $\alpha = (\alpha_1, \ldots, \alpha_d)^\top$ a multiple index with $|\alpha| = \alpha_1 + \cdots + \alpha_d = m$. With similar arguments as for (A.17), $\bar{w} = D^\alpha w$ is a weak solution to PDE (A.17) without a boundary condition. Similar to the derivation for (A.20), $w \in H^{m+2}(V)$ and

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

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 $\eta_1, \ldots, \eta_M$ with the following P-unisolvent property: given any real numbers $\alpha_1, \ldots, \alpha_M$, 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 $p_1, \ldots, p_M \in P$ such that $\eta_i(p_j) = \delta_{ij}$ for $1 \leq i, j \leq M$. Consequently, for all $p \in P$, the following holds:

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

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

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

Lagrange element. If there exist $a_1, \ldots, a_M$ 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 $a_1, \ldots, a_M$. Throughout this paper, we will consider only 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

- $F_T : \hat{T} \to T$, $F_T(\hat{x}) = B_T \hat{x} + b_T$,
- $a_i = F_T(\hat{a}_i)$, $1 \le i \le M$, and
- $p_i(x) = \hat{p}_i(F_T^{-1}(x))$, $1 \le i \le M$.

The mapping F_T is called the 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

- $T = F(\hat{T})$,
- $P = \{p = \hat{p} \circ F^{-1}; \hat{p} \in \hat{P}\}$, and
- $a_i = F(\hat{a}_i)$ 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 $\{(x_1, \ldots, x_d) : \sum_{i=1}^d x_i = 1,\ x_i \ge 0,\ i = 1, \ldots, d\}$ 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 ease of notation, we sometimes omit the index T in P_T and $a_{i,T}$, and write the element as $(T, P, \{p(a_i); 1 \le i \le M\})$.

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 $\dim(\hat{P}) = d(d-1)/2 + d$. The degree of freedom is chosen as follows:

(i) $\hat{a}_i$ is the vector with the ith entry being 1 and all other entries being 0, and
(ii) $\hat{a}_{ij} = \frac{1}{2}(\hat{a}_i + \hat{a}_j)$ is the midpoint of $\hat{a}_i$ and $\hat{a}_j$ for $1 \le i, j \le d$.

Triangularization of a domain. With the reference element specified, we 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, we may choose to partition the domain approximately. We elaborate on the requirement on the partition.

If a domain is a polygon, we 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 an affine family with a mesh size of $\max_{T \in \mathcal{T}_n} \operatorname{diam}(T) = O(2^{-n})$ is used to approximately cover the space U, that is, only a straight triangle is in use, then

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

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, p. 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 $\{(T, P_T, \Sigma_T)\colon T \in \mathcal{T}_n\}$ be a family of finite elements that are isoparametric equivalent to the reference element $(\hat{T}, \hat{P}, \{\hat{p}(\hat{a}_i),\ i = 1, \ldots, M\})$ with $\max_{T\in\mathcal{T}_n} \operatorname{diam}(T) = O(2^{-n})$ satisfying the following requirements.

(a) $\bar{U} = \bigcup_{T\in\mathcal{T}_n} T$.

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

(c) For $T, T' \in \mathcal{T}_n$, $T \neq T'$, $\operatorname{int}(T) \cap \operatorname{int}(T') = \varnothing$, where $\operatorname{int}(T)$ denotes the interior of the triangle T.

(d) If $T \neq T'$ but $T \cap T' \neq \varnothing$, 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 C.1. Among the requirements (a)–(d), (b) and (d) are standard assumptions and can be satisfied by many applications of interest. Assumption (a) 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

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

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

- There exists a constant $\sigma > 0$ such that, for all n and all $T \subset \mathcal{T}_n$,
$$\rho_T \ge \sigma h_T.$$
- For each $T \in \mathcal{T}_n$, let $\tilde{a}_{ij} = \frac{1}{2}(a_i + a_j)$ for all $1 \le i, j \le d$ and a_i, a_j being the vertices of T. We assume that
$$\|a_{ij} - \tilde{a}_{ij}\| = O(2^{-2n})$$
uniformly for all $T \in \mathcal{T}_n$.

Throughout the paper, we consider only the regular isoparametric family. In addition, we 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 for $d = 2$ and a tetrahedron for $d = 3$.

The function space V_n. Based on the 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

$$V_n = \{v \in C(\bar{U})\colon v|_T \in P_T \text{ for each } T \in \mathcal{T}_n \text{ and } v|_{\partial D} = 0\}. \tag{C.1}$$

https://doi.org/10.1017/apr.2018.49 Published online by Cambridge University Press

Isoparametric numerical integral. For isoparametric elements, the numerical integral is calculated by first performing a quadrature approximation over the reference element, and then transforming it to the isoparametric family. We first describe the integral approximation over the reference element $\hat{T} = \{(x_1, \ldots, x_d) : x_i \geq 0, \sum_{i=1}^{d} x_i \leq 1; 1 \leq i \leq d\}$. 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 $\int_{\hat{T}} \hat{\phi}(\hat{x})\, \mathrm{d}\hat{x}$ is approximated by $\sum_{l=1}^{M} \hat{w}_l \hat{\phi}(\hat{b}_l)$ for some weights $\hat{w}_l > 0$, points $\hat{b}_l$, $l = 1, \ldots, M$, and a positive integer M. In order to control the numerical error of the finite element method, we assume that the $\hat{w}_l$ and $\hat{b}_l$ are exact for quadratic functions. That is, if $\hat{\phi}$ is a quadratic function over $\hat{T}$ then $\int_{\hat{T}} \hat{\phi}(\hat{x})\, \mathrm{d}\hat{x} = \sum_{l=1}^{M} \hat{w}_l \hat{\phi}(\hat{b}_l)$. The choice of such a quadrature scheme is not unique. For example, a popular choice for $d = 2$ is $M = 3$, $b_1 = (0.5, 0)$, $b_2 = (0, 0.5)$, $b_3 = (0.5, 0.5)$, and $w_1 = w_2 = w_3 = \frac{1}{6}$. 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 $\int_T \phi(x)\, \mathrm{d}x$ 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 helpful discussions. Xiaoou Li was partially supported by the National Science Foundation (grant number DMS-1712657). Jingchen Liu was partially supported by the National Science Foundation (grant numbers SES-1323977, IIS-1633360, and SES-1826540) and the Army Research Office (grant number W911NF-15-1-0159).

References

[1] Charrier, J. (2012). Strong and weak error estimates for elliptic partial differential equations with random coefficients. *SIAM J. Numer. Anal.* **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 J. Numer. Anal.* **51,** 322–352.
[3] Ciarlet, P. (1991). Basic error estimates for elliptic problems. In *Handbook of Numerical Analysis*, Vol. 2. North-Holland, Amsterdam, pp. 17–351.
[4] Cliffe, K., Giles, M. B., Scheichl, R. and Teckentrup, A. L. (2011). Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Visualization Sci.* **14,** 3–15.
[5] Delhomme, J. P. (1979). Spatial variability and uncertainty in groundwater flow parameters: a geostatistical approach. *Water Resources Res.* **15,** 269–280.
[6] De Marsily, G. *et al.* (2005). Dealing with spatial heterogeneity. *Hydrogeol. J.* **13,** 161–183.
[7] Evans, L. C. (1998). *Partial Differential Equations*. American Mathematical Society, Providence, RI.
[8] Giles, M. B. (2008). Multilevel Monte Carlo path simulation. *Operat. Res.* **56,** 607–617.
[9] Graham, I. G. *et al.* (2011). Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. *J. Comput. Phys.* **230,** 3668–3694.
[10] Knabner, P. and Angermann, L. (2003). *Numerical Methods for Elliptic and Parabolic Partial Differential Equations*. Springer, New York.
[11] Ostoja-Starzewski, M. (2008). *Microstructural Randomness and Scaling in Mechanics of Materials*. Chapman and Hall/CRC Press.
[12] Rhee, C.-H. and Glynn, P. W. (2012). A new approach to unbiased estimation for SDE's. In *Proc. 2012 Winter Simul. Conf.*, IEEE, 7pp.
[13] Rhee, C.-H. and Glynn, P. W. (2013). Unbiased estimation with square root convergence for SDE models. *Operat. Res.* **63,** 1026–1043.
[14] Sobczyk, K. and Kirkner, D. J. (2001). *Stochastic Modeling of Microstructures*. Birkhäuser.
[15] Teckentrup, A. L., Scheichl, R., Giles, M. B. and Ullmann, E. (2013). Further analysis of multilevel Monte Carlo methods for elliptic PDEs with random coefficients. *Numer. Math.* **125,** 569–600.