MULTISCALE MODEL. SIMUL.
Vol. 8, No. 4, pp. 1419–1438

RANDOMIZATION OF FORCING IN LARGE SYSTEMS OF PDEs FOR IMPROVEMENT OF ENERGY ESTIMATES*

C. Y. LEE[†], B. L. ROZOVSKII[†], AND H. M. ZHOU[‡]

Abstract. We consider a class of stochastic PDEs (SPDEs) driven by purely spatial white noise, for which the numerical computation of the energy is desired. Our paper compares the efficiency of two different bases of expansion of white noise, one of a local scale and the other of a "large scale," for approximating the energy of the SPDE, and we will show that the latter basis dramatically improves the approximation of the energy. Such problems with a local scale basis arise in applications such as electromagnetic wave propagation with incoherent sources, but current approaches to computing the energy have found a roadblock in the sheer size of the problem. Thus, knowledge of the improved efficiency of a large scale basis becomes useful in vastly reducing computational cost while attaining highly accurate approximations of the energy.

Key words. stochastic partial differential equations, Wiener chaos expansion, change of basis

AMS subject classifications. 60H15, 60H35, 65C30

DOI. 10.1137/090766292

1. Introduction.

The problem. In this paper, we consider a linear stochastic PDE (SPDE)

$$\frac{\partial}{\partial t} v = \mathcal{A}v + \dot{W}(x), \quad x \in U, \quad t > 0, \tag{1.1}$$

and a system of deterministic PDEs

$$\frac{\partial}{\partial t} v_i(x,t) = \mathcal{A}v_i(x,t) + \rho_i e_i\left(x\right), \quad x \in U, \quad t > 0, \quad i = 1, 2, \ldots, \infty, \tag{1.2}$$

where $U \subset \mathbb{R}^d$ is an open bounded domain, $\mathcal{A}$ is a linear partial differential operator, $\{e_i,\, i \geq 1\}$ is an orthonormal basis in $L_2\left(U\right)$, and $\dot{W}(x)$ is a weighted spatial white noise, given by

$$\dot{W}(x) = \sum \rho_i e_i\left(x\right) \xi_i$$

with $\{\xi_i,\, i \geq 1\}$ being a set of independent Gaussian random variables and $\{\rho_i,\, i \geq 1\}$ being a set of nonnegative weights. If all $\rho_i = 1$, $\dot{W}(x)$ is a standard spatial white noise. Without loss of generality we assume that the initial conditions in (1.1) and (1.2) are zero. In fact, (1.1) and (1.2) are equivalent in that

$$v_i(x,t) = \mathbb{E}[v(x,t)\xi_i] \quad \text{and} \quad v(x,t) = \sum_i v_i(x,t) \left(\dot{W}, e_i\right)_{L_2(U)},$$

[*] Received by the editors July 26, 2009; accepted for publication (in revised form) April 20, 2010; published electronically August 4, 2010.
http://www.siam.org/journals/mms/8-4/76629.html

[†] Division of Applied Mathematics, Brown University, Providence, RI 02912 (chia_ying_lee@brown.edu, rozovsky@dam.brown.edu). The second author's research was supported by NSF grant DMS 0604863, ARO grant W911NF-07-1-0044, and AFOSR grant 5-21024 (inter).

[‡] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (hmzhou@math.gatech.edu). This author's research was supported by NSF Faculty Early Career Development (CAREER) award DMS-0645266.

where $\mathbb{E}$ stands for the expectation. Under very general assumptions, a solution of one of the two equations exists and is unique if and only if the other has a unique solution (see [10]). Moreover, if $\sum \rho_i^2 < \infty$, then the solution is in L_2, and if $\sum \rho_i^2 = \infty$, then the solution is found in a Sobolev space with a negative index.

The equivalence of (1.1) and (1.2) is a very simple implication of the so-called *Wiener chaos expansion* (WCE) for SPDEs. System (1.2) is called the *propagator* for (1.1).

The energy of a solution u of (1.1) is defined by

$$\mathcal{E}[v(t)] := \mathbb{E}\,\|v(\cdot,t)\|^2_{L_2(U)} = \sum_{i=1}^{\infty} \|v_i(\cdot,t)\|^2_{L_2(U)}. \tag{1.3}$$

Clearly, it is independent of the choice of the basis.

Our main goal is to identify bases $\{e_i,\, i \geq 1\}$ as well as estimators $\hat{v}^{(n)}(x,t) = \sum_{i=1}^n v_i(x,t)\rho_i\xi_i$ such that the energy of $\hat{v}_n(x,t)$ efficiently approximates $\mathcal{E}[v(t)]$. For a finite N-dimensional noise $\dot{W}(x)$, we want to study the behavior of the estimators as $N \to \infty$.

Getting a little bit ahead of the story, we remark that, while the energy $\mathcal{E}[v(t)]$ does not depend on the choice of the basis, the rate of convergence of the approximate energy $\sum_{i=1}^n \|v_i(\cdot,t)\|^2_{L_2(U)}$ does and, sometimes, does so quite substantially.

Approximating the energy $\|v(\cdot,t)\|^2_{L_2(U)}$ for system (1.2), and similar systems, requires solving a large number of PDEs that differ only by the forcing terms. For example, the problem of efficient approximation of the energy comes up in the modeling of wave propagation with incoherent sources [8], which appear in a wide range of problems in optics, such as those related to diffuse light [15]. Some popular examples include the Raman photonic crystal spectrometer [11], which is used to measure spatially incoherent light in environmental and biological sensing, as well as fluorescent or bioluminescent tomography [14], which has been used successfully to achieve in-vivo functional imaging in cancer research and drug monitoring. In modeling the performance of new designs for photonic crystal spectrometers, one has to compute the solutions of Maxwell equations, which govern the light propagation in the spectrometer, with spatially incoherent sources $f(x)$. Similarly, current models in fluorescent tomography are based on solving a diffusion approximation of the well-known radiative transport equation, and due to the random phase value it is again natural to model the incoherent fluorescent light source by point sources. Therefore, engineers routinely model incoherence by solving very large systems of equations, each of them excited by a point mass function $f_i(x) = f_i\delta_{x_i}(x)$, $i = 1, \ldots, N$.

On one hand, the incoherence property is well modelled by point sources in (1.2). On the other hand, the sheer number of required point sources sets a computational roadblock. To mitigate the aforementioned numerical complications, it was proposed in [1, 2] to circumvent the local scale problem by replacing the localized forcing terms with a new global scale forcing that efficiently consolidates most of the energy into just a few terms. This was implemented by replacing multiple Maxwell equations with point sources by a single Maxwell equation driven by white noise $\dot{W}_N(x) = \sum_{i=1}^N \xi_i m_i\,(x)$, where $\{\xi_i,\, i = 1, 2, \ldots, N\}$ were independent standard Gaussian random variables and $\{m_i,\, i = 1, 2, \ldots, N\}$ was a subset of a trigonometric basis. The numerical simulations presented in [1, 2] demonstrate a dramatic reduction in computational complexity in evaluating the energy $\|v(\cdot,t)\|^2_{L_2}$ while maintaining a similar level of accuracy of energy approximation. However, papers [1, 2] were not concerned with rigorous theoretical explanations of the validity of the proposed algorithm and the potential scope of its applicability.

In this paper, we present a rigorous approach to the problem of efficient approximation of the energy (1.3) for systems of fairly general evolution equations (1.2). More specifically, we assume that $\mathcal{A}$ is a self-adjoint positive definite elliptic operator of order $2m$. The domain U is an open bounded domain in $\mathbb{R}^d$, and we require $2m/d > 1/2$. Under these assumptions, the eigenfunctions of the operator $\mathcal{A}$ equipped with either periodic or zero Dirichlet boundary conditions form an orthonormal basis in $L_2(U)$.

In section 3, we compare the efficiency of the small scale (point forcing) basis and the "large scale" $\mathcal{A}$-eigenfunction basis, and we deduce our main result—the approximation of the energy using the latter basis yields a 1*st order* improvement over the former (see Theorem 3.1). In fact, we will show that the number of expansion terms under the eigenfunction basis is $\mathcal{O}(1)$ in N, whereas under the point forcing basis it is $\mathcal{O}(N)$. In section 4, we show numerical results for the one-dimensional heat equation under the point forcing and cosine bases that corroborate the theoretical results, and we also show results for the convection-diffusion equations that suggest the applicability of this method to a broader class of parabolic equations. These results could be easily extended to space-time white noise.

We remark that the change of basis method is not the only way to tackle the deterministic system. The key point is the randomization of the system (1.2) to the SPDE (1.1), which can then be handled by various methods, such as WCE or Monte Carlo simulation. While there are numerous works in the literature studying such equations with additive noise, most of these works use a single choice of basis, which is usually a generic basis or the basis derived from the Karhunen–Loève expansion [5, 6]. We point out that, at least in the case of a self-adjoint operator $\mathcal{A}$, the choice of new basis should be related to the eigenfunctions of $\mathcal{A}$ (see section 3), rather than to the basis arising from the Karhunen–Loève expansion of the white noise. Interestingly, [3, 7] specifically chose to use a basis similar to the point forcing basis, but this was only to expedite the use of the finite element method. In [3], the stochastic term was handled by Monte Carlo simulation, and L_2-convergence properties of the solutions were studied. To the best of our knowledge, direct comparison of two bases has not received as much attention.

Everywhere below we will restrict ourselves to the standard spatial white noise. If the energy of the white noise is finite and N is large, similar results could be obtained by simple rescaling.

2. Change of Wiener chaos basis.

We elaborate on the equations and assumptions. Let $U \subset \mathbb{R}^d$ be an open bounded domain. Let $-\mathcal{A}$ be a positive definite self-adjoint elliptic operator of order $2m$, equipped with either periodic or zero Dirichlet boundary conditions. (In the case of periodic boundary conditions, the domain U will be a torus $\mathbb{T}^d$.) We assume the dimensionality condition

$$2m/d > 1/2. \tag{2.1}$$

It is well known that $-\mathcal{A}$ has eigenfunctions $\{\mathfrak{m}_i\}$ that form an orthonormal basis in $L_2(U)$, and the corresponding eigenvalues $\{\lambda_i\}$ behave asymptotically as [13]

$$\lambda_i \sim i^{2m/d}. \tag{2.2}$$

We will refer to $\{\mathfrak{m}_i\}$ as the $\mathcal{A}$-*eigenfunction basis* in $L_2(U)$.

As an unbounded positive definite self-adjoint operator on $L_2(U)$, $-\mathcal{A}$ has a well-defined square root $\Lambda = \sqrt{-\mathcal{A}}$, which has domain $\mathcal{D}(\Lambda) = H^m_{\text{per}}$ or H^m_0. Then Λ induces a Hilbert scale which we denote by $H^\gamma_\mathcal{A}$, $\gamma \in \mathbb{R}$, with norms

$$\|\phi\|^2_{H^\gamma_\mathcal{A}} = \sum_{j=1}^{\infty} \left(\lambda_j^{1/2m}\right)^{2\gamma} \phi_j^2 \tag{2.3}$$

for ϕ of the form $\phi = \sum_{j=1}^J \phi_j \mathfrak{m}_j$, for some $J \in \mathbb{N}$ [9]. $H^\gamma_\mathcal{A}$ is the closure of the set of such ϕ in the norm $\|\cdot\|_{H^\gamma_\mathcal{A}}$. It can be shown that $H^\gamma_\mathcal{A}$ is equivalent to the usual Sobolev scale. In particular, the norm $\|\cdot\|_{H^{-2m}_\mathcal{A}}$ is equivalent to the Sobolev norm

$$\|\phi\|_{H^{-2m}} := \sup_{\psi \in H^{2m}_\cdot} \frac{|\langle \phi, \psi \rangle_{H^{-2m}, H^{2m}}|}{\|\psi\|_{H^{2m}}},$$

where we denoted $H^{2m}_\cdot = H^{2m}_{\text{per}}$ or H^{2m}_0 in the case of periodic or zero Dirichlet boundary conditions, respectively.

At this point, we introduce the related equation driven by an infinite dimensional spatial white noise, which will be used for the error analysis in section 3.1. Define the Gaussian white noise on $L_2(U)$ by the WCE $\dot{W}(x) := \sum_{i=1} \tilde{\xi}_i \mathfrak{m}_i(x)$, where $\tilde{\xi}_i$ are independent and identically distributed (i.i.d.) $\mathcal{N}(0,1)$ random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We consider the equation

$$\frac{\partial u^*}{\partial t} = \mathcal{A} u^* + \dot{W}(x) \tag{2.4}$$

with zero initial conditions and either periodic or zero Dirichlet boundary conditions.[1] Equation (2.4) is solved in the triple $H^{-m} \hookrightarrow L_2 \hookrightarrow H^m$ and should be understood in the weak sense. Its propagator system is

$$\frac{\partial \hat{u}^*_i}{\partial t} = \mathcal{A} \hat{u}^*_i + \mathfrak{m}_i(x). \tag{2.5}$$

The equivalence of the propagator system to the weak solution can be shown. Moreover, there exists a solution u^* such that $u^*(t) \in L_2(\Omega; L_2(U))$ for each $t \in (0, T]$, and the energy $\mathcal{E}[u^*] := \|u^*(t)\|^2_{L_2(\Omega; L_2(U))}$ at any fixed $t \in (0, T]$ is finite.

Fix an arbitrary $N < \infty$. All quantities introduced in the rest of this section depend on this parameter N, but in the future we will suppress explicitly writing this dependence on N if no ambiguity arises.

Let $\mathcal{I} = \{I_i, i = 1, \dots, N\}$ be a partition of U into (small) nonoverlapping subsets with Lebesgue measure $|I_i| \sim 1/N$. We assume the family $\mathcal{I}$ is *quasi-uniform* in N. That is, there exist constants ρ_1, ρ_2 such that

$$\max_i r_i \le (\rho_1 |U|)^{1/d} N^{-1/d},$$
$$\min_i \varepsilon_i \ge (\rho_2 |U|)^{1/d} N^{-1/d},$$

where $r_i = \text{diam}(I_i)$ and ε_i is the radius of the largest sphere B_i contained in I_i. The quasi-uniform assumption implies nondegeneracy, i.e., that there exists ρ_3 such that

[1] In this paper, we will always assume zero initial conditions and periodic or zero Dirichlet boundary conditions, even when not explicitly stated. We also always take $x \in U$ and $t \in (0, T]$ for arbitrary $T < \infty$.

$2\varepsilon_i \geq \rho_3 r_i^{(N)}$ for all i, N. It then follows that

$$\rho_- |U| N^{-1} \leq \min_i |I_i| \leq \max_i |I_i| \leq \rho_+ |U| N^{-1}$$

and

$$\tilde{\rho}_- (r_i)^d \leq |I_i| \leq \tilde{\rho}_+ (r_i)^d,$$

and hence

$$\varepsilon_i \sim r_i \sim N^{-1/d}.$$

Next, we introduce the two bases $\{n_i\}$ and $\{m_i\}$ that will be the focus of our comparative analysis.

1. *Point forcing basis*:

$$n_i(x) = \frac{1}{\sqrt{|I_i|}} \mathbf{1}_{I_i}(x) \quad \text{for } i = 1, \ldots, N, \tag{2.6}$$

and $\{n_i\}_{i=N+1}^{\infty}$ is any basis in $\mathcal{S}_N^{\perp}$, where $\mathcal{S}_N = \operatorname{span}\{n_i,\, i = 1, \ldots, N\}$.

2. *(Discrete) eigenfunction basis in* $\mathcal{S}_N$:

$$m_1 = \mathfrak{m}_1,$$
$$m_i = \frac{1}{Z_i}\left(\mathcal{P}_N \mathfrak{m}_i - \sum_{j=1}^{i-1} (\mathcal{P}_N \mathfrak{m}_i, m_j) m_j\right), \quad i = 2, \ldots, N, \tag{2.7}$$

where $\mathcal{P}_N$ is the L_2 projection onto $\mathcal{S}_N$ and Z_i is the normalization constant. In other words, $\{m_i\}$ is the Gram–Schmidt orthonormalization of the L_2 projections of the first N eigenfunction basis elements onto $\mathcal{S}_N$.

Define the spatial noise $\dot{W}_N(x) = \sum_{i=1}^{N} n_i(x)\eta_i$, where η_i are i.i.d. $\mathcal{N}(0,1)$ random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Clearly, $\dot{W}_N \in L_2(\Omega; H^{\gamma}(U))$ for any $\gamma \leq 0$. The equation in consideration is

$$\frac{\partial u}{\partial t} = \mathcal{A}u + \dot{W}_N(x), \quad x \in U, \quad 0 < t < T, \tag{2.8}$$

with zero initial conditions and either periodic or zero Dirichlet boundary conditions. We solve (2.8) in the triple $H^m(U) \hookrightarrow L_2(U) \hookrightarrow H^{-m}(U)$. The existence and uniqueness of the solution is known (see, e.g., [12]), and it is known that $u \in L_2(\Omega; L_2(0, T; H^m(U))) \cap L_2(\Omega; C^0(0, T; L_2(U)))$. In particular, it makes sense to consider $u(t)$ at any time $t \in (0, T]$. Hereafter, we fix an arbitrary $t \in (0, T]$, and we may suppress explicitly writing the dependence on t if no ambiguity arises.

The framework to allow us to change the basis of the WCE is elementary. We apply the usual change of basis formula to change the spatial basis in the expansion of $\dot{W}_N$,

$$\dot{W}_N(x) = \sum_{i=1}^{N} n_i(x)\eta_i = \sum_{i=1}^{N} m_i(x)\xi_i, \tag{2.9}$$

where $\{\xi_i\}$ is given by the usual change of basis formula

$$\xi_i = \sum_{i=1}^{N} (m_i, n_i)\eta_i.$$

It is clear that $\{\xi_i\} \sim$ i.i.d. $\mathcal{N}(0,1)$, and hence it is a basis in $\operatorname{span}\{\eta_i\} \subset L_2(\Omega)$. So (2.9) defines two equivalent WCEs for $\dot{W}_N$, and the solution of (2.8) also has two WCEs:

$$u(x,t) = \sum_{i=1}^{N} \hat{v}_i(x,t)\eta_i = \sum_{i=1}^{N} \hat{u}_i(x,t)\xi_i. \tag{2.10}$$

Putting the two expansions of u and $\dot{W}_N$ into (2.8) yields two equivalent propagator systems:

$$\frac{\partial \hat{v}_i}{\partial t} = \mathcal{A}\hat{v}_i + n_i(x), \tag{2.11a}$$

$$\frac{\partial \hat{u}_i}{\partial t} = \mathcal{A}\hat{u}_i + m_i(x) \tag{2.11b}$$

for $i = 1, \ldots, N$. In view of (2.10), the energy of the two systems must be equal:

$$\mathcal{E}[u] := \mathbb{E}\|u\|_{L^2}^2 = \sum_{i=1}^{N} \|\hat{v}_i\|_{L_2}^2 = \sum_{i=1}^{N} \|\hat{u}_i\|_{L_2}^2. \tag{2.12}$$

The energy $\mathcal{E}[u]$ is approximated by the energy of a *truncated* system. Truncating the systems (2.11) to $n < N$ equations means to consider the systems

$$\frac{\partial \hat{v}_i}{\partial t} = \mathcal{A}\hat{v}_i + n_i(x), \quad i = 1, \ldots, n, \tag{2.13a}$$

$$\frac{\partial \hat{u}_i}{\partial t} = \mathcal{A}\hat{u}_i + m_i(x), \quad i = 1, \ldots, n. \tag{2.13b}$$

System (2.13a) is the propagator system of

$$\frac{\partial v^{(n)}}{\partial t} = \mathcal{A}v^{(n)} + \dot{W}_n(x), \tag{2.14}$$

whereas system (2.13b) is the propagator system of

$$\frac{\partial u^{(n)}}{\partial t} = \mathcal{A}u^{(n)} + \dot{Y}_n(x), \tag{2.15}$$

where $\dot{Y}_n(x) = \sum_{i=1}^{n} m_i(x)\xi_i$. Obviously, (2.14) and (2.15) are different SPDEs with different energies:

$$\mathcal{E}[v^{(n)}] = \sum_{i=1}^{n} \|\hat{v}_i\|_{L_2}^2 \neq \sum_{i=1}^{n} \|\hat{u}_i\|_{L_2}^2 = \mathcal{E}[u^{(n)}].$$

$\mathcal{E}[v^{(n)}]$ and $\mathcal{E}[u^{(n)}]$ will be taken as an approximation to the true energy $\mathcal{E}[u]$. The absolute and relative errors of the approximations by size n truncations will be denoted as

$$R[v^{(n)}] := \mathcal{E}[u] - \mathcal{E}[v^{(n)}] = \sum_{i=n+1}^{N} \|\hat{v}_i\|_{L_2}^2 \quad \text{and} \quad \bar{R}[v^{(n)}] = \frac{R[v^{(n)}]}{\mathcal{E}[u]}, \tag{2.16a}$$

$$R[u^{(n)}] := \mathcal{E}[u] - \mathcal{E}[u^{(n)}] = \sum_{i=n+1}^{N} \|\hat{u}_i\|_{L_2}^2 \quad \text{and} \quad \bar{R}[u^{(n)}] = \frac{R[u^{(n)}]}{\mathcal{E}[u]} \tag{2.16b}$$

for $n \leq N$. We will compare the performance of the two bases using the relative error of the approximate energy. Given an allowable relative error r, let

$$n_P := \inf\{n : \bar{R}[v^{(n)}] < r\} \quad \text{and} \quad n_E := \inf\{n : \bar{R}[u^{(n)}] < r\} \tag{2.17}$$

be the minimum truncation size of (2.13) that achieves the relative error r for the point forcing and the eigenfunction bases, respectively. Define the *improvement* of the eigenfunction basis over the point forcing basis as

$$\frac{n_P}{n_E}. \tag{2.18}$$

This number is an indication of the computational savings of using the eigenfunction basis for the relative error r.

3. Comparative error analysis and 1st order improvement.

In the foregoing section, all the quantities depend on the number N of subdivisions of U. In this section, we study the asymptotic behavior as $N \to \infty$. We will formulate precise bounds on the relative error and compare the asymptotic behavior of the two bases as $N \to \infty$.

The main goal of this section is to show the 1*st order improvement* of the change of basis method, in the sense of the following theorem.

THEOREM 3.1. *Given a relative error* $r \in (0, 1)$, *we have, at worst,* 1st order improvement *as* $N \to \infty$.

More precisely, there exist constants $0 < C_{0,min} < C_{0,max} \leq 1$, *depending on* r *but independent of* N, *such that for every* $C_0 \in [C_{0,min}, C_{0,max})$ *there exists* $N_0 = N_0(C_0) > 0$ *such that*

$$\frac{n_P}{n_E} \geq C_0 N$$

whenever $N > N_0$. *Moreover,* $N_0 \to \infty$ *as* $C_0 \uparrow C_{0,max}$.

Obviously, 1st order improvement is the best one can hope for, simply because $n_P \leq N$ and $n_E \geq 1$, so that $n_P/n_E \leq N$. The result of Theorem 3.1 states that the constant in front of the 1st order improvement can vary in an interval, with a larger constant holding for larger N.

A big part of the proof of Theorem 3.1 involves studying the decay in n of the relative errors $\bar{R}[v^{(n)}]$ and $\bar{R}[u^{(n)}]$. Theorem 3.1 follows easily from two key facts: first, that the relative error $\bar{R}[v^{(n)}]$ for the point forcing basis decays no faster than linearly; second, that the relative error $\bar{R}[u^{(n)}]$ for the eigenfunction basis decays no slower than superlinearly, on the order of $n^{-\alpha}$ with $\alpha > 0$. (See Figures 1 and 2 for illustration and motivation.) We make these two statements more precise in the following two propositions.

PROPOSITION 3.2. *For the solution of* (2.4), *define the relative error by*

$$\bar{R}[u^{*,(n)}] := \frac{\sum_{i=n+1}^{\infty} \|\hat{u}_i^*\|_{L_2}^2}{\mathcal{E}[u^*]}$$

for $n = 1, 2, \ldots$. *Then*

$$\bar{R}[u^{*,(n)}] \sim n^{-4m/d+1}. \tag{3.1}$$

Given a relative error r,

$$n_0 := \inf\left\{n : \bar{R}[u^{*,(n)}] < r\right\} \sim r^{\frac{d}{d-4m}} \tag{3.2}$$

as $r \downarrow 0$.

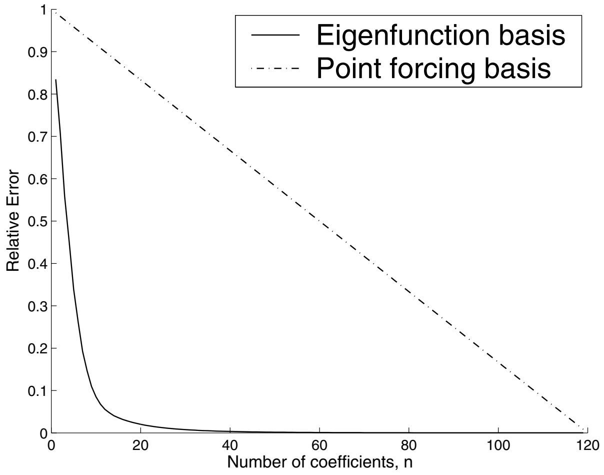

FIG. 1. *Relative errors incurred $\bar{R}[u^{(n)}]$ when the system is truncated to n coefficients, under the point forcing basis (dotted line) and the eigenfunction (cosine) basis (solid line). The convection-diffusion equation was used to produce this data.*

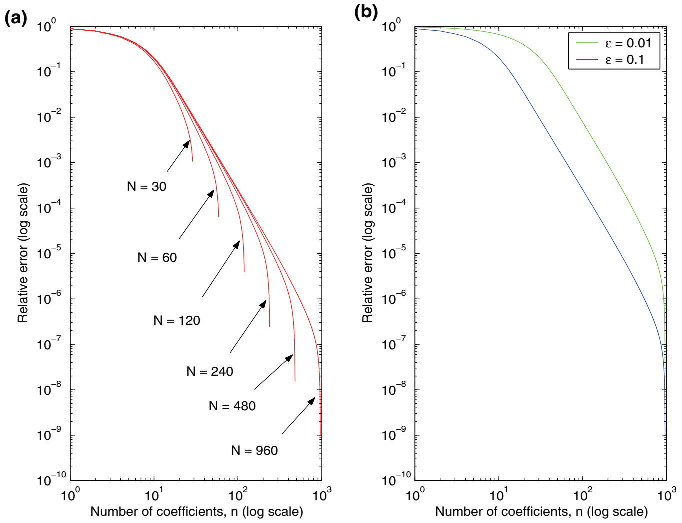

FIG. 2. (a) *Relative errors on log-log axes for increasing values of N, under the cosine basis for the heat equation.* (b) *Relative errors for two values of diffusion coefficients $\epsilon = 0.1, 0.01$. The graph for $\epsilon = 0.01$ lies above the graph for $\epsilon = 0.1$.*

PROPOSITION 3.3. *There exists a constant C independent of n and N such that*

$$\bar{R}[v^{(n)}] \geq L(n) = 1 - nCN^{-1}, \tag{3.3}$$

where $L(n)$ is a straight line passing through $(0,1)$ and with slope $-CN^{-1}$, which tends to 0 as $N \to \infty$.

To show the decay behavior of the relative errors, we will focus on finding bounds on the L_2 norms of the solution modes $\hat{u}_i$ and $\hat{v}_i$. In order to be useful for explaining this contrasting behavior of the two bases, the bounds need to be sensitive to the localness or globalness of the basis and should provide accurate bounds on the solution modes. Error bounds involving $\|n_i\|^2_{L_2}$ and $\|m_i\|^2_{L_2}$ are clearly insensitive to the choice of basis, since both norms equal 1. Standard methods for estimating the time evolution of $\|u(t)\|^2_{L_2}$, such as those involving Gronwall's inequality, may also be inadequate. A case in point is the following.

Suppose $u(t)$ solves the heat equation

$$\frac{\partial u}{\partial t} = \Delta u + f(x), \quad x \in [0, X],$$

with zero initial conditions and periodic boundary conditions (cf. section 4.1). Also assume that $u(t)$ has periodic derivative. Then

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|^2_{L_2} &= \int_U u(x,t)u_t(x,t)\,dx \\ &\leq -\|u_x(t)\|^2_{L_2} + \|u(t)\|_{H^1}\,\|f\|_{H^{-1}} \\ &\leq -\|u_x(t)\|^2_{L_2} + \left(\|u(t)\|^2_{L_2} + \|u_x(t)\|^2_{L_2}\right) + \frac{1}{4}\|f\|^2_{H^{-1}} \\ &= \|u(t)\|^2_{L_2} + \frac{1}{4}\|f\|^2_{H^{-1}}. \end{aligned}$$

Thus, by Gronwall's inequality,

$$\|u(t)\|^2_{L_2} \leq \frac{te^{2t}}{4}\,\|f\|^2_{H^{-1}} = C(t)\|f\|^2_{H^{-1}}.$$

Using n_i or the cosine basis m_i in place of f, the energy of each mode is bounded by

$$\begin{aligned} \|\hat{u}_i(t)\|^2_{L_2} &\leq C(t)\|m_i\|^2_{H^{-1}} \leq C(t)\left(\frac{X}{(i-1)\pi}\right)^2, \\ \|\hat{v}_i(t)\|^2_{L_2} &\leq C(t)\|n_i\|^2_{H^{-1}} \leq C(t). \end{aligned}$$

Then the absolute error of the energy estimate of an n-equation truncated system (for fixed N) decays on the order of

$$R[u^{(n)}] \sim \mathcal{O}\left(\frac{1}{n} - \frac{1}{N}\right), \quad R[v^{(n)}] \sim \mathcal{O}\left(\frac{N-n}{N}\right).$$

The estimate for $R[v^{(n)}]$ is consistent with numerical results, but we will establish this result in more generality. But the estimate for $R[u^{(n)}]$ is merely an upper bound and does not predict the actual $\mathcal{O}\left(n^{-3}\right)$ decay (see Figure 2).

3.1. Fourier techniques for the eigenfunction basis. The Fourier expansion is an effective technique for obtaining exact error estimates. We begin by considering the limiting infinite dimensional case, (2.4), and its propagator system, (2.5). The solution of (2.4) has a formal expansion:

$$u^*(t) = \sum_{i=1}^{\infty} \hat{u}^*_i(t)\xi_i = \sum_{i,j=1}^{\infty} \hat{\hat{u}}^*_{ij}(t)\mathfrak{m}_j\xi_i,$$

where $\hat{\hat{u}}^*_{ij} = (\hat{u}^*_i, \mathfrak{m}_j)$ are the Fourier coefficients of $\hat{u}^*_i$ with respect to the $\mathcal{A}$-eigenfunction basis in $L_2(U)$. Note that only the low order modes $\{\xi_i\mathfrak{m}_j\}_{i,j\geq 1}$ are nonzero because the noise appears additively. From the propagator system (2.5), the Fourier coefficients solve the decoupled system of ODEs

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\hat{\hat{u}}^*_{ij} = -\lambda_j \hat{\hat{u}}^*_{ij} + \delta_{ij}, \\ \hat{\hat{u}}^*_{ij}(0) = 0 \end{cases} \tag{3.4}$$

for $i, j = 1, 2, \ldots$. The solution is

$$\hat{\hat{u}}^*_{ij}(t) = \delta_{ij} \int_0^t e^{-\lambda_j(t-s)}\, ds = \delta_{ij}(t) \frac{1 - e^{-\lambda_j t}}{\lambda_j}.$$

In other words, all the energy is concentrated on the modes $\{\xi_i \mathfrak{m}_i\}$, and

$$\mathbb{E}\|u^*(t)\|^2_{L_2(U)} = \sum_{i=1}^{\infty} \|\hat{u}^*_i(t)\|^2_{L_2(U)} \leq \sum_{i=1}^{\infty} \frac{1}{\lambda_i^2},$$

and the last sum converges because of the asymptotic behavior of the eigenvalues (2.2) and the dimensionality condition (2.1). So $u^*(t) \in L_2(\Omega; L_2(U))$ is square integrable. It also follows from

$$\sum_{i=1}^{\infty} \|\hat{u}^*_i(t)\|^2_{L_2(U)} \geq (1 - e^{-\lambda_1 t})^2 \sum_{i=1}^{\infty} \frac{1}{\lambda_i^2}$$

that (2.1) is necessary for $u^*(t)$ to be square integrable.

The consequence of this computation is that we have precise asymptotic estimates for the truncation error:

$$R[u^{*,(n)}] := \sum_{i=n+1}^{\infty} \|\hat{u}^*_i\|^2_{L_2} = \sum_{i=n+1}^{\infty} (\hat{\hat{u}}^*_{ii})^2 \sim n^{-4m/d+1}. \tag{3.5}$$

The asymptotics in (3.5) obviously hold for $\bar{R}[u^{*,(n)}] = R[u^{*,(n)}]/\mathcal{E}[u^*]$ as well, and we obtain (3.1) in Proposition 3.2. Equation (3.2) then follows by taking the function inverse of the asymptotic bounds in (3.1). Thus,

$$C r^{\frac{d}{d-4m}} \leq n_0 \leq C' r^{\frac{d}{d-4m}} \tag{3.6}$$

for r sufficiently small.

We now look at the finite dimensional model, $N < \infty$. $\dot{W}_N$ can be viewed as a finite dimensional approximation of $\dot{W}$, and we have that $\mathcal{E}[u] \to \mathcal{E}[u^*]$.

Instead of (3.4), the relevant system of ODEs for (2.8) corresponding to the discrete eigenfunction basis in $\mathcal{S}_N$ comes from (2.11b):

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\hat{\hat{u}}_{ij} = -\lambda_j \hat{\hat{u}}_{ij} + (m_i, \mathfrak{m}_j), \\ \hat{\hat{u}}_{ij}(0) = 0 \end{cases} \tag{3.7}$$

for $i = 1, \ldots, N$, $j = 1, 2, \ldots$. Since the projections $\mathcal{P}_N \mathfrak{m}_i \to \mathfrak{m}_i$ in L_2 as $N \to \infty$ for each $i = 1, 2, \ldots$, then, by an inductive procedure, the Gram–Schmidt orthonormalized elements m_i in (2.7) are finite sums of L_2-convergent terms. Hence, $m_i \to \mathfrak{m}_i$ in L_2 as $N \to \infty$ for each $i = 1, 2, \ldots$. For any fixed j, $(m_i, \mathfrak{m}_j) \to \delta_{ij}$ also. Thus, we may expect the behavior of system (3.7) to be close to that of (3.4). Indeed,

$$\hat{\hat{u}}_{ij}(t) = (m_i, \mathfrak{m}_j) \int_0^t e^{-\lambda_j(t-s)}\, ds.$$

Then for each i

$$\|\hat{u}_i\|_{L_2}^2 = \sum_{j=1}^{\infty} (m_i, \mathfrak{m}_j)^2 \left(\int_0^t e^{-\lambda_j (t-s)} \, ds \right)^2$$
$$\longrightarrow \sum_{j \geq 1} \delta_{ij} \left(\int_0^t e^{-\lambda_j (t-s)} \, ds \right)^2 = \|\hat{u}_i^*\|_{L_2}^2$$

as $N \to \infty$ by dominated convergence. Since $\mathcal{E}[u] \to \mathcal{E}[u^*]$, it follows that

$$R[u^{(n)}] = \mathcal{E}[u] - \mathcal{E}[u^{(n)}] \overset{N\to\infty}{\longrightarrow} \mathcal{E}[u^*] - \mathcal{E}[u^{*,(n)}] = R[u^{*,(n)}] \tag{3.8}$$

for every n. In particular, we deduce that the relative error of the size n truncations of the finite, size N system must tend to that of the infinite system, pointwise in n.

We do not assert that $\bar{R}[u^{(n)}] = \mathcal{O}\left(n^{-\frac{4m}{d}+1}\right)$. Nonetheless, we can deduce an asymptotic result for the minimum truncation size, similar to n_0 in (3.2). Recall the minimum truncation size $n_E = n_E(N, r)$ to achieve relative error r, (2.17), for the discrete eigenfunction basis.

Proposition 3.4. *There exists r^* such that, for all $r < r^*$, there exists $N(r)$ such that*

$$Cr^{\frac{d}{d-4m}} \leq n_E \leq C' r^{\frac{d}{d-4m}}$$

whenever $N \geq N(r)$. The constants C, C' are independent of r, N.

Proof. From (3.6), there exists r^* such that

$$Cr^{\frac{d}{d-4m}} \leq n_0(r) \leq C' r^{\frac{d}{d-4m}}$$

whenever $r < r^*$. Fix $\delta \in (0, 1)$. For any $r < r^*/(1+\delta)$, choose $N(r)$ such that

$$\left| \bar{R}[u^{(n)}] - \bar{R}[u^{*,(n)}] \right| < \delta r$$

holds for both $n = n_0((1+\delta)r)$ and $n = n_0((1-\delta)r)$. Then

$$\bar{R}[u^{(n)}]|_{n=n_0((1+\delta)r)} > r \quad \text{and} \quad \bar{R}[u^{(n)}]|_{n=n_0((1-\delta)r)} < r$$

and

$$n_0((1+\delta)r) < n_E \leq n_0((1-\delta)r).$$

Hence

$$C(1+\delta)^{\frac{d}{d-4m}} r^{\frac{d}{d-4m}} \leq n_E \leq C'(1-\delta)^{\frac{d}{d-4m}} r^{\frac{d}{d-4m}},$$

and the result follows with $r^*/(1+\delta)$ in place of r^*. □

3.2. H^{-2m} norm estimates for the point forcing basis.

For the error analysis for the point forcing basis, similar computations for the system of ODEs for the point forcing basis coming from (2.11b) show that $\hat{\hat{v}}_{ij} := (\hat{v}_i, \mathfrak{m}_j)$ satisfies

$$\hat{\hat{v}}_{ij}(t) = (n_i, \mathfrak{m}_j) \int_0^t e^{-\lambda_j (t-s)} \, ds.$$

In particular, the energy of the system is *not* concentrated on $\{\hat{\hat{v}}_{ii},\, i = 1, \ldots, N\}$, and

$$\|\hat{v}_i\|^2_{L_2} = \sum_{j=1}^{\infty} (n_i, \mathfrak{m}_j)^2 \left(\int_0^t e^{-\lambda_j (t-s)}\, ds \right)^2 . \tag{3.9}$$

We have the following lemma.

Lemma 3.5. *Let n_i be defined in* (2.6) *for $i = 1, \ldots, N$. Then we have the bounds*

$$C_1 N^{-2m/d} \le \|n_i\|_{H^{-2m}} \le C_2 N^{-1/2},$$

where C_1, C_2 are independent of i and N.

Proof. For the lower bound, consider the mollifier ζ_ε with support in $B(0, \varepsilon)$, and let α_i be the center of the largest sphere B_i contained in I_i with radius ε_i. Then, denoting $H^{2m}_{\cdot} = H^{2m}_{\text{per}}$ or H^{2m}_0,

$$\begin{aligned} \|n_i\|_{H^{-2m}} &= \sup_{\psi \in H^{2m}_{\cdot}(U)} \frac{|\langle n_i, \psi\rangle|}{\|\psi\|_{H^{2m}(U)}} \\ &\ge \|\zeta_{\varepsilon_i}(\cdot - \alpha_i)\|^{-1}_{H^{2m}(U)} \int_U n_i(x) \zeta_{\varepsilon_i}(x - \alpha_i)\, dx \\ &= \|\zeta_{\varepsilon_i}\|^{-1}_{H^{2m}(\mathbb{R}^d)} (n_i * \zeta_{\varepsilon_i})(\alpha_i) \ge C N^{-2m/d}. \end{aligned}$$

The last inequality holds because it can be computed that $\|\zeta_{\varepsilon_i}\|_{H^{2m}} \sim \varepsilon_i^{-(2m+d/2)}$ and $(n_i * \zeta_{\varepsilon_i})(\alpha_i) = n_i(\alpha_i) \sim N^{1/2} \sim \varepsilon_i^{-d/2}$.

For the upper bound,

$$\begin{aligned} \|n_i\|_{H^{-2m}} &= \sup_{\psi \in H^{2m}_{\cdot}} \frac{|\langle n_i, \psi\rangle|}{\|\psi\|_{H^{2m}}} \\ &\le |I_i|^{1/2} \sup_{\psi \in H^{2m}_{\cdot}} \frac{\frac{1}{|I_i|}\int_{I_i} |\psi|\, dx}{\|\psi\|_{H^{2m}}} \\ &\le C N^{-1/2}, \end{aligned}$$

where the C is independent of i and N, since by the Sobolev embedding every $\psi \in H^{2m}_0(U)$, or every $\psi \in H^{2m}_{\text{per}}(U)$ in the periodic case with a rectangular domain, belongs to $C^{0,1/2}(\bar{U})$. □

Corollary 3.6. *For each $i = 1, \ldots, N$, we have the bounds*

$$C_3 N^{-4m/d} \le \|\hat{v}_i\|^2_{L_2} \le C_4 N^{-1},$$

where C_3, C_4 are independent of i and N.

Proof. From the definition of the $H^\gamma_{\mathcal{A}}$ norm, (2.3),

$$\|\hat{v}_i\|^2_{L_2} \ge \sum_{j=1}^{\infty} \left((n_i, \mathfrak{m}_j) \frac{1 - e^{-\lambda_1 t}}{\lambda_j} \right)^2 = (1 - e^{-\lambda_1 t})^2 \|n_i\|^2_{H^{-2m}_{\mathcal{A}}}$$

and

$$\|\hat{v}_i\|^2_{L_2} \le \sum_{j=1}^{\infty} \left((n_i, \mathfrak{m}_j) \frac{1}{\lambda_j} \right)^2 = \|n_i\|^2_{H^{-2m}_{\mathcal{A}}}.$$

By the equivalence of the $H^{\gamma}_{\mathcal{A}}$ norms and the Sobolev norms, and from Lemma 3.5, it follows that

$$C_3 N^{-4m/d} \leq \|\hat{v}_i\|^2_{L_2} \leq C_4 N^{-1}. \qquad \square$$

The lower bound in Corollary 3.6 gives another way to see that the solution is not square integrable if the dimensionality condition (2.1) is not met. This lower bound also gives a lower bound for the relative error of the point forcing basis. Interestingly, a more informative lower bound on the relative error can be derived from the upper bound in Corollary 3.6:

$$\begin{aligned} \bar{R}[v^{(n)}] &= \frac{\mathcal{E}[u] - \sum_{i=1}^{n} \|\hat{v}_i\|^2_{L_2}}{\mathcal{E}[u]} \\ &\geq \frac{\mathcal{E}[u] - nC_4N^{-1}}{\mathcal{E}[u]} = 1 - n\frac{C_4}{N\mathcal{E}[u]} =: L(n). \end{aligned} \tag{3.10}$$

Since the constant C_4 is independent of n and N, the relative error is bounded from below by a straight line $L(n)$ passing through the point $(0, 1)$ and with slope $-C_4/(N\mathcal{E}[u])$, which tends to 0 as $N \to \infty$. We have just shown Proposition 3.3.

We now prove Theorem 3.1.

Proof of Theorem 3.1. From (3.10), the linear lower bound $L(n)$ attains relative error r for $n \geq n_L$, where

$$n_L = \frac{(1-r)\mathcal{E}[u]}{C_4} N = \tilde{C}(N)N$$

is the value such that $L(n_L) = r$. Then, since $\bar{R}[u^{(n)}] \geq L(n)$,

$$n_P \geq n_L = \tilde{C}(N)N.$$

$\tilde{C}(N)$ depends on N because $\mathcal{E}[u]$ depends on N. We next show a series of estimates for $\tilde{C}(N)$ to remove the dependence on N. First, $\mathcal{E}[u_{\{N\}}] \geq \mathcal{E}[u_{\{N=1\}}]$ for all N, so

$$\tilde{C}(N) \geq \tilde{C}(1)$$

for all N. Now, since $\mathcal{E}[u] \to \mathcal{E}[u^*]$, for any $\epsilon \in (0, \mathcal{E}[u^*] - \mathcal{E}[u_{\{N=1\}}])$, there exists $N(\epsilon)$ such that $\mathcal{E}[u] \geq \mathcal{E}[u^*] - \epsilon$. So

$$\tilde{C}(N) \geq \frac{(1-r)(\mathcal{E}[u^*] - \epsilon)}{C_4}$$

whenever $N > N(\epsilon)$. Denote $\tilde{C}(\infty) = \frac{(1-r)\mathcal{E}[u^*]}{C_4}$. As ϵ ranges from 0 to $\mathcal{E}[u^*] - \mathcal{E}[u_{\{N=1\}}]$, the right-hand side of the last inequality ranges from $\tilde{C}(\infty)$ to $\tilde{C}(1)$. Clearly, $N(\epsilon)$ increases to ∞ as $\epsilon \downarrow 0$. In other words, for any $C \in [\tilde{C}(1), \tilde{C}(\infty))$, there exists $N(C)$ such that

$$n_P \geq \tilde{C}(N)N \geq CN$$

whenever $N > N(C)$. Moreover, $N(C)$ increases to ∞ as $C \uparrow \tilde{C}(\infty)$.

For n_E, recall $n_0 = \inf\{n : \bar{R}[u^{*,(n)}] < r\}$, (3.2). Let $\epsilon_0 = r - \bar{R}[u^{*,(n_0)}] > 0$. From (3.8), $\bar{R}[u^{(n_0)}] \to \bar{R}[u^{*,(n_0)}]$ as $N \to \infty$. So there exists $N(n_0) > 0$ such that

$$\bar{R}[u^{(n_0)}] < \bar{R}[u^{*,(n_0)}] + \epsilon_0 = r$$

whenever $N > N(n_0)$. Hence, $n_E \leq n_0$ if $N > N(n_0)$.

Combining the two inequalities for n_P, n_E,

$$\frac{n_P}{n_E} \geq \frac{CN}{n_0} = C_0 N$$

whenever $N > N_0(C_0) := \max\{N(C), N(n_0)\}$. Hence, $C_0 \in [C(1)/n_0, C(\infty)/n_0)$ and $N_0 \to \infty$ as $C_0 \uparrow C(\infty)/n_0$. $\square$

The next result gives upper and lower bounds on the improvement in terms of the relative error r.

Corollary 3.7. *There exist $r^* \in (0,1)$ and constants $0 < C_{*,min} < C_{*,max} \leq C_* \leq 1$ such that, for every $r < r^*$ and every $C_0 \in [C_{*,min}, C_{*,max})$, there exists $N_0 = N_0(r, C_0) > 0$ such that*

$$C_0 r^{-\frac{d}{d-4m}} N \leq \frac{n_P}{n_E} \leq C_* r^{-\frac{d}{d-4m}} N$$

whenever $N > N_0$. Moreover, $N_0 \to \infty$ as $C_0 \uparrow C_{*,max}$ or as $r \downarrow 0$.

Proof. In the proof of Theorem 3.1, the inequalities hold if we replace $\tilde{C}(N)$ with $\tilde{C}_*(N) := \frac{(1-r^*)\mathcal{E}[u]}{C_4}$ so that, for any $C \in [\tilde{C}_*(1), \tilde{C}_*(\infty))$, there exists $N(C)$ such that

$$n_P \geq \tilde{C}_*(N) N \geq CN$$

whenever $N > N(C)$. Also $N(C)$ increases to ∞ as $C \uparrow \tilde{C}_*(\infty)$. From Proposition 3.4,

$$\frac{n_P}{n_E} \geq \frac{CN}{C' r^{\frac{d}{d-4m}}} = C_0 N r^{-\frac{d}{d-4m}}$$

whenever $N > N_0(r, C_0)$.

Also from Proposition 3.4, and since $n_P \leq N$,

$$\frac{n_P}{n_E} \leq \frac{N}{C r^{\frac{d}{d-4m}}} = C_* N r^{-\frac{d}{d-4m}}. \qquad \square$$

If $r_1 < r_2$, then $r_1^{-\frac{d}{d-4m}} < r_2^{-\frac{d}{d-4m}}$, so Corollary 3.7 indicates that one would expect a slower convergence to 1st order improvement for a smaller relative error, in accordance with trend (T3) in the numerical simulations. We also note that the interval endpoints in Theorem 3.1 and Corollary 3.7 are inversely proportional to C_4 from Corollary 3.6, which is in turn inversely proportional to the norm of $\mathcal{A}$. This point is corroborated by the numerical result that showed that the improvement is better for a larger diffusivity constant (cf. trend (T1)).

3.3. The non–self-adjoint case.

If $\mathcal{A}$ is not self-adjoint or not positive definite, precise bounds on the error decay may not be readily available, but under additional assumptions we can still deduce certain asymptotic results similar to the positive definite self-adjoint case, such as 1st order improvement. Assume $\mathcal{A}$ is a $2m$th order elliptic operator, and solve the SPDE (2.8) in the triple $H^m \hookrightarrow L_2 \hookrightarrow H^{-m}$. Also assume a more stringent dimensionality condition:

$$m/d > 1/2. \tag{3.11}$$

We split $\mathcal{A} = \mathcal{A}_0 + \mathcal{A}_1$, where $\mathcal{A}_0$ contains the highest $2m$th order terms and $\mathcal{A}_1$ are all lower order terms. Then $-\mathcal{A}_0$ is positive definite self-adjoint and generates an eigenfunction basis $\{\mathfrak{m}_i\}$ with eigenfunctions $\{\lambda_i\}$ satisfying (2.2). Similarly to (2.3), $\mathcal{A}_0$ defines a scale of Hilbert spaces $H^\gamma_{\mathcal{A}_0}$, with norm $\|\phi\|^2_{H^\gamma_{\mathcal{A}_0}} = \sum_{j=1}^\infty (\phi, \mathfrak{m}_j)^2 \lambda_j^{\gamma/m}$, that is equivalent to the Sobolev scale H^γ.

In the infinite dimensional case with white noise (2.4), the existence and uniqueness of the solution u^* is shown in [12] because the asymptotics of the eigenvalues (2.2) and the new dimensionality condition (3.11) imply that $\dot{W} \in L_2(\Omega; H^{-m}(U))$. Applying the usual deterministic parabolic estimates to the propagator system (2.5),

$$\sup_{t\in(0,T]} \|\hat{u}_i^*(t)\|_{L_2(U)} + \|\hat{u}_i^*(t)\|_{L_2(0,T;H^m(U))} \le C\|\mathfrak{m}_i\|_{L_2(0,T;H^{-m}(U))}.$$

In particular, since $\hat{u}_i^*(t)$ is continuous in t, for each $t \in (0,T]$,

$$\|\hat{u}_i^*(t)\|^2_{L_2(U)} \le C\|\mathfrak{m}_i\|^2_{H^{-m}} \le C'\|\mathfrak{m}_i\|^2_{H^{-m}_{\mathcal{A}_0}} \le C'\lambda_i^{-1}.$$

Then we have a result analogous to Proposition 3.2. For the error

$$R[u^{*,(n)}] := \sum_{i=n+1}^{\infty} (\hat{u}_{ii}^*)^2 \le Cn^{-2m/d+1}, \tag{3.12}$$

and for $n_0 := \min\{n : \bar{R}[u^{*,(n)}] < r\}$,

$$n_0 \le Cr^{\frac{d}{d-2m}}.$$

In the finite dimensional case (2.8), we again have $\mathcal{E}[u] \to \mathcal{E}[u^*]$. For the discrete eigenfunction basis,

$$\left|\|\hat{u}_i\|^2_{L_2} - \|\hat{u}_i^*\|^2_{L_2}\right| \le C\|\hat{u}_i - \hat{u}_i^*\|_{L_2} \le C'\|m_i - \mathfrak{m}_i\|_{H^{-m}} \overset{N\to\infty}{\longrightarrow} 0$$

so that $\bar{R}[u^{(n)}] \to \bar{R}[u^{*,(n)}]$ as $N \to \infty$ for each n. For the point forcing basis,

$$\|\hat{v}_i(t)\|^2_{L_2(U)} \le C\|n_i\|^2_{H^{-m}} \le C'N^{-1},$$

where the last inequality follows by an argument similar to the upper bound in Lemma 3.5. The proof of Theorem 3.1 follows through identically, so the statement of 1st order improvement applies to the non–self-adjoint case as well, provided (3.11) holds.

However, this argument by parabolic estimates works only when (3.11) holds; the behavior when $1/4 < m/d \le 1/2$, which was covered in the self-adjoint case, is not addressed here. This should not be a surprise because the parabolic estimates are essentially Gronwall-type estimates, which we have noted in the beginning of section 3 to give suboptimal error bounds. The main difference between the two analyses is the estimation of the forcing terms in the H^{-m} norm in the parabolic estimate case, rather than the H^{-2m} norm in the self-adjoint case. Hence, the parabolic estimates provide only upper bounds on $R[u^{*,(n)}]$ that are $\mathcal{O}\left(n^{-2m/d+1}\right)$, which is less favorable and less precise than the $o(n^{-4m/d+1})$ decay found in the self-adjoint case. Nonetheless, we conjecture that the asymptotic behavior of $R[u^{*,(n)}]$ should in principle be dominated by the self-adjoint part $\mathcal{A}_0$, even though this is not reflected with the parabolic estimates (see section 4.3).

4. Examples and simulations. The change of basis strategy is applied to some simple equations to illustrate the efficiency of the point forcing bases and cosine bases in (4.1), (4.2) for approximating the energy of the system (2.11a) or (2.11b). For the heat equation, we observe results that corroborate the analysis in section 3. We also present numerical results for convection-diffusion equations that share very similar comparative properties to the pure diffusion case and extend the discussion to the connection with the pure convection equation.

For our numerical simulations, we take the interval $U = [0, X]$, and we let $\mathcal{I}_N = \{I_i,\ i = 1, \dots, N\}$ be a uniform partition of U into intervals of length X/N. We consider the operator with $\mathcal{A} = \epsilon\Delta$ with periodic boundary conditions, whose eigenfunctions are the usual cosine basis. ϵ is a small diffusivity coefficient. The two bases on $\mathcal{S}_N := \operatorname{span}\{n_i,\ i = 1, \dots, N\}$ are the following:

1. *Point forcing basis*:

$$n_i(x) = \sqrt{\frac{N}{X}}\mathbf{1}_{I_i}(x) \quad \text{for } i = 1, \dots, N. \tag{4.1}$$

2. *Cosine basis in* $\mathcal{S}_N$: The eigenfunction basis in $L_2([0, X])$ is the usual cosine basis:

$$\begin{aligned}\mathfrak{m}_1(x) &= \sqrt{\frac{1}{X}},\\ \mathfrak{m}_i(x) &= \sqrt{\frac{2}{X}}\cos\left(\frac{(i-1)\pi x}{X}\right), \quad i = 2, 3, \dots.\end{aligned}$$

Define the *cosine basis in* $\mathcal{S}_N$ as the Gram–Schmidt orthonormalization of the L_2 projections of the first N cosine basis elements onto $\mathcal{S}_N$:

$$\begin{aligned}m_1 &= \mathfrak{m}_1,\\ m_i &= \frac{1}{Z_i}\left(\mathcal{P}_N\mathfrak{m}_i - \sum_{j=1}^{i-1}(\mathcal{P}_N\mathfrak{m}_i, m_j)m_j\right),\end{aligned} \tag{4.2}$$

where $\mathcal{P}_N$ is the L_2 projection onto $\mathcal{S}_N$ and Z_i is the normalization constant.

We study the equation

$$\frac{\partial u}{\partial t} = \epsilon\Delta u + \dot{W}_N \tag{4.3}$$

with zero initial conditions and periodic boundary conditions. The two WCEs for $\dot{W}_N$ are given in (2.9). Since $\dot{W}_N$ can also be viewed as a finite dimensional truncation of the one-dimensional white noise $\dot{W}$, we can give ξ_i and η_i precise expressions:

$$\xi_i := \int_U m_i(x)\, dW(x) \quad \text{and} \quad \eta_i := \int_U n_i(x)\, dW(x),$$

where $W(x)$ is a Brownian motion on U and from which we can check by direct computation that $\xi_i = \sum_j (m_i, n_j)\eta_j$. Equations (2.10), (2.11), and (2.12) hold.

Note that, strictly speaking, the analysis of section 3 does not apply to (4.3) because $-\Delta$ with periodic boundary conditions is not strictly positive definite—it has a zero eigenvalue. Nonetheless, we can still apply the ideas from section 3 to obtain analogous results for the error decay and 1st order improvement. Equations

(2.4)–(3.8) hold true with appropriate changes to the infinite summations, while for the point forcing basis a result analogous to Corollary 3.6 is

$$C_3 N^{-1} \leq \|\hat{v}_i\|_{L_2} \leq C_4 N^{-1}.$$

A lower bound is obtained by

$$\|\hat{v}_i\|_{L_2}^2 \geq (n_i, m_1)^2 t^2 = N^{-1} t^2.$$

For the upper bound, we integrate by parts backwards twice to find

$$(n_i, \lambda_j^{-1} m_j) = (-1)^{j-1} \lambda_j^{-1} f_i'(X) + (f_i, m_j), \quad j \geq 2, \tag{4.4}$$

for some function $f_i(x)$ such that $f_i'' = n_i$. (This step takes the place of invoking the H^{-2} norm of n_i.) It can be directly computed that

$$f_i(x) = \begin{cases} 0, & x \leq \frac{(i-1)X}{N}, \\ \frac{1}{2}\sqrt{\frac{N}{X}}\left(x - \frac{(i-1)X}{N}\right), & \frac{(i-1)X}{N} < x \leq \frac{iX}{N}, \\ \sqrt{\frac{X}{N}}\left(x + \frac{(\frac{1}{2}-i)X}{N}\right)^2, & x > \frac{iX}{N}, \end{cases}$$

so $f_i'(X) = \sqrt{X/N}$ and $\|f_i\|_{L_2}^2 = \cdots \leq \frac{17X^4}{15}\frac{1}{N}$. The upper bound follows by squaring (4.4) and summing over j.

4.1. Heat equation. For the heat equation (4.3), we show in Figure 2(a) the relative error of the truncated system under the cosine basis for different values of N. We observe that, for each n, the relative error increases pointwise to a limit as $N \to \infty$. We assume that the $N = 960$ error plot is representative of the error in the limit as $N \to \infty$, at least for n not near 960. When $n > 10$, the relative error decays linearly on the log-log axes, with a gradient of ≈ -3; i.e., $\bar{R}[u^{(n)}] \sim \mathcal{O}\left(n^{-3}\right)$. This same order of decay is seen for $\epsilon = 0.01$ only when $n > 40$ (Figure 2(b)), and the actual relative error is larger than for $\epsilon = 0.1$. Both these orders of decay are consistent with (3.5) when $m = d = 1$.

In contrast, the relative error decays linearly on the linear axes for the point forcing basis (cf. Figure 1) and does not exhibit the same limiting behavior as the error plots for the cosine basis do. In fact, in this case of periodic boundary conditions, the relative error plot is simply a straight line of slope $-N^{-1}$ joining the points $(0, 1)$ and $(N, 0)$ because the energy of each $\|\hat{v}_i\|_{L_2}^2$ is equal. For a given level of relative error and for large values of N, n_P for the point basis scales on the order of $\mathcal{O}(N)$, whereas n_E for the cosine basis scales with $\mathcal{O}(1)$. As a result, this implies the 1st order convergence seen in Table 1.

Table 1(b) shows the improvements of the cosine basis for 5% error. We highlight several trends.

(T1) For fixed N, the improvement increases for larger ϵ. This increase is most significant for large N.

(T2) We have 1*st order improvement*: doubling N increases the improvement by a factor that approaches double as N becomes large.

(T3) 1st order improvement is seen for a smaller error of 1% (data not shown), but the convergence to 1st order improvement is slower.

Table 1
Improvement n_P/n_E in the number of basis elements required to attain 5% error.

	(a) Convection-diffusion equation			(b) Heat equation	
N	$\epsilon = 0.1$	$\epsilon = 0.01$	$\epsilon = 0$	$\epsilon = 0.1$	$\epsilon = 0.01$
30	2.6364	2.4167	2.4167	1.8125	1.0741
60	4.2846	4.1429	3.8000	3.4118	1.3571
120	8.7692	7.6000	4.7917	6.3889	2.3
240	17.5385	12.6667	8.4815	12.7222	4.3019
480	35.0769	21.7619	15.7241	25.3889	8.4444
960	70.2308	43.4762	30.4333	50.6667	16.8889

Numerical scheme. The discontinuous Galerkin dG(1) scheme with a 2nd order Runge–Kutta time stepping scheme [4] was used in this computation. For each number N of forcing terms, we took N spatial grid points and used $X = 2\pi$, $T = 0.5$. The simulations were also done using a fixed number of grid points (960 grid points) for all values of N, but little difference was found in the quantitative and qualitative behaviors of the estimates.

4.2. Convection-diffusion equations. We applied the same change of basis method for the stochastic convection-diffusion equation

$$\frac{\partial}{\partial t}u + bu_x = \epsilon u_{xx} + \dot{W}_N \tag{4.5}$$

with zero initial conditions and periodic boundary conditions. We performed simulations with constant convection speed $b_0 = 1.47$ and small diffusive coefficients $\epsilon = 0.01, 0.1$.

Figure 1 shows the behavior of the relative errors of the two bases on linear axes. Under the point forcing expansion, the relative error of the truncated system decays linearly in n, whereas the relative error under the cosine expansion decays superlinearly. The improvement is also found for varying sizes of the full system, $N = 30, 60, \ldots, 960$ (Table 1(a)).

4.3. Further remarks. As noted in section 3.3, it is not straightforward to deduce precise error estimates for general equations where $\mathcal{A}$ does not provide an eigenfunction basis. If the equation is simple enough, the error decay rate can be found from the explicit solution. In the case of (4.5),

$$\begin{aligned}\frac{\partial}{\partial t}\|\hat{u}_i\|_{L_2}^2 &= \int_U 2\hat{u}_i(-b\hat{u}_{i,x} + \epsilon\hat{u}_{i,xx} + m_i)\,dx \\ &= \int_U b_x\hat{u}_i^2 + 2\epsilon\hat{u}_i\hat{u}_{i,xx} + 2\hat{u}_i m_i\,dx.\end{aligned}$$

If $\epsilon = 0$,

$$\|\hat{u}_i(t)\|_{L_2}^2 = \|\hat{u}_i(0)\|_{L_2}^2 + 2\int_0^t \int_U \hat{u}_i m_i\,dx\,dt = 2\int_0^t (\hat{u}_i(\cdot,\tau), m_i)\,d\tau,$$

so the error of each mode depends only on the coefficients $\hat{\hat{u}}_{ii}(\tau) := (\hat{u}_i(\tau), m_i)$ up to time t. By explicitly solving the convection equation,

$$\hat{\hat{u}}_{ii}(t) = \frac{X}{(i-1)\pi b}\sin\left(\frac{(i-1)\pi tb}{X}\right)$$

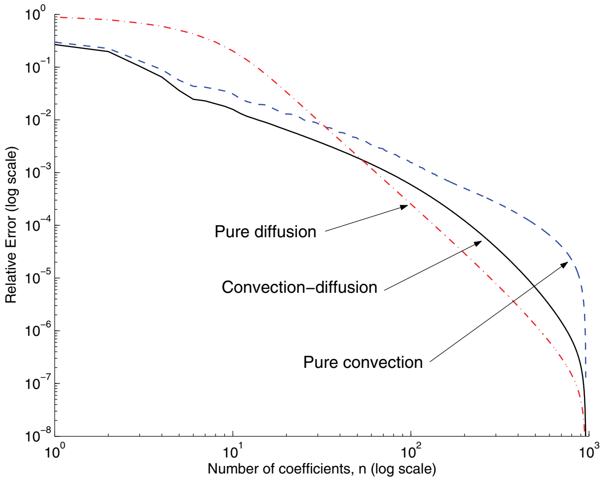

FIG. 3. *Log scale plots of the relative errors incurred by the truncation of the convection-diffusion system, as well as the pure diffusion and the pure convection systems. The convection and diffusion coefficients are $b = 6b_0$ and $\epsilon = 0.1$, respectively.*

and hence

$$\begin{aligned} R[u^{(n)}] &= \sum_{i=n+1}^{N} \|\hat{u}_i\|_{L_2}^2 = \sum_{i=n+1}^{N} 2\int_0^t \hat{\hat{u}}_{ii}^{(N)}\, d\tau \\ &= 2\sum_{i=n+1}^{N} \left(\frac{X}{(i-1)\pi b}\right)^2 \left(1 - \cos\frac{(i-1)\pi tc}{X}\right) \\ &\sim \mathcal{O}\left(\frac{1}{n} - \frac{1}{N}\right) \overset{N\to\infty}{\longrightarrow} \mathcal{O}\left(\frac{1}{n}\right). \end{aligned}$$

An approximately $\mathcal{O}\left(n^{-1}\right)$ decay for the pure diffusion case is seen in Figure 3—this is the decay rate predicted by the parabolic estimate analysis in section 3.3. If $\epsilon > 0$, the decay rate seems to be a hybrid between the convection and the diffusion parts—for small n, the $\mathcal{O}\left(n^{-1}\right)$ decay from the convection part dominates, while for large n the decay shows better agreement with the $\mathcal{O}\left(n^{-3}\right)$ decay from the diffusion part. Evidently, the analysis in section 3.3 is unable to capture the intermediate and asymptotic behaviors of the error decay for the convection-diffusion equation.

5. Conclusion. We have studied the relative efficiencies of two bases of white noise expansion and their implications for approximating the energy. The 1st order improvement in Theorem 3.1 and Corollary 3.7 has direct implications for reducing computational cost by requiring the solution of just $n \ll N$ equations. Of course, the randomization of a deterministic system of equations is also applicable to any problem where the quantity of interest is independent of basis. However, in practice, understanding the precise behavior of the error decay is still necessary for choosing the optimal truncation size. Self-adjoint problems provide an eigenfunction basis as a natural choice, from which precise error estimates can be found. For certain other general equations, the change of basis method might still be applicable with similar asymptotic behavior, but their exact analysis is more difficult.

REFERENCES

[1] M. Badieirostami, A. Adibi, H. Zhou, and S. Chow, *Efficient modeling of spatially incoherent sources based on Wiener chaos expansion method for the analysis of photonic crystal spectrometers*, Proc. SPIE Int. Soc. Opt. Eng., 6480 (2007), pp. 648018-1–8.

[2] M. Badieirostami, A. Adibi, H.-M. Zhou, and S.-N. Chow, *Wiener chaos expansion and simulations of electromagnatic wave propagation excited by a spatially incoherent source*, Multiscale Model. Simul., 8 (2010), pp. 591–604.

[3] Y. Cao, H. Yang, and L. Yin, *Finite element methods for semilinear elliptic stochastic partial differential equations*, Numer. Math., 106 (2007), pp. 181–198.

[4] B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., *Discontinuous Galerkin Methods: Theory, Computation and Applications*, Springer, Berlin, 2000.

[5] A. Debussche and J. Printems, *Weak order for the discretization of the stochastic heat equation*, Math. Comp., 78 (2009), pp. 845–863.

[6] P. Frauenfelder, C. Schwab, and R. A. Todor, *Finite elements for elliptic problems with stochastic coefficients*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 205–228.

[7] R. G. Ghanem and P. D. Spanos, *Polynomial chaos in stochastic finite elements*, J. Appl. Mech., 57 (1990), pp. 197–202.

[8] J. D. Kraus, *Electromagnetics*, McGraw–Hill, New York, 1991.

[9] S. G. Kreĭn, Ju. I. Petunin, and E. M. Semenov, *Interpolation of Linear Operators*, AMS, Providence, RI, 1982.

[10] S. V. Lototsky and B. L. Rozovskii, *Stochastic partial differential equations driven by purely spatial noise*, SIAM J. Math. Anal., 41 (2009), pp. 1295–1322.

[11] L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics*, Cambridge University Press, Cambridge, UK, 1995.

[12] B. L. Rozovskii, *Stochastic Evolution Systems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[13] M. A. Shubin, *Pseudodifferential Operators and Spectral Theory*, Springer, Berlin, 1987.

[14] R. Weissleder, *Molecular imaging in cancer*, Science, 312 (2006), pp. 1168–1171.

[15] A. Yodh and B. Chance, *Spectroscopy and imaging with diffusing light*, Phys. Today, 48 (1995), pp. 34–40.