Sparse spectral methods for solving high-dimensional and multiscale elliptic PDEs

Craig Gross[1*] and Mark Iwen[1,2]

[1]Department of Mathematics, Michigan State University, 619 Red Cedar Road, East Lansing, MI, 48824.

[2]Department of Computational Mathematics, Science and Engineering, Michigan State University, 428 S Shaw Lane, East Lansing, MI, 48824.

*Corresponding author(s). E-mail(s): grosscra@msu.edu;
Contributing authors: iwenmark@msu.edu;

Abstract

In his monograph *Chebyshev and Fourier Spectral Methods*, John Boyd claimed that, regarding Fourier spectral methods for solving differential equations, "[t]he virtues of the Fast Fourier Transform will continue to improve as the relentless march to larger and larger [bandwidths] continues" [3, pg. 194]. This paper attempts to further the virtue of the Fast Fourier Transform (FFT) as not only bandwidth is pushed to its limits, but also the dimension of the problem. Instead of using the traditional FFT however, we make a key substitution: a high-dimensional, *sparse Fourier transform* (SFT) paired with randomized rank-1 lattice methods. The resulting *sparse spectral method* rapidly and automatically determines a set of Fourier basis functions whose span is guaranteed to contain an accurate approximation of the solution of a given elliptic PDE. This much smaller, near-optimal Fourier basis is then used to efficiently solve the given PDE in a runtime which only depends on the PDE's data compressibility and ellipticity properties, while breaking the curse of dimensionality and relieving linear dependence on any multiscale structure in the original problem. Theoretical performance of the method is established herein with convergence analysis in the Sobolev norm for a general class of non-constant diffusion equations, as well as pointers to technical extensions of the convergence analysis to more general advection-diffusion-reaction equations. Numerical experiments demonstrate good empirical performance on several multiscale and high-dimensional example problems, further showcasing the promise of the proposed methods in practice.

Keywords: Spectral methods, sparse Fourier transforms, high-dimensional function approximation, elliptic partial differential equations, compressive sensing, rank-1 lattices

MSC Classification: 65N35 , 65T40 , 35J15 , 65D40 , 35J05

Communicated by Tino Ullrich.

1 Introduction

Consider as a model problem an elliptic PDE with periodic boundary conditions

$$-\nabla \cdot (a \nabla u) = f \tag{1}$$

where, for $\mathbb{T} := \mathbb{R}/\mathbb{Z}$ taken to be the one-dimensional torus, $a, f : \mathbb{T}^d \to \mathbb{R}$ are the PDE data, and $u : \mathbb{T}^d \to \mathbb{R}$ is the solution. Herein we propose a two stage method for solving such PDE. First, we use recently developed SFT methods for high-dimensional functions [26] to approximate the Fourier data of both the diffusion coefficient a and the forcing function f. So long as the PDE data, a and f, are well represented by sparse Fourier approximations, we then provide a technique for using the SFT output to find a relatively small number of Fourier coefficients that are guaranteed to reconstruct an accurate approximation of the solution u. In all, this results in a sublinear-time, curse-of-dimensionality-breaking spectral method for solving non-constant diffusion equations under periodic boundary conditions. Moreover, the technique presented is theoretically sound, with H^1 convergence guarantees provided.

These convergence guarantees hinge on a novel analysis of the Fourier-Galerkin representation of a non-constant diffusion operator where we are able to fully characterize the Fourier compressibility of the solution to (1) in terms of the Fourier compressibility of the PDE data. Additionally, we provide algorithmic improvements to the SFT developed in [26] that allow the method to run in fully sublinear-time (with respect to the size of the initial frequency set of interest). This is accompanied by new L^∞ error guarantees for this SFT which, in addition to the original L^2 guarantees, allow for the final H^1 convergence analysis of the spectral method. We also provide implementations of our methods along with various numerical experiments. Of special note, we conclude by further extending our methods beyond the simple diffusion equation (1) to also apply to multiscale and/or high-dimensional advection-diffusion-reaction equations including, e.g., the governing equations for flow dynamics in a porous medium used in hydrological modeling [42].

Solving (1) using a traditional Fourier spectral method amounts to replacing the data and the solution with their Fourier series, simplifying the left-hand side into a single Fourier series, matching the Fourier coefficients of both sides, and solving the resulting system of equations for the Fourier coefficients of u. See Section 5 for further explanation of this Galerkin formulation and the related formulations discussed below.

Two main sources of approximation error arise when implementing this technique computationally. The first is due to truncating the Fourier series involved to a finite number of terms. The second is due to numerically approximating the Fourier coefficients of the PDE data. Due to the rich theory of traditional spectral methods, these two sources of error can directly quantify the error of the resulting approximation of u.

Lemma 1 (Strang's lemma, [11]). *Let $u^{\text{truncation}}$ be the function which has the same Fourier series as u but truncated in some manner, and $a^{\text{approximate}}$*

and $f^{\text{approximate}}$ *be computed using approximations of the Fourier series of* a *and* f *truncated in the same way as* $u^{\text{truncation}}$*. Then the procedure outlined above produces a solution* u^{spectral} *which satisfies*

$$\left\|u - u^{\text{spectral}}\right\|_{H^1} \lesssim_{a,f} \left\|u - u^{\text{truncation}}\right\|_{H^1} + \left\|a - a^{\text{approximate}}\right\|_{L^\infty} + \left\|f - f^{\text{approximate}}\right\|_{L^2}$$

where the exact notion of the periodic Sobolev space H^1 *is discussed further in Section 3, and* $\lesssim_{a,f}$ *denotes an upper bound with constants that depend on the PDE data.*

This is a rough simplification of *Strang's lemma* [11], which is itself a generalization of the well-known *Céa's lemma* (the specific version of this lemma used in this paper is presented and proven in Lemma 6 below). Effectively, it states that the spectral method solution is optimal up to its Fourier series truncation and the approximation of the PDE data a and f. Thus, analyzing convergence reduces to estimating these two errors.

This outline provides the three primary ingredients for this paper:

1. a truncation method and the resulting error analysis (Section 6),
2. a (sparse) Fourier series approximation technique (Sections 7 and 8), and
3. a version of Strang's lemma that ties everything together (Section 9).

The final method is given in Algorithm 1. Its convergence guarantee in Corollary 5 shows that the error in approximating u converges like the (near-optimal) convergence rates of the SFT approximation error of a and f in addition to an exponentially decaying term related to the ellipticity properties of a.

The sections preceding the main theoretical analysis listed above include background on sparse spectral methods and motivation for our techniques (Section 2), setting the notation and PDE setup (Sections 3 and 4 respectively), and the aforementioned Galerkin formulation of our model PDE underpinning the spectral method approach (Section 5). The paper is closed with a numerics section (Section 10) describing the implementation of our technique and a variety of numerical experiments demonstrating the theory.

2 Background and motivation

We now outline some of the previous literature on spectral methods with an emphasis on exploiting sparsity. Along the way, various shortcomings will arise, and we will use these as opportunities to motivate and explain our approach in the sequel.

2.1 Convergence and computational complexity

Using a d-dimensional FFT (see, e.g., [39, Section 5.3.5] for details) to compute $a^{\text{approximate}}$ and $f^{\text{approximate}}$ in the procedure suggested in Lemma 1 naturally enforces a Fourier series truncation. A d-dimensional FFT using a tensorized grid of K uniformly spaced points in each dimension will produce approximate

Fourier coefficients indexed by frequencies in the d-dimensional hypercube on the integer lattice $\mathbb{Z}^d$ of sidelength K (note that when we refer to "bandwidth" in a multidimensional sense, we are still referring to the sidelength K of the hypercube containing these integer frequencies). The cost of each d-dimensional FFT in general requires more than K^d operations, as does the linear-system solve (in the absence of any sparsity or other tricks). Thus, not only do traditional Fourier spectral methods suffer from the curse of dimensionality, but even in moderate dimensions, multiscale problems (i.e., PDE data which require very high bandwidth to be fully resolved) can result in intractable computations.

Note that a standard FFT requires more than K^d operations in the discussion above exactly because we implicitly chose to expand our PDE data and solution with respect to an impractically huge set of K^d Fourier basis functions there. What if we instead expand all of a, f, and u in terms of the union of their individual best possible $s \ll K^d$ Fourier basis functions from this larger set? Note that doing so would automatically lead to each term on the right hand side of Lemma 1 becoming related to a nonlinear best s-term approximation error with respect to the Fourier basis in the sense of, e.g., Cohen et al [13]. Furthermore, whenever these errors decayed fast enough in s it would in fact imply that each of a, f, and u was effectively sparse/compressible in the Fourier basis, allowing the theory of compressive sensing to imply the sufficiency of a small discretization of (1). Of course, this procedure is not terribly useful in practice unless one can actually rapidly discover the best possible subset of $s \ll K^d$ Fourier basis functions for each function involved above via, e.g., compressive sensing.

A naive application of standard compressive sensing theory in pursuit of this strategy flounders in at least two ways here, however: First, though extremely successful at reducing the number of linear measurements needed in order to reconstruct a given function, standard compressive sensing recovery algorithms such as basis pursuit must still individually represent all K^d basis functions (in this simple case) during the function's numerical approximation. As a result, no dramatic runtime speedups can be expected here without additional modifications. Second, standard compressive sensing theory also generally requires direct linear measurements (in the form of, e.g., point samples) to be gathered from the function whose sparse approximation one seeks. In the case of (1) this may be trivially possible for both a and f, but is not generally possible for the a priori unknown solution u that one aims to compute (at least, not without additional innovations). Of course these difficulties can be overcome to various degrees even when using standard compressive sensing reconstruction strategies, and at least one such approach for doing so will be discussed below in Section 2.5.

In this paper, however, we instead circumvent the two difficulties mentioned above by using modified sparse Fourier transform methods. SFTs [2, 19, 20, 29, 30, 37] are compressive sensing algorithms which are highly specialized to take advantage of the number theoretic and algebraic structure of

the Fourier basis as much as possible. As a result, SFTs rarely have to consider Fourier basis functions individually during the reconstruction process, and so can simultaneously reduce both their measurement needs *and* computational complexities to effectively depend only on the number of important Fourier series coefficients in the function one aims to approximate. In the present setting, this means that SFT algorithms will run in sublinear $o(K^d)$-time, more or less automatically sidestepping the reconstruction runtime issues plaguing standard compressive sensing recovery algorithms which must represent each of the K^d-basis functions individually as they run. To circumvent the issues related to not being able to measure the solution u directly, we then use yet another approach. Instead of attempting to apply compressive sensing methods to u at all, we instead use the more easily discovered most-significant Fourier basis elements of a and f to predict in advance where the most significant Fourier basis elements of u must reside by analyzing the structure of (1). Of course, once we have discovered which Fourier basis elements are important in representing u in this fashion, standard Galerkin techniques can then be used to solve a small truncated discretization of (1) thereafter.

2.2 Prior attempts to relieve dependence on bandwidth via SFT-type methods

A key work pioneering the use of SFTs in computing solutions to PDEs is due to Daubechies, et al. [15]. This work mostly focuses on time-dependent, one-dimensional problems where the spectral scheme is formulated as alternating Fourier-projections and time-steps. Thus, there is no need to impose an a priori Fourier basis truncation on the solution. The proposed projection step instead utilizes an SFT at each time step to adaptively retain the most significant frequencies throughout the time-stepping procedure. Time-independent problems like (1) can then be handled by stepping in time until a stationary solution is obtained.

A simplified form of this algorithm is shown to succeed numerically in [15], and it is also analyzed theoretically in the case where the diffusion coefficient consists of a known, fine-scale mode superimposed over lower frequency terms. There, the Fourier-projection step can be considered to be fixed. However, removing the known fine-scale assumption leads to many difficulties, including the possibility of sparsity-induced omissions in early time steps cascading into larger errors later on. In this paper, on the other hand, we focus on the case of time-independent problems. This allows us to utilize SFTs only once initially. By doing so we avoid the possibility of SFT-induced error accumulation over many time steps. The main difficulty in our analysis then becomes determining how the Fourier-sparse representations of the PDE data discovered by high-dimensional SFTs can be used to rapidly find a suitable Fourier representation of the solution. This takes the form of mixing the Fourier supports of a and f into *stamping sets* (discussed in detail in Section 6) on which we can analyze the projection error of the solution. In fact, these stamping sets can

be viewed as a modification and generalization of the techniques used in the one-dimensional and known fine-scale analysis from [15].

2.3 Attempts to relieve the curse of dimensionality

High-dimensional PDEs are important modeling tools in many fields. Common examples include the Black-Scholes equation in mathematical finance [28], the Hamilton-Jacobi-Bellman equations in game theory and control theory [31], the Fokker-Planck equation in mathematical physics [12], and the electronic Schrödinger equation in quantum chemistry [45, 46]. The wide reach of high-dimensional PDEs has thus spurred the need for numerical methods that avoid the curse of dimensionality (see, e.g., [17] for a broad overview of modern approaches).

In the specific case of Fourier spectral methods, many attempts to overcome the curse of dimensionality have focused on using basis truncations which allow for an efficient high-dimensional Fourier transform. One of the most popular techniques is the sparse grid spectral method, which computes Fourier coefficients on the hyperbolic cross [10, 14, 22–24, 33, 43]. In general, a sparse grid method reduces the number of sampling points necessary to approximate the PDE data to $\mathcal{O}(K \log^{d-1}(K))$, where K acts as a type of bandwidth parameter. Algorithms to compute spectral representations using these sparse sampling grids run with similar complexity. When used in conjunction with spectral methods for solving PDE, these sparse grid Fourier transforms produce solution approximations with error estimates similar to the full d-dimensional FFT-versions reduced by factors only on the order of $1/\log^{d-1}(K)$.

In the context of sparse grid Fourier transforms, these methods compute Fourier coefficients with frequencies on hyperbolic crosses of similar cardinality to the number of sampling points. These hyperbolic crosses have intimate links with the space of bounded mixed derivative, in the sense that they are the optimal Fourier-approximation spaces for this class. Thus, sparse grid Fourier spectral methods are particularly apt for problems where the solution is of bounded mixed derivative, as this produces an optimal $u - u^{\text{truncation}}$ term in Lemma 1 above.

Though sparse-grid spectral methods can efficiently solve a variety of high-dimensional problems, there are clear downsides for the types of problems we target in this paper. While many problems fit the bounded mixed derivative assumption, and therefore have accurate Fourier representations on the hyperbolic cross, the multiscale, Fourier-sparse problems that we are interested are especially problematic. In fact, since a hyperbolic cross of bandwidth K contains only those frequencies $\mathbf{k} \in \mathbb{Z}^d$ with $\prod_{i=1}^{d} |k_i| = \mathcal{O}(K)$, d-dimensional frequencies active in all dimensions can have only $\|\mathbf{k}\|_\infty = \mathcal{O}(K^{1/d})$. Thus, in a multiscale problem with even one frequency that interacts in all dimensions, a hyperbolic cross is required with a bandwidth exponential in d to properly resolve the data. This then forces the traditionally curse-of-dimensionality-mitigating $\log^{d-1}(K)$ terms characteristic of sparse grid methods to be at least on the order of d^{d-1}.

2.4 More on high-dimensional Fourier transforms

As outlined in Section 2.2 above, this paper uses sparse Fourier transforms to create an adaptive basis truncation suited to the PDE data. This mimics a similar evolution in the field of high-dimensional Fourier transforms from sparse grids to more flexible techniques [16, 24, 27, 34–36, 38, 39]. In particular, the high-dimensional sparse Fourier transforms discussed in Section 7 originate from a link between early high-dimensional quadrature techniques and Fourier approximations on the hyperbolic cross [34, 35]. Instead of sampling functions on sparse grids, these methods sample high-dimensional functions along a rank-1 lattice. Rank-1 lattices are described by sampling M points in $\mathbb{T}^d$ in the direction of a generating vector $\mathbf{z} \in \mathbb{N}^d$, that is, using the sampling set

$$\Lambda(\mathbf{z}, M) := \left\{ \frac{j}{M}\mathbf{z} \bmod \mathbf{1} \mid j \in \{0, \dots, M-1\} \right\}.$$

So long as a rank-1 lattice satisfies certain properties with respect to a frequency space of interest $\mathcal{I} \in \mathbb{Z}^d$, these sampling points are sufficient to compute the Fourier coefficients of a function on $\mathcal{I}$ with a length-M univariate FFT. Though many references take $\mathcal{I}$ to be the hyperbolic cross to leverage the well-studied regularity properties and cardinality bounds similarly enjoyed in the sparse-grid literature, rank-1 lattice results are available for arbitrary frequency sets. The computationally efficient extension of these techniques via sparse Fourier transforms in [26] as well as the randomization trick presented in Section 8 take this frequency set flexibility to its limit, allowing $\mathcal{I}$ to be the a priori unknown set of the most important Fourier coefficients of the function to be approximated. This again suggests the applicability of these methods over sparse grid (or other non-sparsity exploiting) Fourier transforms in the context of multiscale problems involving even a small number of Fourier coefficients in extremely high dimensions.

2.5 Additional links to compressive sensing

As discussed above, the SFT literature overlaps considerably with the language and techniques of compressive sensing. As detailed in Section 7 below, the high-dimensional SFT we use in this paper provides error bounds with best s-term approximation, compressive-sensing-type error guarantees [13]. As a result, the Fourier coefficients of the PDE data are approximated with errors depending on the compressibility of their true Fourier series, and then the compressibility of the PDE's solution in the Fourier basis is inferred from the Fourier compressibility of the data in a direct and constructive fashion.

Another very successful line of work, however, aims to more directly apply standard compressive sensing reconstruction methods to the general spectral method framework for solving PDEs. Referred to as CORSING [4–6, 8, 9], these techniques use compressed sensing concepts to recover a sparse representation of the solution to the system of equations derived from the (Petrov-)Galerkin formulation of a PDE. These methods have been further

extended to the case of pseudospectral methods in [7], in which a simpler-to-evaluate matrix equation is subsampled and used as measurements for a compressive sensing algorithm (as an aside, [7] and discussions with the author served as a primary inspiration for this paper). This compressive spectral collocation method works by finding the largest Fourier-sine coefficients of the solution with frequencies in the integer hypercube with bandwidth K by applying Orthogonal Matching Pursuit (OMP) on a set of samples of the PDE data. By using OMP, the method is able to succeed with measurements on the order of $\mathcal{O}(d\exp(d)s\log^3(s)\log(K))$ where s is the imposed sparsity level of the solution's Fourier series. Thus, while the $\mathcal{O}(K^d)$ dependence from a traditional Fourier (pseudo)spectral method is avoided and the method adapts well to large bandwidths, the curse of dimensionality is still apparent.

In the preparation of this paper, the authors became aware of an improvement on [7] that addresses the curse of dimensionality and is therefore well-suited for similar types of problems discussed in this paper. In [44], the approach of approximating Fourier-sine coefficients on a full hypercube is replaced with approximating Fourier coefficients on a hyperbolic cross. This has the effect of converting the linear dependence on d in the sampling complexity to a $\log(d)$ due to cardinality estimates of the hyperbolic cross. However, the $\exp(d)$ term is refined using a different technique. The key theoretical ingredient for being able to apply compressive sensing to these problems is bounding the Riesz constants of the basis functions that result after applying the differential operator [8]. A careful estimation of these constants on the Fourier basis on the hyperbolic cross is able to entirely remove the exponential in d dependence, leading to a sampling complexity on the order of $\mathcal{O}(C_a s\log(d)\log^3(s)\log(K))$, where C_a involves terms depending on ellipticity and compressibility properties of a. Notably, this estimation procedure has connections to our stamping set techniques described in Section 6.

On the other hand, though focusing on the hyperbolic cross in compressive spectral collocation breaks the curse of dimensionality in the sampling complexity, the method still suffers from the inability to generalize to multiscale problems or generic frequency sets of interest like those described in 2.3. Additionally, as previously mentioned in this section, the compressive-sensing algorithm used for recovery (in this case OMP) suffers from a computational complexity on the order of the cardinality of the truncation set of interest. For the hyperbolic cross, this can still be, at worst, exponential in d. Finally, the error estimates are presented in terms of the compressibility of the Fourier series of the solution u, which may not be known a priori from the PDE data. We expect that there may be some way to link our stamping theory and convergence estimates with the compressive sensing theory to refine and generalize both approaches.

3 Notation

Define the one-dimensional torus to be $\mathbb{T} := \mathbb{R}/\mathbb{Z}$. Unless otherwise stated, all functions are complex-valued and defined on the torus $\mathbb{T}^d$. For example, we take the inner product for $u, v \in L^2 := L^2(\mathbb{T}^d; \mathbb{C})$ to be

$$\langle u, v \rangle_{L^2} := \int_{\mathbb{T}^d} u(\mathbf{x}) \overline{v}(\mathbf{x}) \, d\mathbf{x}.$$

Additionally, unless otherwise stated, all multiindexed infinite sequences are complex-valued and indexed on $\mathbb{Z}^d$. For example, we take the inner product for $\hat{u}, \hat{v} \in \ell^2 := \ell^2(\mathbb{Z}^d; \mathbb{C})$ to be

$$\langle \hat{u}, \hat{v} \rangle_{\ell^2} := \sum_{\mathbf{k} \in \mathbb{Z}^d} \hat{u}_{\mathbf{k}} \overline{\hat{v}}_{\mathbf{k}}.$$

All finite length vectors/tensors will be denoted in boldface and when required, will be implicitly extended to larger index sets by taking on the value zero wherever they are not originally defined. We also denote the complex-valued finite-length vectors or infinite-length sequences supported on a set $\mathcal{D}$ as $\mathbb{C}^{\mathcal{D}}$. Since sparse approximations will be an important tool in our final algorithm, we also define the best s-term approximation of a sequence $\hat{u}$ as $\hat{u}$ restricted to its s largest magnitude entries and denote this as $\hat{u}_s^{\text{opt}}$.

We now define periodic Sobolev spaces (see also [4, Section 2.1] and [33, Appendix A.2.2]).

Definition 1. For $u \in L^2$ and $\alpha \in \mathbb{N}_0^d$ a multiindex, if there exists a $v \in L^2$ such that

$$\langle v, \phi \rangle_{L^2} = (-1)^{|\alpha|} \langle u, \partial^\alpha \phi \rangle_{L^2} \qquad \text{for all } \phi \in C^\infty \subseteq L^2,$$ [1]

we call v the *weak α derivative of u*, and write $\partial^\alpha u := v$. We define the inner product

$$\langle u, v \rangle_{H^1} := \langle u, v \rangle_{L^2} + \int_{\mathbb{T}^d} \nabla u(\mathbf{x}) \cdot \overline{\nabla v}(\mathbf{x}) \, d\mathbf{x},$$

(where all derivatives are taken in the weak sense) and have the associated norm $\|u\|_{H^1} := \sqrt{\langle u, u \rangle_{H^1}}$. The *periodic Sobolev space* H^1 is defined as $H^1 := \{u \in L^2 \mid \|u\|_{H^1} < \infty\}$.

In order to set our notation for Fourier coefficients and series, we first note the density of trigonometric monomials in L^2 and H^1.

Theorem 1. *The space of all infinitely differentiable periodic functions C^∞ is dense in L^2 and H^1. In particular, space of trigonometric monomials $\{e_{\mathbf{k}}(\mathbf{x}) :=$*

[1] Here $C^\infty := \left\{\phi : \mathbb{T}^d \to \mathbb{C} \mid \partial^\alpha \phi \text{ is continuous } \forall \alpha \in \mathbb{N}_0^d\right\}$.

$e^{2\pi i \mathbf{k}\cdot\mathbf{x}} \in C^\infty \mid k \in \mathbb{Z}^d\}$ *is a basis for* C^∞*, an orthonormal basis for* L^2*, and an orthogonal basis for* H^1*.*

Definition 2. For any $u \in L^1$, and any $\mathbf{k} \in \mathbb{Z}^d$, we define the $\mathbf{k}$*th Fourier coefficient*

$$\hat{u}_\mathbf{k} = \langle u, e_\mathbf{k} \rangle_{L^2} = \int_{\mathbb{T}^d} u(\mathbf{x}) e^{-2\pi i \mathbf{k}\cdot\mathbf{x}} \, d\mathbf{x}.$$

If $u \in L^2$, the orthonormality of the trigonometric monomials in Theorem 1 allows us to write the *Fourier series for* u,

$$u(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{Z}^d} \hat{u}_\mathbf{k} e_\mathbf{k}(\mathbf{x}).$$

We also note the well-known Parseval–Plancherel identity for use later.

Proposition 1 (Parseval–Plancherel identity)**.** *If* $u \in L^2$*, then* $\hat{u} \in \ell^2$ *with* $\|u\|_{L^2} = \|\hat{u}\|_{\ell^2}$*. If* $v \in L^2$*, then* $\langle u, v\rangle_{L^2} = \langle \hat{u}, \hat{v}\rangle_{\ell^2}$*.*

Definition 3. We additionally define the *mean-zero periodic Sobolev space* H as $H^1/\mathbb{R}$ where the representative u is chosen so that $\hat{u}_\mathbf{0} = 0$, endowed with the inner product[2]

$$\langle u, v\rangle_H := \int_{\mathbb{T}^d} \nabla u(\mathbf{x}) \cdot \overline{\nabla v}(\mathbf{x}) \, d\mathbf{x}.$$

In the sequel, we will often consider restrictions in frequency space denoted by, e.g., $\hat{u}|_\mathcal{D}$, where $\mathcal{D} \subseteq \mathbb{Z}^d$. We will simultaneously consider this to be an element of $\mathbb{C}^\mathcal{D}$ and a complex valued sequence on $\mathbb{Z}^d$ with zero entries on $\mathbb{Z}^d \setminus \mathcal{D}$. When $\hat{u}$ represents the Fourier coefficients of a function u, we define the associated restriction

$$u|_\mathcal{D} := \sum_{\mathbf{k}\in\mathbb{Z}^d} (\hat{u}|_\mathcal{D})_\mathbf{k} \, e_\mathbf{k} = \sum_{\mathbf{k}\in\mathcal{D}} \hat{u}_\mathbf{k} e_\mathbf{k},$$

where the fact that $\mathcal{D} \subseteq \mathbb{Z}^d$ is treated as a set of frequencies indicates that we are restricting u in frequency, not space. Given a hatted sequence $\hat{v}$ or vector $\hat{\mathbf{v}}$, the associated function with Fourier series $\sum_{\mathbf{k}\in\mathbb{Z}^d} \hat{v}_\mathbf{k} e_\mathbf{k}$ will always be implicitly labeled using the non-hatted, roman font letter (in this example, v).

4 Elliptic PDE setup

We begin with a model elliptic partial differential equation.

[2] note that by Proposition 1, $\langle u, v\rangle_H \simeq \langle u, v\rangle_{H^1}$ for $u, v \in H$.

Definition 4. For some $a : \mathbb{T}^d \to \mathbb{R}$ sufficiently smooth, define the *linear, elliptic partial differential operator in divergence form* $\mathcal{L}[a] : C^2 \to C^0$ by

$$\mathcal{L}[a]u = -\nabla \cdot (a\nabla u)\,.$$

If for some $f : \mathbb{T}^d \to \mathbb{R}$ sufficiently smooth, $u \in C^2$ satisfies

$$\mathcal{L}[a]u = f, \tag{SF}$$

we say that u *solves the given elliptic PDE with periodic boundary conditions in the strong form.*

Now, after multiplying by the complex conjugate of a test function $v \in H^1(\mathbb{T}^d)$ and integrating by parts, we define the bilinear form associated to $\mathcal{L}[a]$ as $\mathfrak{L}[a] : H^1 \times H^1 \to \mathbb{C}$ with

$$\mathfrak{L}[a](u, v) := \int_{\mathbb{T}^d} a(\mathbf{x})\nabla u(\mathbf{x}) \cdot \overline{\nabla v}(\mathbf{x})\, d\mathbf{x},$$

and we say that $u \in H^1$ *solves the given elliptic PDE with periodic boundary conditions in the weak form* if

$$\mathfrak{L}[a](u, v) = \langle f, v\rangle_{L^2} \quad \text{for all } v \in H^1. \tag{WF}$$

For our purposes, we will take $a \in L^\infty(\mathbb{T}^d;\ \mathbb{R})$, and $f \in L^2(\mathbb{T}^d;\ \mathbb{R})$.

By the conditions specified in the Lax-Milgram theorem (see, e.g., [18]), we are guaranteed that a unique mean-zero solution to (WF) exists so long as the right-hand side and test space is also mean-zero. See [4, Proposition 2.1] for a more specific formulation in our setting and its proof.

Proposition 2. *For $a \in L^\infty(\mathbb{T}^d;\ \mathbb{R})$, $\mathfrak{L}[a]$ is continuous with continuity constant $\beta \le \|a\|_{L^\infty}$, that is*

$$|\mathfrak{L}[a](u, v)| \le \beta\|u\|_H\|v\|_H \qquad \text{for all } u, v \in H. \tag{2}$$

Additionally, if $a(\mathbf{x}) \ge a_{\min} > 0$ a.e. on $\mathbb{T}^d$, then $\mathfrak{L}[a]$ is also coercive with coercivity constant $\alpha \ge a_{\min}$, that is

$$|\mathfrak{L}[a](u, u)| \ge \alpha\|u\|_H^2 \qquad \text{for all } u \in H. \tag{3}$$

Under conditions (2) and (3), if $f \in L^2(\mathbb{T}^d;\ \mathbb{R})$ is mean-zero, that is, $\hat{f}_0 = 0$, then (WF) has unique, mean-zero solution $u \in H$ satisfying

$$\|u\|_H \le \frac{\|f\|_{L^2}}{\alpha}. \tag{4}$$

5 Galerkin spectral methods

By Theorem 1 one may rewrite the weak PDE (WF) in the new form

$$\mathfrak{L}[a](u, e_{\mathbf{k}}) = \langle f, e_{\mathbf{k}} \rangle_{L^2} =: \hat{f}_{\mathbf{k}} \quad \text{for all } \mathbf{k} \in \mathbb{Z}^d.$$

Now rewriting the bilinear form on the left-hand side and using the Fourier series representations of a and u, we obtain

$$\begin{aligned}
\mathfrak{L}[a](u, e_{\mathbf{k}}) &= \sum_{\mathbf{l}_1, \mathbf{l}_2 \in \mathbb{Z}^d} \hat{a}_{\mathbf{l}_1} \hat{u}_{\mathbf{l}_2} \int_{\mathbb{T}^d} e_{\mathbf{l}_1}(\mathbf{x}) \nabla e_{\mathbf{l_2}}(\mathbf{x}) \cdot \overline{\nabla e_{\mathbf{k}}}(\mathbf{x})\, d\mathbf{x} \\
&= \sum_{\mathbf{l}_1, \mathbf{l}_2 \in \mathbb{Z}^d} (2\pi)^2 (\mathbf{l_2} \cdot \mathbf{k}) \hat{a}_{\mathbf{l}_1} \hat{u}_{\mathbf{l}_2} \delta_{\mathbf{l}_1, \mathbf{k} - \mathbf{l}_2} \\
&= \sum_{\mathbf{l} \in \mathbb{Z}^d} (2\pi)^2 (\mathbf{l} \cdot \mathbf{k}) \hat{a}_{\mathbf{k}-\mathbf{l}} \hat{u}_{\mathbf{l}} \\
&=: (L[\hat{a}]\hat{u})_{\mathbf{k}},
\end{aligned}$$

where $L[\hat{a}]$ is an operator in ℓ^2. This leads to the *Galerkin form* of our PDE,

$$L[\hat{a}]\hat{u} = \hat{f}. \tag{GF}$$

The computational advantages of (GF) are clear. By numerically approximating $\hat{a}$ and $\hat{f}$ (thereby also truncating $L[\hat{a}]$), we arrive at a discretized, finite system of equations that can be solved for the Fourier coefficients of our solution.

We will use a fast sparse Fourier transform (SFT) for functions of many dimensions to approximate our PDE data which then leads to a sparse system of equations that we can quickly solve to approximate $\hat{u}$. This SFT will use the values of a and f at equispaced nodes on a randomized rank-1 lattice in $\mathbb{T}^d$, and therefore, our technique is effectively a pseudospectral method where the discretization of the solution space $\{\hat{u} \mid u \in H\}$ is adapted to the PDE data.

Before we move to the detailed discussion of this SFT, we provide a more detailed analysis of the Galerkin operator in Section 6 to help us analyze the resulting spectral method. But first, we note that $L[\hat{a}]$ also captures the behavior of $\mathfrak{L}[a]$ as a bilinear form.

Proposition 3. *For $\hat{u}, \hat{v} \in \ell^2$ with $u, v \in H$,*

$$\mathfrak{L}[a](u, v) = \langle L[\hat{a}]\hat{u}, \hat{v} \rangle_{\ell^2}.$$

Proof By the Fourier series representation of v,

$$\mathfrak{L}[a](u, v) = \sum_{\mathbf{k} \in \mathbb{Z}^d} \mathfrak{L}[a](u, e_{\mathbf{k}}) \overline{\hat{v}}_{\mathbf{k}} = \sum_{\mathbf{k} \in \mathbb{Z}^d} (L[\hat{a}]\hat{u})_{\mathbf{k}}\, \overline{\hat{v}}_{\mathbf{k}} = \langle L[\hat{a}]\hat{u}, \hat{v} \rangle_{\ell^2}.$$

□

6 Stamping sets and truncation analysis

Notably, (GF) gives us insight into the frequency support of $\hat{u}$. The structure outlined in the following proposition is crucial in constructing a fast spectral method that exploits Fourier-sparsity.

Proposition 4. *For any set $F \subseteq \mathbb{Z}^d$ and $N \in \mathbb{N}_0$, recursively define the sets*

$$\begin{aligned} \mathcal{S}^N[\hat{a}](F) &:= \begin{cases} F & \text{if } N = 0 \\ \mathcal{S}^{N-1}[\hat{a}](F) + \operatorname{supp}(\hat{a}) & \text{if } N > 0 \end{cases}, \\ \mathcal{S}^\infty[\hat{a}](F) &:= \bigcup_{N=0}^{\infty} \mathcal{S}^N[\hat{a}](F), \end{aligned} \tag{5}$$

where here, addition is the Minkowski sum of sets. Under the conditions of Proposition 2, $\operatorname{supp}(\hat{u}) \subseteq \mathcal{S}^\infty[\hat{a}](\operatorname{supp}(\hat{f}))$.

Proof The fact that a is strictly positive implies that $\hat{a}_{\mathbf{0}} \neq 0$, and the fact that a is real implies $\operatorname{supp}(\hat{a}) = -\operatorname{supp}(\hat{a})$. Now, for any $\mathbf{k} \in \mathbb{Z}^d \setminus \{\mathbf{0}\}$, we may rearrange the equality $(L[\hat{a}]\hat{u})_{\mathbf{k}} = \hat{f}_{\mathbf{k}}$ to obtain

$$\begin{aligned} \hat{u}_{\mathbf{k}} &= \frac{\hat{f}_{\mathbf{k}} - \sum_{\mathbf{l} \in (\{\mathbf{k}\} + \operatorname{supp}(\hat{a})) \setminus \{\mathbf{k}\}} (2\pi)^2 (\mathbf{l} \cdot \mathbf{k}) \hat{a}_{\mathbf{k}-\mathbf{l}} \hat{u}_{\mathbf{l}}}{(2\pi)^2 (\mathbf{k} \cdot \mathbf{k}) \hat{a}_{\mathbf{0}}} \\ &= \frac{\hat{f}_{\mathbf{k}} - \sum_{\mathbf{l} \in \operatorname{supp}(\hat{a}) \setminus \{\mathbf{0}\}} (2\pi)^2 (\mathbf{k} \cdot \mathbf{k} - \mathbf{l} \cdot \mathbf{k}) \hat{a}_{\mathbf{l}} \hat{u}_{\mathbf{k}-\mathbf{l}}}{(2\pi)^2 (\mathbf{k} \cdot \mathbf{k}) \hat{a}_{\mathbf{0}}}. \end{aligned}$$

Thus, $\hat{u}_{\mathbf{k}}$ explicitly depends only on the values of $\hat{u}$ on $\mathcal{S}^1[\hat{a}](\{\mathbf{k}\}) \setminus \{\mathbf{k}\}$, which themselves then depend only on values of $\hat{u}$ on $\mathcal{S}^2[\hat{a}](\{\mathbf{k}\})$, and so on. This decouples the system of equations $L[\hat{a}]\hat{u}$ into a disjoint collection of systems of equations, one for each class of frequencies $\mathcal{S}^\infty[\hat{a}](\{\mathbf{k}\})$. Since Proposition 2 implies that $\hat{v} = 0$ is the unique solution of $L[\hat{a}]\hat{v} = 0$, the unique solution of the system of equations for $\hat{u}$ on $\mathcal{S}^\infty[\hat{a}](\{\mathbf{k}\})$ for any $\mathbf{k} \notin \operatorname{supp}(\hat{f})$ is $\hat{u}|_{\mathcal{S}^\infty[\hat{a}](\{\mathbf{k}\})} = 0$. Therefore, $\operatorname{supp} \hat{u} \subseteq \mathcal{S}^\infty[\hat{a}](\operatorname{supp}(\hat{f}))$ as desired. □

In what follows, when the set F and Fourier coefficients $\hat{a}$ are clear from context, we suppress them in the notation given by (5) so that $\mathcal{S}^N := \mathcal{S}^N[\hat{a}](F)$. Intuitively, we can imagine constructing $\mathcal{S}^N$ by first creating a "rubber stamp" in the shape of $\operatorname{supp}(\hat{a})$. This rubber stamp is then stamped onto every frequency in $F =: \mathcal{S}^0$ to construct $\mathcal{S}^1$. Then, this process is repeated, stamping each element of $\mathcal{S}^1$ to produce $\mathcal{S}^2$, and so on. For this reason, we will colloquially refer to these as "stamping sets." Figure 1 gives an example of this stamping procedure for $d = 2$.

Remark 1. Note that the inclusion $\operatorname{supp}(\hat{u}) \subseteq \mathcal{S}^\infty$ can be shown to be sharp using more direct applications of Fourier series. Indeed, consider an ODE of the form (SF) with

$$a(x) = 3 + 2\cos(2\pi k_a x), \quad f(x) = \sin(2\pi k_f x),$$

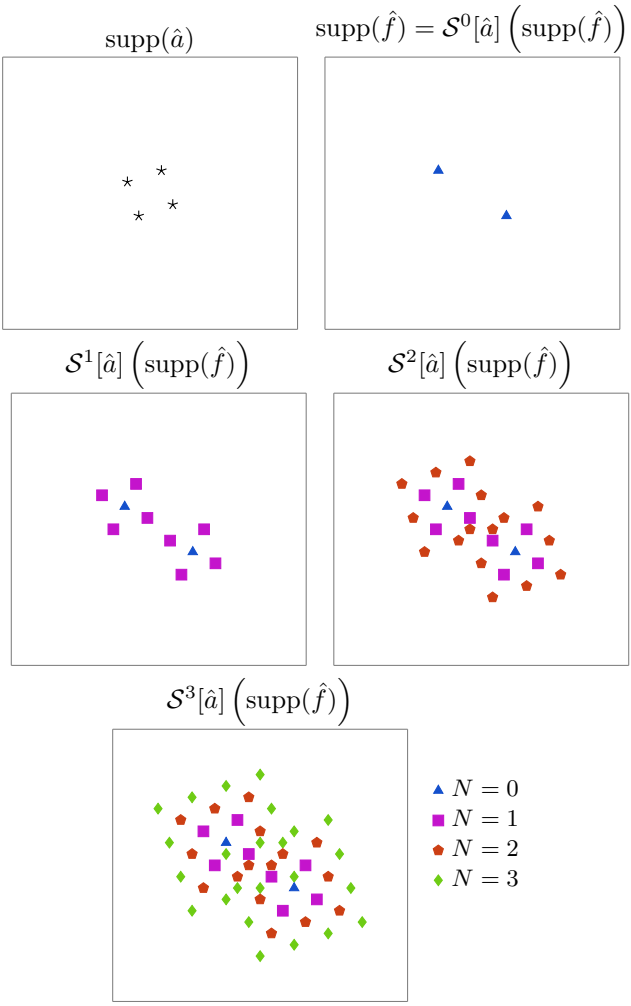


Fig. 1: New frequencies in each stamping level up to $N = 3$ where $N = 0$ is $\mathrm{supp}(\hat{f})$.

with $k_a, k_f \in \mathbb{N}$. Letting $g(x) = 1/a(x)$, taking Fourier series of both sides of (SF), solving for $\hat{u}$, and simplifying with the convolution theorem implies $\mathrm{supp}(\hat{u}) = \mathrm{supp}(\hat{f} * \hat{g})$, where $*$ is the discrete convolution. It can be shown by directly computing a complex contour integral that $\hat{g}_{k_a n} \neq 0$ for all $n \in \mathbb{Z}$. Thus, $\hat{f} * \hat{g}$ (and therefore $\hat{u}$) is supported on $k_f + k_a\mathbb{Z} = \mathcal{S}^\infty$.

On the other hand, there are cases when $\mathrm{supp}(\hat{u})$ is strictly contained in $\mathcal{S}^\infty$. Consider the case where we choose some Fourier sparse a and u, and let $\mathrm{supp}(L[\hat{a}]\hat{u}) =: \hat{f}$. By construction, u solves (GF) with a and f as PDE data. The proof of Proposition 5 below, however, implies that $\hat{f}$ is also sparse and

therefore $\left|\mathcal{S}^N\right| < \infty$ for any stamping level N. Since $|\mathrm{supp}(\hat{u})|$ is also finite, there must be some minimal N such that $\mathrm{supp}(\hat{u}) \subseteq \mathcal{S}^N \subsetneq \mathcal{S}^\infty$.

A key approach of our further analysis will be analyzing the decay of $\hat{u}$ on successive stamping levels. The stamping level will become the driving parameter in the spectral method rather than bandwidth in a traditional spectral method. Before moving onto this analysis however, we provide an upper bound for the cardinality of the stamping sets. This will ultimately be used to upper bound the computational complexity of our technique. The proof of this bound is given in Appendix A.

Lemma 2. *Suppose that* $\mathbf{0} \in \mathrm{supp}(\hat{a})$, $\mathrm{supp}(\hat{a}) = -\mathrm{supp}(\hat{a})$, *and* $|\mathrm{supp}(\hat{a})| = s$. *Then*

$$\left|\mathcal{S}^N[\hat{a}](\mathrm{supp}(\hat{f}))\right| \leq 6\left|\mathrm{supp}(\hat{f})\right| \left(\frac{\max(s, 2N)}{\sqrt{2}}\right)^{\min(s,2N)}. \tag{6}$$

Proposition 4 gives us a natural way to consider truncations of the solution u in frequency space. We will use these truncations to discretize the Galerkin formulation (GF) in Section 9 below. In order to analyze the error in the resulting spectral method algorithm, we will need quantitative bounds on how the solution decays outside of the frequency sets $\mathcal{S}^N := \mathcal{S}^N[\hat{a}](\mathrm{supp}(\hat{f}))$. For $\mathcal{S}^N$ to be finite, we assume in this section that $\mathrm{supp}\,\hat{a}$ and $\mathrm{supp}\,\hat{f}$ are finite. This assumption will be lifted later via Lemma 5.

We begin with a technical result regarding the interplay between $L[\hat{a}]$ and the supports of vectors that it acts on.

Proposition 5. *For any* $\hat{v}$ *with* $\mathrm{supp}(\hat{v}) \subseteq \mathcal{S}^n \setminus \mathcal{S}^{n-1}$, $\mathrm{supp}(L[\hat{a}]\hat{v}) \subseteq \mathcal{S}^{n+1} \setminus \mathcal{S}^{n-2}$.

Proof For any $\mathbf{k} \in \mathbb{Z}^d$, consider

$$\begin{aligned}(L[\hat{a}]\hat{v})_{\mathbf{k}} &= \sum_{\mathbf{l}\in\mathbb{Z}^d} (2\pi)^2(\mathbf{l}\cdot\mathbf{k})\hat{a}_{\mathbf{k}-\mathbf{l}}\hat{v}_{\mathbf{l}} \\ &= \sum_{\mathbf{l}\in(\{\mathbf{k}\}-\mathrm{supp}(\hat{a}))\cap\mathrm{supp}(\hat{v})} (2\pi)^2(\mathbf{l}\cdot\mathbf{k})\hat{a}_{\mathbf{k}-\mathbf{l}}\hat{v}_{\mathbf{l}} \\ &= \sum_{\mathbf{l}\in(\{\mathbf{k}\}-\mathrm{supp}(\hat{a}))\cap(\mathcal{S}^n\setminus\mathcal{S}^{n-1})} (2\pi)^2(\mathbf{l}\cdot\mathbf{k})\hat{a}_{\mathbf{k}-\mathbf{l}}\hat{v}_{\mathbf{l}}.\end{aligned}$$

This sum is nonempty only if $\mathbf{k}$ is such that there exists $\mathbf{l} \in \mathcal{S}^n \setminus \mathcal{S}^{n-1}$ and $\mathbf{k}_a^* \in \mathrm{supp}(\hat{a})$ with $\mathbf{k} = \mathbf{l} + \mathbf{k}_a^*$. By definition of $\mathbf{l} \in \mathcal{S}^n \setminus \mathcal{S}^{n-1}$, n is the minimal number such that

$$\mathbf{l} = \mathbf{k}_f + \sum_{m=1}^{n} \mathbf{k}_a^m, \text{ where } \mathbf{k}_f \in \mathrm{supp}(\hat{f}),\ \mathbf{k}_a^m \in \mathrm{supp}(\hat{a}) \text{ for all } m = 1, \ldots, n$$

holds. In particular, this implies that $\mathbf{k}_a^m \neq \mathbf{0}$ for all $m = 1, \ldots, n$.

There are now two cases. First, if $\mathbf{k}_a^* = -\mathbf{k}_a^m$ for any m, $\mathbf{k} = \mathbf{l} + \mathbf{k}_a^* \in \mathcal{S}^{n-1}\setminus\mathcal{S}^{n-2}$, and the proposition is satisfied. On the other hand, we consider the case when $\mathbf{k}_a^*$

does not negate any $\mathbf{k}_a^m$ involved in the sum equalling $\mathbf{l}$. If $\mathbf{k}_a^* = \mathbf{0}$, then clearly $\mathbf{k} = \mathbf{l} \in \mathcal{S}^n \setminus \mathcal{S}^{n-1}$. In any other case, we represent

$$\mathbf{k} = \mathbf{k}_f + \sum_{m=1}^{n} \mathbf{k}_a^m + \mathbf{k}_a^* =: \mathbf{k}_f + \sum_{m=1}^{n+1} \mathbf{k}_a^m,$$

where $n+1$ is the smallest number for which this holds. Thus, $\mathbf{k} \in \mathcal{S}^{n+1} \setminus \mathcal{S}^n$. Altogether then, the only possible $\mathbf{k}$ values such that the sum is nonzero are those in $\mathcal{S}^{n+1} \setminus \mathcal{S}^{n-2}$, completing the proof. □

Noting that $\operatorname{supp}(L[\hat{a}]\hat{u}) = \operatorname{supp}(\hat{f})$, we observe the following interesting relationship between the values of $\hat{u}$ on neighboring stamping levels. Below, to simplify notation, for all $m, n \in \mathbb{N}_0$, we set

$$b_{m,n} := \langle L[\hat{a}]\hat{u}_{\mathcal{S}^m \setminus \mathcal{S}^{m-1}}, \hat{u}_{\mathcal{S}^n \setminus \mathcal{S}^{n-1}} \rangle_{\ell^2},$$

with the convention that $\mathcal{S}^{-1} = \emptyset$.

Corollary 1. *For all $n \in \mathbb{N}_0$,*

$$b_{n+1,n} + b_{n,n} + b_{n-1,n} = \begin{cases} \langle \hat{f}, \hat{u}|_{\mathcal{S}^0} \rangle_{\ell^2} & \text{if } n = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Proof By Proposition 5, $\hat{u}|_{\mathcal{S}^n \setminus \mathcal{S}^{n-1}}$ is ℓ^2-orthogonal to $L[\hat{a}]\hat{u}|_{\mathcal{S}^m \setminus \mathcal{S}^{m-1}}$ for all $m \notin \{n-1, n, n+1\}$. In our simplified notation, $b_{m,n} = 0$ for all $m \notin \{n-1, n, n+1\}$. Thus

$$\langle \hat{f}, \hat{u}|_{\mathcal{S}^n \setminus \mathcal{S}^{n-1}} \rangle_{\ell^2} = \langle L[\hat{a}]\hat{u}, \hat{u}|_{\mathcal{S}^n \setminus \mathcal{S}^{n-1}} \rangle_{\ell^2} = \sum_{m=0}^{\infty} b_{m,n} = b_{n+1,n} + b_{n,n} + b_{n-1,n}.$$

The proof is finished by noting that

$$\langle \hat{f}, \hat{u}|_{\mathcal{S}^n \setminus \mathcal{S}^{n-1}} \rangle_{\ell^2} = \begin{cases} \langle \hat{f}, \hat{u}|_{\mathcal{S}^0} \rangle & \text{if } n = 0 \\ 0 & \text{otherwise} \end{cases},$$

which follows from the definition of $\mathcal{S}^n$ in (5). □

We are now ready to estimate $\hat{u}|_{\mathcal{S}^n \setminus \mathcal{S}^{n-1}}$ in terms of its neighbors $\hat{u}|_{\mathcal{S}^{n+1} \setminus \mathcal{S}^n}$ and $\hat{u}|_{\mathcal{S}^{n-1} \setminus \mathcal{S}^{n-2}}$. The standard approach would be to use a combination of coercivity and continuity (see, e.g., the proof of Lemma 6 or [11, Section 6.4] for other examples): where α and β are respectively the coercivity and continuity constants in Propositon 2, for $n > 0$,

$$\begin{aligned} \alpha \left\| u|_{\mathcal{S}^n \setminus \mathcal{S}^{n-1}} \right\|_H^2 &\le |b_{n,n}| \\ &\le |b_{n+1,n}| + |b_{n-1,n}| \\ &\le \beta \left\| u|_{\mathcal{S}^n \setminus \mathcal{S}^{n-1}} \right\|_H \left(\left\| u|_{\mathcal{S}^{n+1} \setminus \mathcal{S}^n} \right\|_H + \left\| u|_{\mathcal{S}^{n-1} \setminus \mathcal{S}^{n-2}} \right\|_H \right), \end{aligned}$$

and we obtain

$$\left\|u|_{\mathcal{S}^n\setminus\mathcal{S}^{n-1}}\right\|_H \leq \frac{\beta}{\alpha}\left(\left\|u|_{\mathcal{S}^{n+1}\setminus\mathcal{S}^n}\right\|_H + \left\|u|_{\mathcal{S}^{n-1}\setminus\mathcal{S}^{n-2}}\right\|_H\right).$$

However, we will hope to iterate this bound, and the fact that $\beta \geq \alpha$ will not allow for us to show any decay as $n \to \infty$. Thus, we require a slightly subtler estimate than simply using continuity.

Proposition 6. *For $n > 0$, we have*

$$|b_{n\pm1,n}| \leq \|a - \hat{a}_{\mathbf{0}}\|_{L^\infty}\left\|u|_{\mathcal{S}^n\setminus\mathcal{S}^{n-1}}\right\|_H\left\|u|_{\mathcal{S}^{n\pm1}\setminus\mathcal{S}^{n\pm1-1}}\right\|_H.$$

Proof Restricting all sums to the support of the vectors they index, we have

$$b_{n\pm1,n} = \sum_{\mathbf{k}\in\mathcal{S}^n\setminus\mathcal{S}^{n-1}}\sum_{\mathbf{l}\in(\mathbf{k}-\operatorname{supp}(\hat{a}))\cap(\mathcal{S}^{n\pm1}\setminus\mathcal{S}^{n\pm1-1})}(2\pi)^2(\mathbf{l}\cdot\mathbf{k})\hat{a}_{\mathbf{k}-\mathbf{l}}\hat{u}_{\mathbf{l}}\overline{\hat{u}}_{\mathbf{k}}.$$

Clearly, choosing $\mathbf{l} = \mathbf{k} \in \mathcal{S}^n\setminus\mathcal{S}^{n-1}$ would not allow for $\mathbf{l} \in \mathcal{S}^{n\pm1}\setminus\mathcal{S}^{n\pm1-1}$. Thus, no term multiplying $\hat{a}_{\mathbf{k}-\mathbf{k}} = \hat{a}_{\mathbf{0}}$ will appear in this sum. We then have the equivalence

$$b_{n\pm1,n} = \langle L[\hat{a} - \hat{a}_{\mathbf{0}}]\hat{u}|_{\mathcal{S}^{n\pm1}\setminus\mathcal{S}^{n\pm1-1}}, \hat{u}|_{\mathcal{S}^n\setminus\mathcal{S}^{n-1}}\rangle_{\ell^2},$$

which by the standard argument for continuity, implies

$$|b_{n\pm1,n}| \leq \|a - \hat{a}_{\mathbf{0}}\|_{L^\infty}\left\|u|_{\mathcal{S}^n\setminus\mathcal{S}^{n-1}}\right\|_H\left\|u|_{\mathcal{S}^{n\pm1}\setminus\mathcal{S}^{n\pm1-1}}\right\|_H,$$

as desired. □

The same argument preceding Proposition 6 then gives the desired "neighbor" estimate.

Corollary 2. *For all $n > 1$,*

$$\left\|u|_{\mathcal{S}^n\setminus\mathcal{S}^{n-1}}\right\|_H \leq \frac{\|a - \hat{a}_{\mathbf{0}}\|_{L^\infty}}{a_{\min}}\left(\left\|u|_{\mathcal{S}^{n+1}\setminus\mathcal{S}^n}\right\|_H + \left\|u|_{\mathcal{S}^{n-1}\setminus\mathcal{S}^{n-2}}\right\|_H\right).$$

We now have the pieces to state an estimate of the truncation error.

Lemma 3. *Let a, f, and u be as in Proposition 2. Assume*

$$3\|a - \hat{a}_{\mathbf{0}}\|_{L^\infty} < a_{\min} \tag{7}$$

Then

$$\|u - u|_{\mathcal{S}^N}\|_H \leq \left(\frac{\|a - \hat{a}_{\mathbf{0}}\|_{L^\infty}}{a_{\min} - 2\|a - \hat{a}_{\mathbf{0}}\|_{L^\infty}}\right)^{N+1}\frac{\|f\|_{L^2}}{a_{\min}}.$$

Proof We begin by breaking $\operatorname{supp}(\hat{u}) \setminus \mathcal{S}^N$ into $\bigcup_{n=N+1}^{\infty} \left(\mathcal{S}^n \setminus \mathcal{S}^{n-1}\right)$, the sets of new contributions (which holds due to Proposition 4). Thus

$$\|u - u|_{\mathcal{S}^N}\|_H \leq \sum_{n=N+1}^{\infty} \left\|u|_{\mathcal{S}^n \setminus \mathcal{S}^{n-1}}\right\|_H =: T_N.$$

Applying the neighbor bound, Corollary 2, (where we define $A := \|a - \hat{a}_{\mathbf{0}}\|_{L^\infty}/a_{\min}$), we have

$$\begin{aligned} T_N &\leq A \left(\sum_{n=N+1}^{\infty} \left\|u|_{\mathcal{S}^{n+1} \setminus \mathcal{S}^n}\right\|_H + \sum_{n=N+1}^{\infty} \left\|u|_{\mathcal{S}^{n-1} \setminus \mathcal{S}^{n-2}}\right\|_H \right) \\ &= A\left(T_{N+1} + T_{N-1}\right) \\ &= 2AT_N + A\left(\left\|u|_{\mathcal{S}^N \setminus \mathcal{S}^{N-1}}\right\|_H - \left\|u|_{\mathcal{S}^{N+1} \setminus \mathcal{S}^N}\right\|_H \right). \end{aligned}$$

After rearranging, and ignoring the negative term, we find

$$T_N \leq \frac{A}{1-2A} \left\|u|_{\mathcal{S}^N \setminus \mathcal{S}^{N-1}}\right\|_H. \tag{8}$$

Noting that we always have

$$\left\|u|_{\mathcal{S}^N \setminus \mathcal{S}^{N-1}}\right\|_H \leq T_{N-1}, \tag{9}$$

iterating (8) and (9) in turn gives

$$\|u - u|_{\mathcal{S}^N}\|_H \leq T_N \leq \left(\frac{A}{1-2A}\right)^{N+1} \|u|_{\mathcal{S}^0}\|_H \leq \left(\frac{A}{1-2A}\right)^{N+1} \frac{\|f\|_{L^2}}{a_{\min}},$$

where the final inequality follows by bounding $\|u|_{\mathcal{S}^0}\|_H \leq \|u\|_H$ from above by (4). □

Remark 2. Condition (7) is necessary for there to be truncation decay in the stamping level in the upper bound provided. However, as shown in numerical examples (see Section 10.4), there are situations where a proxy for this condition is not satisfied, but there can still be decay in the stamping level. Thus, we are lead to believe that (7) is an artifact of the proof, and a less restrictive condition may apply.

7 Previous results on SFTs

In [26], two methods for high-dimensional SFTs are presented, each with a deterministic and Monte Carlo variant. Here, we use the faster of the two algorithms (at the cost of slightly suboptimal error guarantees). We focus on only the Monte Carlo variant as the improvements to this technique described in Section 8 below use an additional layer of randomization.

This method relies on applying one-dimensional SFTs to samples of a high-dimensional function along special sets called *reconstructing rank-1 lattices*.

Definition 5. Given a number of sampling points $M \in \mathbb{N}$ and a generating vector $\mathbf{z} \in \{1, \dots M-1\}^d$, we define the *rank-1 lattice* $\Lambda(\mathbf{z}, M)$ as the set

$$\Lambda(\mathbf{z}, M) := \left\{ \frac{j}{M}\mathbf{z} \bmod \mathbf{1} \mid j \in \{0, \dots, M-1\} \right\} \subseteq \mathbb{T}^d.$$

Additionally, given a set of frequencies $\mathcal{I} \subseteq \mathbb{Z}^d$, we say that $\Lambda(\mathbf{z}, M)$ is a *reconstructing rank-1 lattice for* $\mathcal{I}$ if

$$\mathbf{l} \cdot \mathbf{z} \not\equiv \mathbf{k} \cdot \mathbf{z} \bmod M \quad \text{for all } \mathbf{l} \neq \mathbf{k} \in \mathcal{I}.$$

The fundamental idea of a reconstructing rank-1 lattice is that it takes a multivariate function $g : \mathbb{T}^d \to \mathbb{R}$ and gives the locations for M equispaced samples of the univariate function $t \mapsto g(t\mathbf{z})$. The univariate Fourier content of these samples can then be assigned to the original function g with the reconstructing property ensuring that no multidimensional frequencies of interest are aliased together in the one-dimensional analysis. For the following theorem, we assume that we know a reconstructing rank-1 lattice exists for a given frequency set of interest, $\mathcal{I}$. This assumption will be lifted in the following section.

The following theorem is a restatement of [26, Corollary 2] with minor simplifications and improvements (most notably, L^∞ error bounds). The proof of these improvements is given in Appendix B.

Theorem 2 ([26], Corollary 2)**.** *Let $\mathcal{I} \subseteq \mathbb{Z}^d$ be a frequency set of interest with expansion defined as $K := \max_{j\in\{1,\dots,d\}}(\max_{\mathbf{k}\in\mathcal{I}} k_j - \min_{\mathbf{l}\in\mathcal{I}} l_j)$ (i.e., the side-length of the smallest hypercube containing $\mathcal{I}$), and $\Lambda(\mathbf{z}, M)$ be a reconstructing rank-1 lattice for $\mathcal{I}$.*

There exists a fast, randomized SFT which, given $\Lambda(\mathbf{z}, M)$, sampling access to $g \in L^2$, and a failure probability $\sigma \in (0, 1]$, will produce a $2s$-sparse approximation $\hat{\mathbf{g}}^s$ of $\hat{g}$ and function $g^s := \sum_{\mathbf{k}\in\operatorname{supp}(\hat{\mathbf{g}}^s)} \hat{g}^s_{\mathbf{k}} e_{\mathbf{k}}$ approximating g satisfying

$$\|g - g^s\|_{L^2} \leq \|\hat{g} - \hat{\mathbf{g}}^s\|_{\ell^2} \leq (25 + 3K)\left[\frac{\|\hat{g}|_{\mathcal{I}} - (\hat{g}|_{\mathcal{I}})^{\mathrm{opt}}_s\|_1}{\sqrt{s}} + \sqrt{s}\|\hat{g} - \hat{g}|_{\mathcal{I}}\|_1\right]$$

with probability exceeding $1 - \sigma$. If $g \in L^\infty$, then we additionally have

$$\|g - g^s\|_{L^\infty} \leq \|\hat{g} - \hat{\mathbf{g}}^s\|_{\ell^1} \leq (35 + 3K)\left[\left\|\hat{g}|_{\mathcal{I}} - (\hat{g}|_{\mathcal{I}})^{\mathrm{opt}}_s\right\|_1 + s\|\hat{g} - \hat{g}|_{\mathcal{I}}\|_1\right]$$

with the same probability estimate. The total number of samples of g and computational complexity of the algorithm can be bounded above by

$$\mathcal{O}\left(ds\log^3(dKM)\log\left(\frac{dKM}{\sigma}\right)\right).$$

8 Improvements with randomized lattices

To use the previous SFT algorithm, we need to know a reconstructing rank-1 lattice in advance. Though there are deterministic algorithms to construct a reconstructing rank-1 lattice given any frequency set $\mathcal{I}$ (for example, the

component-by-component construction [32, 39]), these algorithms are superlinear in $|\mathcal{I}|$ as they effectively search the frequency space for collisions throughout construction.

This section presents an alternative based on choosing a random lattice. This lattice is chosen by drawing $\mathbf{z}$ from a uniform distribution over $\{1, \dots, M-1\}^d$ for M sufficiently large. Below, we provide probability estimates for when this lattice is reconstructing for a frequency set $\mathcal{I}$.

Lemma 4. *Let $K := \max_{j\in\{1,\dots d\}}(\max_{\mathbf{k}\in\mathcal{I}} k_j - \min_{\mathbf{l}\in\mathcal{I}} l_j)$ be the expansion of the frequency set $\mathcal{I} \subseteq \mathbb{Z}^d$. Let $\sigma \in (0, 1]$, and fix M to be the smallest prime greater than $\max(K, \frac{|\mathcal{I}|^2}{\sigma})$. Then drawing each component of $\mathbf{z}$ i.i.d. uniformly from $\{1, \dots M-1\}$ gives that $\Lambda(\mathbf{z}, M)$ is a reconstructing rank-1 lattice for $\mathcal{I}$ with probability at least $1 - \sigma$.*

Proof In order to show that $\Lambda(\mathbf{z}, M)$ is reconstructing for $\mathcal{I}$, it suffices to show that for any $\mathbf{k} \neq \mathbf{l} \in \mathcal{I}$, $\mathbf{k}\cdot\mathbf{z} \not\equiv \mathbf{l}\cdot\mathbf{z} \bmod M$. Thus, we are interested in showing that $\mathbb{P}[\exists \mathbf{k} \neq \mathbf{l} \in \mathcal{I} \text{ s.t. } (\mathbf{k}-\mathbf{l})\cdot\mathbf{z} \equiv \mathbf{0} \bmod M]$ is small.

If $\mathbf{k}, \mathbf{l} \in \mathcal{I}$ are distinct, at least one component $k_j - l_j$ is nonzero. Since $M > K$, we therefore have that $k_j - l_j \not\equiv 0 \bmod M$, and since M is prime, $k_j - l_j$ has a multiplicative inverse modulo M. Then $\mathbb{P}[(\mathbf{k}-\mathbf{l})\cdot\mathbf{z} \equiv \mathbf{0} \bmod M] = \mathbb{P}\left[z_j = \left((k_j - l_j)^{-1} \sum_{i\in\{1,\dots d\}, i\neq j}(k_i - l_i)z_i \bmod M\right)\right]$. Since z_j is uniformly distributed in $\{1, \dots M-1\}$, this probability is $\frac{1}{M-1}$. By the union bound,

$$\begin{aligned}\mathbb{P}[\exists \mathbf{k} \neq \mathbf{l} \in \mathcal{I} \text{ s.t. } (\mathbf{k}-\mathbf{l})\cdot\mathbf{z} \equiv \mathbf{0} \bmod M] &\leq \sum_{\mathbf{k}\neq\mathbf{l}\in\mathcal{I}} \mathbb{P}[(\mathbf{k}-\mathbf{l})\cdot\mathbf{z} \equiv \mathbf{0} \bmod M] \\ &\leq \frac{|\mathcal{I}|^2}{M-1} \\ &\leq \sigma\end{aligned}$$

as desired.

□

One important consequence of Lemma 4 is that we no longer need to provide the frequency set of interest in Theorem 2. Having chosen K, the expansion, and s, the sparsity level, we can always take $\mathcal{I}$ to be the frequencies corresponding to the largest s Fourier coefficients of the function g in the hypercube $[-K/2, K/2]^d$. Lemma 4 then implies that a randomly generated lattice with length $\max(K, s^2/\sigma)$ will be reconstructing for these optimal frequencies with probability σ. We summarize this in the following corollary.

Corollary 3. *Fix a multivariate bandwidth K. For a multivariate function's Fourier series $\hat{g}$, then define $\hat{g}|_K := \hat{g}|_{[-K/2,K/2]^d}$. Now also fix a sparsity level s and a probability of failure $\sigma \in (0, 1]$, and suppose that you have sampling access to a given $g \in L^2$. Then, there exists a fast, randomized SFT which will produce a $2s$-sparse approximation $\hat{\mathbf{g}}^s$ of $\hat{g}$ as well as a function $g^s :=$*

$\sum_{\mathbf{k}\in\mathrm{supp}(\hat{\mathbf{g}}^s)} \hat{g}^s_{\mathbf{k}} e_{\mathbf{k}}$ *approximating* g *that will satisfy*

$$\|g - g^s\|_{L^2} \leq \|\hat{g} - \hat{\mathbf{g}}^s\|_{\ell^2} \leq (25 + 3K)\sqrt{s}\left\|\hat{g} - (\hat{g}|_K)^{\mathrm{opt}}_s\right\|_{\ell^1}$$

with probability at least $1 - \sigma$*. If* $g \in L^\infty$*, then* g^s *and* $\hat{g}^s$ *will also satisfy the upper bound*

$$\|g - g^s\|_{L^\infty} \leq \|\hat{g} - \hat{\mathbf{g}}^s\|_{\ell^1} \leq (35 + 3K)s\left\|\hat{g} - (\hat{g}|_K)^{\mathrm{opt}}_s\right\|_{\ell^1}$$

with the same probability estimate. Furthermore, both the total number of samples of g *and computational complexity of the algorithm is always bounded above by*

$$\mathcal{O}\left(ds\log^3(dK\max(K, s/\sigma))\log\left(\frac{dK\max(K, s/\sigma)}{\sigma}\right)\right).$$

If, e.g., we fix σ *(to, say,* $\sigma = 0.95$*), this reduces to a complexity of*

$$\mathcal{O}\left(ds\log^4(dK\max(K, s))\right).$$

9 A sparse spectral method via SFTs

Let $\hat{\mathbf{a}}^s$ and $\hat{\mathbf{f}}^s$ be s-sparse approximations of $\hat{a}$ and $\hat{f}$ respectively. We will use these approximations to discretize the Galerkin formulation (GF) of our PDE. The first step is to reduce to the case where the PDE data is Fourier-sparse which is motivated by the following lemma.

Lemma 5. *Let* $a' := a|_{\mathrm{supp}\,\hat{\mathbf{a}}^s}$ *and* $f' := f|_{\mathrm{supp}\,\hat{\mathbf{f}}^s}$*. Suppose that* a' *and* f' *satisfy the conditions of Proposition 2 and let* u' *be the unique solution of the resulting elliptic PDE, which we write in Galerkin form as*

$$L[\hat{a}']\hat{u}' = \hat{f}'. \tag{10}$$

Then

$$\|u - u'\|_H \leq \frac{\|f - f'\|_{L^2}}{a_{\min}} + \frac{\|a - a'\|_{L^\infty}\|f'\|_{L^2}}{a_{\min}a'_{\min}}.$$

Proof We begin by observing

$$L[\hat{a}](\hat{u} - \hat{u}') = L[\hat{a}]\hat{u} - L[\hat{a}']\hat{u}' - L[\hat{a} - \hat{a}']\hat{u}' = \hat{f} - \hat{f}' - L[\hat{a} - \hat{a}']\hat{u}',$$

and therefore

$$\left|\langle L[\hat{a}](\hat{u} - \hat{u}'), \hat{u} - \hat{u}'\rangle\right| \leq \left|\langle \hat{f} - \hat{f}', \hat{u} - \hat{u}'\rangle\right| + \left|\langle L[\hat{a} - \hat{a}']\hat{u}', \hat{u} - \hat{u}'\rangle\right|.$$

After an application of Proposition 3 to convert the ℓ^2 inner products into bilinear forms, we can make use of coercivity, (3), continuity, (2) and the Cauchy-Schwarz inequality to produce the H approximation

$$a_{\min}\|u - u'\|_H \leq \left\|\hat{f} - \hat{f}'\right\|_{\ell^2} + \|a - a'\|_{L^\infty}\|u'\|_H.$$

An application of the stability estimate (4) gives the desired bound

$$\|u - u'\|_H \leq \frac{\|f - f'\|_{L^2}}{a_{\min}} + \frac{\|a - a'\|_{L^\infty}\|f'\|_{L^2}}{a_{\min}a'_{\min}}.$$

□

We can now replace the trial and test spaces in (WF) with finite dimensional approximations so as to convert (GF) to a matrix equation. Inspired by Proposition 4 and the truncation error analysis in Section 6, we use the space of functions whose Fourier coefficients are supported on $\mathcal{S}^N := \mathcal{S}^N[\hat{a}](\operatorname{supp} \hat{f})$. By doing so, we discretize the Galerkin formulation of the problem (GF) into the finite system of equations

$$(\mathbf{L}_N \hat{\mathbf{u}})_{\mathbf{k}} := \sum_{\mathbf{l} \in \mathcal{S}^N} (2\pi)^2 (\mathbf{l} \cdot \mathbf{k}) \hat{a}_{\mathbf{k}-\mathbf{l}} \hat{u}_{\mathbf{l}} = \hat{f}_{\mathbf{k}} \quad \text{for all } \mathbf{k} \in \mathcal{S}^N. \tag{11}$$

However, in practice, we do not know $\hat{a}$ and $\hat{f}$ exactly (and indeed, they may not be exactly sparse). Thus, we substitute the SFT approximations $\hat{\mathbf{a}}^s$ and $\hat{\mathbf{f}}^s$, defining the new finite-dimensional operator $\mathbf{L}_{N,s} : \mathbb{C}^{\mathcal{S}^N} \to \mathbb{C}^{\mathcal{S}^N}$ by

$$(\mathbf{L}_{N,s} \hat{\mathbf{u}})_{\mathbf{k}} := \sum_{\mathbf{l} \in \mathcal{S}^N} (2\pi)^2 (\mathbf{l} \cdot \mathbf{k}) \hat{a}^s_{\mathbf{k}-\mathbf{l}} \hat{u}_{\mathbf{l}} \quad \text{for all } \mathbf{k} \in \mathcal{S}^N.$$

Our new approximate solution will be $\hat{\mathbf{u}}^{N,s} \in \mathbb{C}^{\mathcal{S}^N}$ which solves

$$\mathbf{L}_{N,s} \hat{\mathbf{u}}^{N,s} = \hat{\mathbf{f}}^s. \tag{12}$$

We summarize our technique in Algorithm 1.

Algorithm 1 Sparse spectral method

Input: PDE data a and f, a sparsity parameter s, a bandwidth parameter K, and a stamping level N

Output: Fourier coefficients $\hat{\mathbf{u}}^{s,N}$ of approximate solution

1: $\hat{\mathbf{a}}^s \leftarrow \mathrm{SFT}[s, K](a)$ // SFT is the algorithm in [26] using a random rank-1 lattice (cf. Section 8)
2: $\hat{\mathbf{f}}^s \leftarrow \mathrm{SFT}[s, K](f)$
3: Compute $\mathcal{S}^N[\hat{\mathbf{a}}^s](\operatorname{supp}(\hat{\mathbf{f}}^s))$ // see, e.g., (5) or (A3)
4: $(\mathbf{L}_{N,s})_{\mathbf{k} \in \mathcal{S}^N, \mathbf{l} \in \mathcal{S}^N} \leftarrow (2\pi)^2 (\mathbf{l} \cdot \mathbf{k}) \hat{a}^s_{\mathbf{k}-\mathbf{l}}$
5: $\hat{\mathbf{u}}^{N,s} \leftarrow \mathbf{L}_{N,s} \backslash \hat{\mathbf{f}}^s$ // using MATLAB backslash notation for matrix solve

Showing that $u^{N,s}$ converges to u now relies on a version of Strang's lemma [11, Equation (6.4.46)]. We make the assumption here that $\operatorname{supp}(\hat{a}) = \operatorname{supp}(\hat{\mathbf{a}}^s)$ and $\operatorname{supp}(\hat{f}) = \operatorname{supp}(\hat{\mathbf{f}}^s)$ so that our use of $\mathcal{S}^N$ is unambiguous. However, this assumption will be lifted by Lemma 5 in Corollary 4 below.

Lemma 6 (Strang's Lemma)**.** *Suppose that* $\operatorname{supp}(\hat{a}) = \operatorname{supp}(\hat{\mathbf{a}}^s)$ *and that* $\operatorname{supp}(\hat{f}) = \operatorname{supp}(\hat{\mathbf{f}}^s)$*. Also suppose that* $a^s \geq a^s_{\min} > 0$ *on* $\mathbb{T}^d$*. Let* u *and* $u^{N,s}$ *be as above. Then*

$$\left\|u - u^{N,s}\right\|_H \leq \left(1 + \frac{\|a\|_{L^\infty}}{a^s_{\min}}\right) \left\|u|_{\mathbb{Z}^d \setminus \mathcal{S}^N}\right\|_H + \frac{\|a - a^s\|_{L^\infty}}{a^s_{\min}} \|u|_{\mathcal{S}^N}\|_H + \frac{\|f - f^s\|_{L^2}}{a^s_{\min}}.$$

Proof We let $\hat{\mathbf{e}} := \hat{\mathbf{u}}^{N,s} - \hat{u}|_{\mathcal{S}^N}$, and consider

$$\begin{aligned}
\mathbf{L}_{N,s}\hat{\mathbf{e}} &= \mathbf{L}_{N,s}\hat{\mathbf{u}}^{N,s} - (L[\hat{\mathbf{a}}^s]\hat{u}|_{\mathcal{S}^N})|_{\mathcal{S}^N} \\
&= \hat{\mathbf{f}}^s - \hat{f} + (L[\hat{a}]\hat{u})|_{\mathcal{S}^N} - (L[\hat{\mathbf{a}}^s]\hat{u}|_{\mathcal{S}^N})|_{\mathcal{S}^N} \\
&= \hat{\mathbf{f}}^s - \hat{f} + (L[\hat{a}]\hat{u}|_{\mathbb{Z}^d \setminus \mathcal{S}^N})|_{\mathcal{S}^N} + (L[\hat{a}]\hat{u}|_{\mathcal{S}^N} - L[\hat{\mathbf{a}}^s]\hat{u}|_{\mathcal{S}^N})|_{\mathcal{S}^N} \\
&= \hat{\mathbf{f}}^s - \hat{f} + (L[\hat{a}]\hat{u}|_{\mathbb{Z}^d \setminus \mathcal{S}^N})|_{\mathcal{S}^N} + (L[\hat{a} - \hat{\mathbf{a}}^s]\hat{u}|_{\mathcal{S}^N})|_{\mathcal{S}^N}.
\end{aligned}$$

Noting that $\mathbf{L}_{N,s}\hat{\mathbf{e}} = (L[\hat{\mathbf{a}}^s]\hat{\mathbf{e}})|_{\mathcal{S}^N}$ and owing to coercivity of $L[\hat{\mathbf{a}}^s]$, we have

$$\begin{aligned}
a^s_{\min}\|e\|_H^2 &\leq \left|\langle \mathbf{L}_{N,s}\hat{\mathbf{e}}, \hat{\mathbf{e}}\rangle\right| \\
&\leq \left\|f^s - f\right\|_{L^2}\|e\|_H + \|a\|_{L^\infty}\left\|u|_{\mathbb{Z}^d \setminus \mathcal{S}^N}\right\|_H \|e\|_H \\
&\quad + \left\|a - a^s\right\|_{L^\infty}\|u|_{\mathcal{S}^N}\|_H\|e\|_H.
\end{aligned}$$

The result then follows from rearranging to estimate $\|e\|_H$ and using the triangle inequality to estimate $\left\|u - u^{N,s}\right\|_H \leq \|u - u|_{\mathcal{S}^N}\|_H + \|e\|_H$. □

We can now thread all of our results together into a final convergence analysis. The first corollary below is a more direct application of Strang's lemma which is then followed by another corollary which takes advantage of the SFT recovery results. We will also return to the setting where a and f are not necessarily Fourier sparse. Thus, for a^s and f^s Fourier sparse approximations of a and f, we again let $a' = a|_{\operatorname{supp}\hat{\mathbf{a}}^s}$ and $f' = f|_{\operatorname{supp}\hat{\mathbf{f}}^s}$ as in Lemma 5.

Corollary 4. *Suppose* a*,* f *and* a^s*,* f^s *respectively satisfy the conditions of Proposition 2. Additionally, suppose that*

$$4 \sum_{\mathbf{k} \in \operatorname{supp}(\hat{\mathbf{a}}^s) \setminus \{\mathbf{0}\}} |\hat{a}_{\mathbf{k}}| \leq \hat{a}_{\mathbf{0}}. \tag{13}$$

Then with u *the exact solution to* (WF) *and* $u^{N,s}$ *the output of Algorithm 1, we have*

$$\begin{aligned}
\left\|u - u^{N,s}\right\|_H \leq{}& \frac{\|f - f'\|_{L^2}}{a_{\min}} + \frac{\|a - a'\|_{L^\infty}\|f'\|_{L^2}}{a_{\min}a'_{\min}} \\
&+ \left(1 + \frac{\|a'\|_{L^\infty}}{a^s_{\min}}\right)\left(\frac{\|a' - \hat{a}'_{\mathbf{0}}\|_{L^\infty}}{a'_{\min} - 2\|a' - \hat{a}'_{\mathbf{0}}\|_{L^\infty}}\right)^{N+1} \frac{\|f'\|_{L^2}}{a'_{\min}}
\end{aligned}$$

$$+ \frac{\|a' - a^s\|_{L^\infty} \|f'\|_{L^2}}{a^s_{\min} a_{\min}} + \frac{\|f' - f^s\|_{L^2}}{a^s_{\min}}$$

Proof The condition (13) ensures that a' is coercive, and therefore a' and f' also satisfy Proposition 2. Additionally, since

$$\hat{a}_{\mathbf{0}} - \|\hat{a}' - \hat{a}'_{\mathbf{0}}\|_{\ell^1} \leq \hat{a}_{\mathbf{0}} - \|a' - \hat{a}'_{\mathbf{0}}\|_{L^\infty} \leq \hat{a}'_{\min}$$

this allows the use of Lemma 3, which upper bounds the truncation error in Lemma 6. Combining Lemma 5 with this bound from Lemma 6 and applying the stability estimate from Proposition 2 finishes the proof. □

Remark 3. In order for this bound to hold, it is necessary for the weak forms of both

$$\mathcal{L}[a]u = f \text{ and } \mathcal{L}[a^s]u^s = f^s$$

to be well-posed, that is, satisfy the continuity and coercivity conditions of Proposition 2. In practice, this condition is not much more restrictive than assuming only the original equation is well-posed as long as the diffusion coefficient is Fourier-compressible and the sparsity level s is large enough to ensure that a^s stays strictly positive. In fact, (13) allows for the simple (if pessimistic) check after computing $\hat{\mathbf{a}}^s$ that $\|\hat{\mathbf{a}}^s - \hat{a}^s_{\mathbf{0}}\|_{\ell^1} < |\hat{a}^s_{\mathbf{0}}|$ to ensure the positivity of a^s.

With minor modifications, we can rewrite this upper bound to pass all dependence on sparsity through the error in approximating a and f via SFTs.

Corollary 5. *Under the same conditions as Corollary 4 above substituting* (13) *with*

$$3\|\hat{a} - \hat{a}_{\mathbf{0}}\|_{\ell^1} + \|\hat{a} - \hat{\mathbf{a}}^s\|_{\ell^1} < a_{\min}, \tag{14}$$

we have

$$\begin{aligned}\left\|u - u^{N,s}\right\|_H \leq \left(1 + \frac{\|\hat{a}\|_{\ell^1}}{a_{\min} - \|\hat{a} - \hat{\mathbf{a}}^s\|_{\ell^1}}\right) \frac{\|f\|_{L^2}}{a_{\min} - \|\hat{a} - \hat{\mathbf{a}}^s\|_{\ell^1}} \\ \times \left(\frac{\|f - f^s\|_{L^2}}{\|f\|_{L^2}} + \|a - a^s\|_{L^\infty} \right. \\ \left. + \left(\frac{\|\hat{a} - \hat{a}_{\mathbf{0}}\|_{\ell^1}}{a_{\min} - 2\|\hat{a} - \hat{a}_{\mathbf{0}}\|_{\ell^1} - \|\hat{a} - \hat{\mathbf{a}}^s\|_{\ell^1}}\right)^{N+1}\right).\end{aligned}$$

Proof Since $\hat{a}' = \hat{a}|_{\operatorname{supp} \hat{\mathbf{a}}^s}$,

$$\begin{aligned}\|a - a'\|_{L^\infty} \leq \|\hat{a} - \hat{a}'\|_{\ell^1} \leq \|\hat{a} - \hat{\mathbf{a}}^s\|_{\ell^1}, \\ \|a' - a^s\|_{L^\infty} \leq \|\hat{a}' - \hat{\mathbf{a}}^s\|_{\ell^1} \leq \|\hat{a} - \hat{\mathbf{a}}^s\|_{\ell^1},\end{aligned}$$

and analogously to show that $\|f - f'\|_{L^2}$ and $\|f' - f^s\|_{L^2}$ are bounded above by $\|f - f^s\|_{L^2}$. Additionally,

$$\begin{aligned}a^s \geq a - \|a - a^s\|_{L^\infty} \geq a - \|\hat{a} - \hat{\mathbf{a}}^s\|_{\ell^1} \text{ and} \\ a' \geq a - \|a - a'\|_{L^\infty} \geq a - \|\hat{a} - \hat{\mathbf{a}}^s\|_{\ell^1}\end{aligned}$$

giving $\min(a^s_{\min}, a'_{\min}) \geq a_{\min} - \|\hat{a} - \hat{\mathbf{a}}^s\|_{\ell^1}$. The rest follows from applications of (4) and rearranging. □

Remark 4. Though this final bound is difficult to parse, we can focus our attention on the final factor

$$\frac{\|f - f^s\|_{L^2}}{\|f\|_{L^2}} + \|a - a^s\|_{L^\infty} + \left(\frac{\|\hat{a} - \hat{a}_{\mathbf{0}}\|_{\ell^1}}{a_{\min} - 2\|\hat{a} - \hat{a}_{\mathbf{0}}\|_{\ell^1} - \|\hat{a} - \hat{\mathbf{a}}^s\|_{\ell^1}}\right)^{N+1}, \quad (15)$$

since the other factors are more or less fixed. The first two terms are respectively controlled by having good SFT approximations to f in the L^2 norm and a in the L^∞ norm. In our algorithm, these terms can be reduced by increasing the bandwidth K and the sparsity s. As a reminder, the errors in these approximations given in Theorem 2 are near optimal, as

$$\|f - f^s\|_{L^2} \leq (25 + 3K)\sqrt{s}\left\|\hat{f} - \left(\hat{f}|_K\right)_s^{\text{opt}}\right\|_{\ell^1}$$

and

$$\|a - a^s\|_{L^\infty} \leq (35 + 3K)s\left\|\hat{a} - (\hat{a}|_K)_s^{\text{opt}}\right\|_{\ell^1}$$

with high probability.

The final term is controlled by properties of a as well as the final stamping level used. Overall, the convergence is exponential in N, the stamping level. This convergence is accelerated as the base of the exponent decreases: effectively, this happens as the diffusion coefficient approaches a large constant. Indeed, the numerator can be thought of as an upper bound for the absolute deviation of a from its mean while the denominator grows with the minimum of a.

Remark 5. The computational complexity of Algorithm 1 is

$$\mathcal{O}\left(ds\log^4(dK\max(K,s)) + s^3\max(s,2N)^{3\min(s,2N)}\right).$$

This is due to the two SFTs and a matrix solve of a $\left|\mathcal{S}^N\right| \times \left|\mathcal{S}^N\right|$ system. Note that computing the stamping set can be done by enumerating the frequencies using the techniques in Lemma 8 and therefore is subject to the same upper bound as given in Lemma 2 for a stamp set's cardinality. Recall also that the SFT complexity can be tuned to produce SFT approximations satisfying the above bounds with higher probability.

We do not analyze the complexity of the matrix solve in depth, and instead resort to the upper bound given by Gaussian elimination on the dense matrix, $\mathcal{O}\left(s^3\max(s,2N)^{3\min(s,2N)}\right)$. However, $\mathbf{L}_{N,s}$ is relatively sparse for larger stamping levels. As the capabilities of sparse solvers depend strongly on analyzing the graph connecting interacting rows in $\mathbf{L}_{N,s}$ (cf. [21, Chapter 11]), we expect that the analysis of an efficient sparse solver could be carried out using much of the same analysis of stamping sets performed in Section 6.

Remark 6. This paper considers the theory for solving the simple diffusion equation (1). However, these techniques extend to more complex advection-diffusion-reaction (ADR) equations. The test problem is then

$$-\nabla\cdot(a(\mathbf{x})\nabla u(\mathbf{x}))+\mathbf{b}(\mathbf{x})\cdot\nabla u(\mathbf{x})+c(\mathbf{x})u(\mathbf{x})=f(\mathbf{x}) \text{ for all } \mathbf{x}\in\mathbb{T}^3. \quad (16)$$

As before $a, f, u : \mathbb{T}^d \to \mathbb{R}$ are the diffusion coefficient, forcing function, and solution respectively. These are now joined by an advection field $\mathbf{b} : \mathbb{T}^d \to \mathbb{R}^d$ and an additional reaction coefficient $c : \mathbb{T}^d \to \mathbb{R}$. For more on the properties and well-posedness of this periodic ADR equation, we refer to [4].

Adapting Algorithm 1 for solving ADR equations requires two modifications:

1. When computing the approximations $\hat{\mathbf{a}}^s, \hat{\mathbf{f}}^s$ via SFT, additionally compute $\hat{\mathbf{b}}^s := (\hat{\mathbf{b}}^s_j)_{j=1}^d$, an approximation to the Fourier coefficients of each component of $\mathbf{b}$, and compute $\hat{\mathbf{c}}^s$, an approximation to $\hat{c}$.
2. Redefine the "stamp" used to define $\mathcal{S}^N[\hat{\mathbf{a}}^s](\mathrm{supp}(\hat{\mathbf{f}}^s))$ by including the supports of $\hat{\mathbf{b}}^s$ and $\hat{\mathbf{c}}^s$. Mathematically, we define

$$\mathcal{S}^N[\hat{\mathbf{a}}^s,\hat{\mathbf{b}}^s,\hat{\mathbf{c}}^s](\mathrm{supp}(\hat{\mathbf{f}}^s)) := \begin{cases} \mathrm{supp}(\hat{\mathbf{f}}^s) & \text{if } N=0 \\ \mathcal{S}^{N-1}+\mathrm{supp}(\hat{\mathbf{a}}^s)+\sum_{j=1}^d \mathrm{supp}(\hat{\mathbf{b}}^s_j)+\mathrm{supp}(\hat{\mathbf{c}}^s) & \text{if } N>0 \end{cases}$$

where, as usual, we suppress the Fourier coefficients when clear from context.

The convergence analysis for this method is much the same as that leading to Corollary 5 where terms like $\|a-a^s\|_{L^\infty}$ are replaced by the term $\max\left\{\|a-a^s\|_{L^\infty}, \|\|\mathbf{b}-\mathbf{b}^s\|_{\ell^2}\|_{L^\infty}, \|c-c^s\|_{L^\infty}\right\}$ and similarly for the mean-zero version of a used in the exponentially decaying term. For full details see [25].

10 Numerics

This section gives examples of the algorithm summarized above applied to various problems. We begin with an overview of our implementation as well as some techniques used to evaluate the accuracy of our approximations. We then present solutions to univariate and very high-dimensional multiscale problems with both exactly sparse and Fourier-compressible data. We then close with an extension of our methods to a three-dimensional advection-diffusion-reaction equation.

10.1 Code and testing overview

We implement Algorithm 1 described above in MATLAB using an object-oriented approach, with all code publicly available.[3] All SFTs are computed using the rank-1 lattice sparse Fourier code from [26].[4]

In order to evaluate the quality of our approximations, we need to choose an appropriate metric. Letting $u^{s,N}$ be the approximation returned by our algorithm, the ideal choice would be $\left\|u - u^{s,N}\right\|_H$. However, for the types of problems we will be investigating, the true solution u is unavailable to us. Instead, we will use a proxy that takes advantage of the stability result in Proposition 2.

Lemma 7. *Let u be the true solution to (GF) and $u^{s,N}$ be the approximation returned by solving (12). Define $\hat{f}^{s,N} := L[\hat{a}]\hat{u}^{s,N}$ with $f^{s,N} = \mathcal{L}[a]u^{s,N}$. Then*

$$\left\|u - u^{s,N}\right\|_H \leq \frac{\left\|f - f^{s,N}\right\|_{L^2}}{a_{\min}} = \frac{\left\|\hat{f} - \hat{f}^{s,N}\right\|_{\ell^2}}{a_{\min}}.$$

Proof The result follows from the fact that $\hat{u} - \hat{u}^{s,N}$ solves $L[\hat{a}]\left(\hat{u} - \hat{u}^{s,N}\right) = \hat{f} - L[\hat{a}]\hat{u}^{s,N} = \hat{f} - \hat{f}^{s,N}$ and applying Proposition 2. □

In the sequel, we will ignore $a_{\min}$ since we are mostly interested in convergence properties in s and N and we will compute the relative error

$$\frac{\left\|f - f^{s,N}\right\|_{L^2}}{\|f\|_{L^2}} \text{ or } \frac{\left\|\hat{f} - \hat{f}^{s,N}\right\|_{\ell^2}}{\left\|\hat{f}\right\|_{\ell^2}}$$

as our proxy instead. Whenever $\hat{f}$ and $\hat{a}$ are exactly sparse, the numerator of the second term can be computed exactly due to the fact that $\operatorname{supp}(\hat{f}^{s,N})$ is known to be contained in $\mathcal{S}^{N+1}$ (cf. Proposition 5). However, in the non-sparse setting, even though $f - f^{s,N}$ can be evaluated pointwise, computing an accurate approximation of its norm on $\mathbb{T}^d$ is challenging for large d. For this reason, we approximate the norm via Monte Carlo sampling. We also furnish the cases where exactly computing $\left\|\hat{f} - \hat{f}^{s,N}\right\|_{\ell^2}$ is possible with the pointwise Monte Carlo estimates to show that in practice, Monte Carlo sampling does as well as the exact computation.

[3] https://gitlab.com/grosscra/SparseADR
[4] this code is publicly available at https://gitlab.com/grosscra/Rank1LatticeSparseFourier

10.2 Univariate compressible

We begin by replicating the lone numerical example of solving an elliptic problem in [15, Section 5.1]. In this case, we solve the univariate problem

$$\begin{aligned} -(a(x)u'(x))' &= f(x) \text{ for all } x \in \mathbb{T}, \text{ where} \\ a(x) &= \frac{1}{10}\exp\left(\frac{0.6 + 0.2\cos(2\pi x)}{1 + 0.7\sin(256\pi x)}\right), \\ f(x) &= \exp(-\cos(2\pi x)) - \int_{\mathbb{T}} \exp(-\cos(2\pi y))\, dy \end{aligned} \tag{17}$$

(note that the only difference from [15] is that we use the domain $\mathbb{T} = [0, 1]$ rather than $[0, 2\pi]$). This data is not Fourier sparse, but is compressible. In the original paper, a bandwidth of $K = 1\,536$ is considered and approximations with 9 and 17 Fourier coefficients are used.

We first construct a high accuracy approximation of the solution to (17) by numerically integrating on an extremely fine mesh of 10 000 points. This allows us to forgo our proxy error described in Lemma 7. As in [15], the bandwidth of our SFT used is set to $K = 1\,536$. Due to our SFT returning a $2s$ sparse approximation, we use $s = 4$ and $s = 8$ to compare with the 9 and 17 terms respectively considered in the original paper, and also provide an example with $s = 12$. We set the stamping level to $N = 1$ throughout, which, as discussed in the introduction, is similar to the technique used in [15].

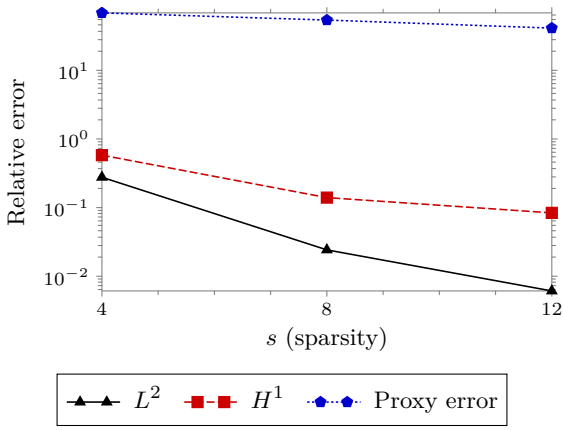
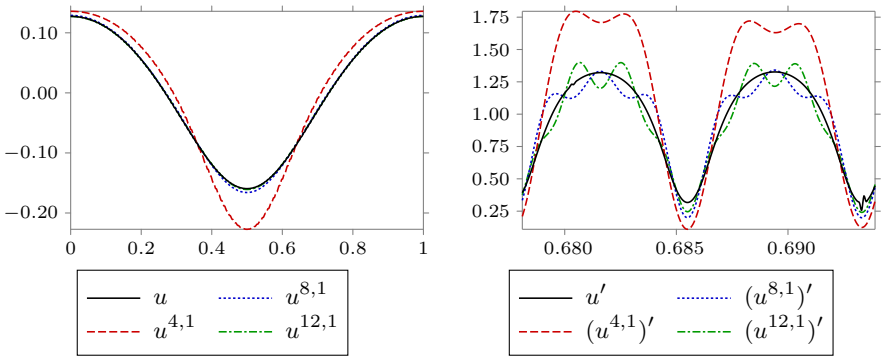


Fig. 2: Errors in approximating the solution to (17).

The relative errors approximated in L^2 and H^1 are given in Figure 2. The original paper does not give numerical results, and instead, gives qualitative results, comparing the approximate solutions and their derivatives with the true solution and its derivative. We have replicated this qualitative analysis in Figure 3 with similar results.



(a) Approximate solutions of (17). (b) Approximate derivatives of (17).

Fig. 3: Qualitative results.

Figure 2 also shows the error computed via the proxy described by Lemma 7, and in particular, how pessimistic the proxy error can be. In this case, the small errors in the derivative (visualized in Figure 3b) are compounded by passing the approximate solution through the operator where a' is often large relative to a. In future examples, we will see that the convergence of the proxy error is much more tolerable.

10.3 Multivariate exactly sparse

10.3.1 Low sparsity

Moving to the multivariate case, we start with a simple example with exactly sparse data. Our goal is to solve

$$\begin{aligned} -\nabla \cdot (a(\mathbf{x})\nabla u(\mathbf{x})) &= f(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathbb{T}^d, \text{ where} \\ a(\mathbf{x}) = \hat{a}_0 + c_a \cos(2\pi \mathbf{k}_a \cdot \mathbf{x}), &\quad f(x) = \sin(2\pi \mathbf{k}_f \cdot \mathbf{x}). \end{aligned} \tag{18}$$

We draw $c_a \sim \text{Unif}([-1,1])$, keep it constant for each dimension, and set $\hat{a}_{\mathbf{0}} = 4$ so that our problem remains elliptic (in the specific example below, $c_a \approx -0.6$). For dimensions varying from $d = 1$ to $d = 1\,024$, we then draw $\mathbf{k}_a, \mathbf{k}_f \sim \text{Unif}\left([-499, 500]^d \cap \mathbb{Z}^d\right)$. The PDE (18) is then solved for stamping levels $N = 1, \ldots, 5$. The bandwidth of the SFT is set to 1000 and the sparsity is set to 2. We then compute a Monte Carlo approximation of the proxy error choosing 200 points drawn uniformly from $\mathbb{T}^d$ and also compute the proxy error exactly by virtue of the sparsity of a and f. The results are given in Figure 4.

We see that the results do not depend on the dimension of the problem. Since all dependence on d is in the runtime of the SFT, we also observe that in practice, after the SFTs of the data have been computed, re-solving the problem on different stamping levels takes about the same amount of time for each d. The error also converges exponentially in the stamping level as

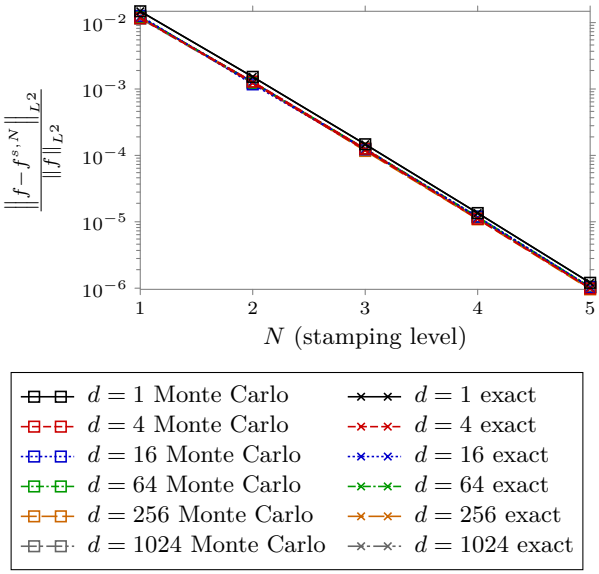


Fig. 4: Proxy error solving (18) with $d = 1, 4, 16, 64, 256, 1\,024$ and $N = 1, \ldots, 5$.

suggested by the theoretical error guarantees. Notably, we also see that the Monte Carlo approximation with 200 points captures the same proxy error as the exact computation.

10.3.2 High sparsity

We expand on the exactly sparse case by testing a diffusion coefficient with much higher sparsity. Here, we solve (18) with

$$a(\mathbf{x}) = \hat{a}_{\mathbf{0}} + \sum_{\mathbf{k} \in \mathcal{I}_a} c_{\mathbf{k}} \cos(2\pi \mathbf{k} \cdot \mathbf{x}). \tag{19}$$

The vector of coefficients is drawn as $\mathbf{c} \sim \text{Unif}\left([-1, 1]^{25}\right)$ once and reused in each test. For every d, the frequencies $\mathbf{k} \in \mathcal{I}_a$ are each drawn uniformly from $[-499, 500]^d \cap \mathbb{Z}^d$ as before with $|\mathcal{I}_a| = 25$. Here $\hat{a}_0 = 4 \lceil \|\mathbf{c}\|_2 \rceil$ to ensure ellipticity. Again, the bandwidth of the SFT algorithm is set to $1\,000$, but the sparsity is now fixed to 26. We only consider stamping levels up to $N = 3$ due to memory considerations (see the following section, Section 10.3.3, for details). The results are given in Figure 5.

Again, we see that the results do not depend on the spatial dimension except for the notable example of $d = 1$. The $d = 1$ case suffers from similar issues in a pessimistic proxy error as in Figure 2. Specifically, the right hand-side for this example was generated with frequency $k_f = -10$ and is therefore relatively low-frequency. Thus, the high-frequency modes leading to errors in

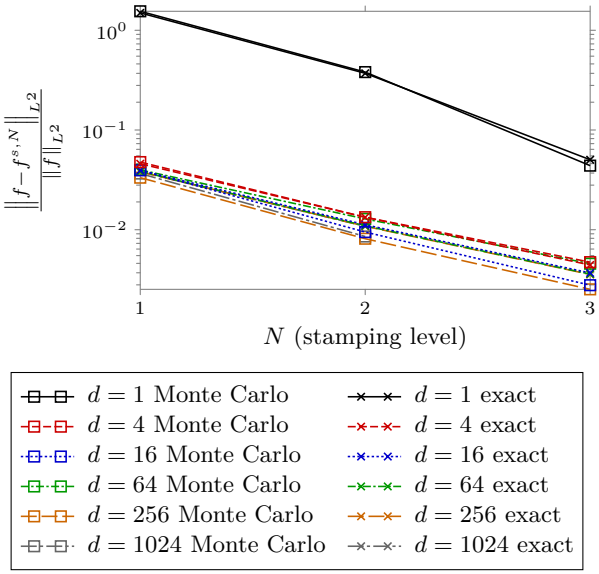


Fig. 5: Proxy error solving (18) with diffusion coefficient (19) in dimensions $d = 1, 4, 64, 256, 1\,024$ and stamping levels $N = 1, \ldots, 3$.

the approximate solution are amplified by the high-frequencies in a when computing $f^{s,N}$. Indeed, in further experiments (not pictured here), increasing the frequencies of f or decreasing the frequencies of a result in a lower proxy error.

For the other dimensions, the slight offsets in the exact proxy error can be attributed to the randomized frequencies as well as slight variations in the randomized SFT code. We do see slightly more variance in the proxy error computed using Monte Carlo sampling however. This is to be expected for data with more varied frequency content, and as such, in future experiments, we increase the number of sampling points.

10.3.3 Stamp size and complexity comparisons

We also use this exactly sparse setting to provide insight into the memory and computational complexity of Algorithm 1. First, Figures 6 and 7 show the cardinality of the stamping sets used in Figures 4 and 5 respectively. We also show the upper bound for the stamp set cardinality provided in (6) as well as the more accurate combinatorial bound (A1) used to prove this bound.

In the low sparsity case depicted in Figure 6, we see that the stamp set sizes do not depend on dimension. This contrasts with the fact that the $d = 1$ case flattens out quickly in the high sparsity setting depicted in Figure 7 while the sizes for the remaining dimensions are indistinguishable and grow exponentially in the stamping level N.

However, this is to be expected, since in one dimension, there is significant overlap in stamping levels that does not occur when stamping sets in higher

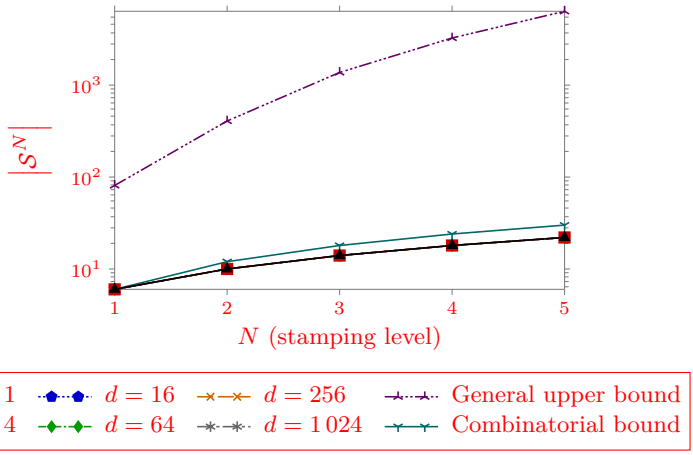


Fig. 6: Cardinality of the stamp sets used in Figure 4 compared against the combinatorial upper bound (A1) and the more general upper bound (6).

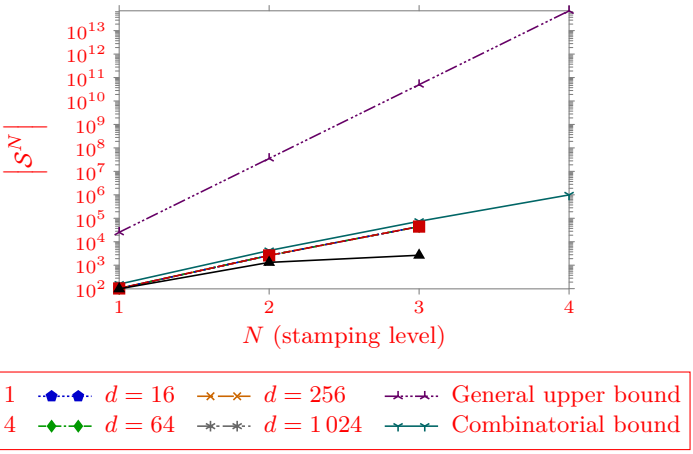


Fig. 7: Cardinality of the stamp sets used in Figure 5 compared against the combinatorial upper bound (A1) and the more general upper bound (6).

dimensions are able to "spread out" in those additional dimensions. In the low sparsity case, the same amount of overlapping as in one dimension also happens in higher dimensions since the stamping sets effectively grow in the direction dictated by the few frequencies of the diffusion coefficient. In examples not pictured here, when diffusion coefficients are randomly generated with frequencies in a small band, stamping sets in lower dimensions (e.g., $d < 4$) will sometimes grow somewhat slowly due to this overlap. But generally, stamping sets in larger dimensions will grow in cardinality somewhat quickly.

We also see that the general upper bound (6) is quite rough, though it does properly capture the polynomial versus exponential asymptotic growth in stamping level in the low versus high sparsity cases, respectively. The combinatorial bound (A1) on the other hand is very accurate for predicting the size of a "fully spread out" stamping set.

In Figure 7, we not only depict the sizes of the stamping sets used to solve the PDE (that is, $N = 1, 2, 3$) but also include the size at $N = 4$ which is used to exactly compute the proxy error function $f^{26,3}$ (cf. Section 10.1). In particular, this demonstrates the fact that the algorithm is memory limited by the size of the stamping set for large sparsity values. In the $N = 3$ case, a full stamping set has cardinality $\approx 4 \times 10^4$, which, when used to construct the $\left|\mathcal{S}^3\right| \times \left|\mathcal{S}^3\right|$ Galerkin operator of double precision floats, uses approximately 16 GB in a dense matrix representation. For $N = 4$, the stamp set size is $\approx 6 \times 10^5$, corresponding to an approximately 2.5 TB dense matrix (and the sparse matrix barely fits into the 1 TB of memory on the compute node that we use).

In Figure 8, we also show the relationship between the runtime for constructing and solving the discretized PDE and the runtime in computing the SFT approximation of the data in the high sparsity example from Figure 5. In general, the SFT dominates the runtime, though at stamping level $N = 3$, the ratio of solve to SFT runtime does exceed one in the cases of $d = 1, 4$ (and nearly $d = 16$). We also see the dependence of dimension on the SFT step (the decreasing vertical intercept of each line) and the relative independence of dimension on the stamping and solving procedure (the fixed shape of each line) with the notable exception of $d = 1$ due to the reduced stamp set size. However, we do warn that the parameters for the SFT routines are chosen for higher success rates rather than highly optimized runtimes and the Galerkin operator construction step is parallelized over 64 cores. Using a better optimized and/or parallelized SFT routine may give more competitive runtime splits.

10.4 Multivariate compressible

In order to test Fourier-compressible data which is not exactly sparse, we use a series of tensorized, periodized Gaussians. Here, we present the only details necessary to demonstrate our algorithm's effectiveness on Fourier-compressible data, but for a fuller treatment on the Fourier properties of periodized Gaussians, see e.g., [37, Section 2.1].

Here, we define the periodic Gaussian $G_r : \mathbb{T} \to \mathbb{R}$ by

$$G_r(x) = \frac{\sqrt{2\pi}}{r} \sum_{m=-\infty}^{\infty} \mathrm{e}^{-\frac{(2\pi)^2 (x-m)}{2r^2}}$$

where the dilation-type parameter r allows us to control the effective support of $\hat{G}_r$. In practice, we truncate the infinite sum to $m \in \{-10, \ldots, 10\}$ as additional terms do not change the output up to machine precision. Note

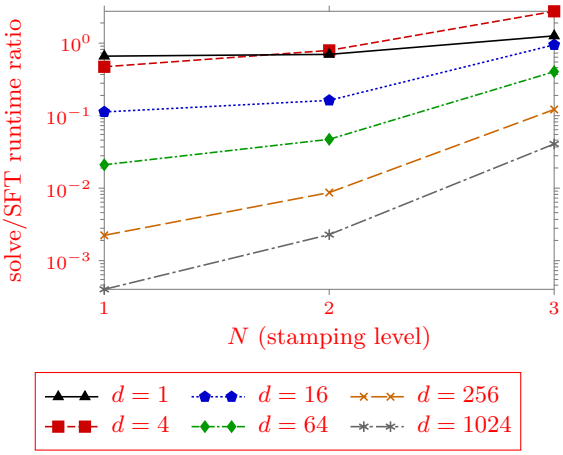


Fig. 8: The ratio of runtimes of Lines 3–5 in Algorithm 1 (the construction and solution of the discretized Galerkin equation) versus Lines 1–2 (the SFT of the PDE data).

here that the nonstandard multiplicative factors help control the behavior of the function in frequency rather than space. Given a multivariate modulating frequency $\mathbf{k} \in \mathbb{Z}^d$, we define the modulated, tensorized, periodic Gaussian by

$$G_{r,\mathbf{k}}(\mathbf{x}) = \prod_{j=1}^{d} \mathrm{e}^{2\pi \mathrm{i} k_j x_j} G_r(x_j).$$

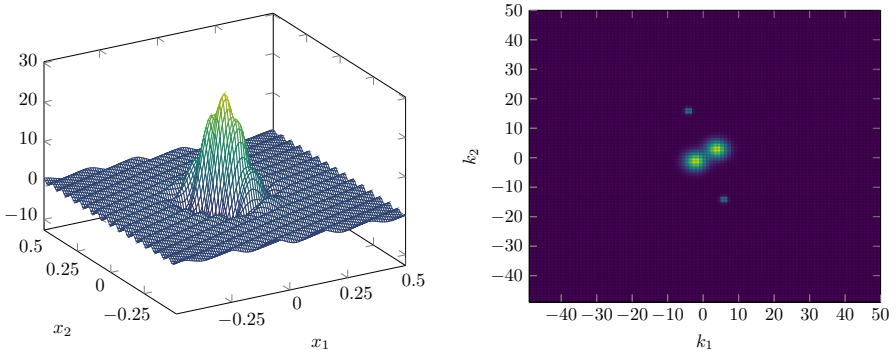
Finally, given a set of frequencies $\mathcal{I} \subseteq \mathbb{Z}^d$, dilation parameters $\mathbf{r} \in \mathbb{R}_+^{\mathcal{I}}$, and coefficients $\mathbf{c} \in \mathbb{R}^{\mathcal{I}}$, we can define Gaussian series

$$G_{\mathbf{c},\mathbf{r}}^{\mathcal{I}}(\mathbf{x}) := \sum_{\mathbf{k}\in\mathcal{I}} c_{\mathbf{k}} G_{r_{\mathbf{k}},\mathbf{k}}(\mathbf{x}).$$

Depending on the severity of the dilations chosen (i.e., $r_{\mathbf{k}} \gg 1$), this can well approximate a Fourier series with frequencies in $\mathcal{I}$. On the other hand, a less severe dilation results in Fourier coefficients with magnitudes forming less concentrated Gaussians centered around the "frequencies" $\mathbf{k} \in \mathcal{I}$ and $-\mathbf{k}$. An example of a series with its associated Fourier transform is given in Figure 9.

In our first experiment, we fix $d = 2$ and vary both stamping level and sparsity to again solve (18). The diffusion coefficient in (18) is replaced with a two-term Gaussian series $a = c_0 + G_{\mathbf{c},\mathbf{r}}^{\mathcal{I}}$, where

$$\begin{aligned} \mathcal{I} \sim \mathrm{Unif}\left(\left([-24,25]^2 \cap \mathbb{Z}^2\right)^2\right), \qquad \mathbf{c} \sim \mathrm{Unif}\left([-1,1]^2\right), \\ \mathbf{r} = 1.1^2\mathbf{1}, \qquad c_0 = 10\left\lceil \|\mathbf{c}\|_2 \right\rceil. \end{aligned} \tag{20}$$



(a) $c_1 G_{r_1,\mathbf{k}_1} + c_2 G_{r_2,\mathbf{k}_2}$ (b) $c_1 \hat{G}_{r_1,\mathbf{k}_1} + c_2 \hat{G}_{r_2,\mathbf{k}_2}$

Fig. 9: An example Gaussian series with $c_1 = c_2 = 1$, $r_1 = 0.5$, $r_2 = 2$, $\mathbf{k}_1 = (3, 2)$, and $\mathbf{k}_2 = (-5, 15)$. The first term corresponds to the wider Gaussian shape and more spread out portions of the Fourier transform. The second term contributes to the highly oscillatory parts and the isolated spikes in the Fourier transform.

Note the increased constant factor from our previous examples to decrease the likelihood of sparse approximations of a not satisfying the ellipticity property. The Fourier transform of the resulting a used for the following test is depicted in Figure 10 below.

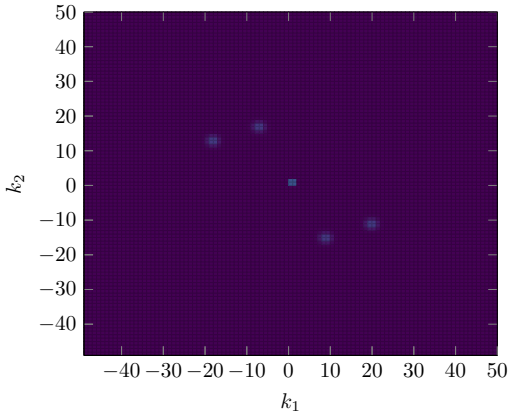


Fig. 10: The specific $\hat{a}$ used in examples depicted in Figure 11.

The diffusion equation is then solved across various sparsities with increasing stamping level. The bandwidth parameter of the SFT is set to $K = 100$ to account for the wider effective support of $\hat{a}$. The Monte Carlo proxy error is computed with 1 000 samples and depicted in Figure 11.

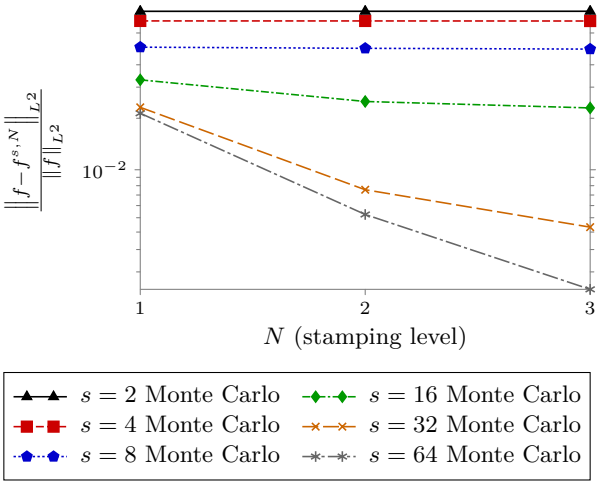


Fig. 11: Proxy error solving (18) with Gaussian series diffusion coefficient with sparsity levels $s = 2, 4, 8, 16, 32, 64$, and stamping levels $N = 1, \ldots, 3$.

Here, the stamping level does not affect convergence until the sparsity is above $s \geq 16$. This demonstrates the tradeoff between sparsity and stamping level in regards to the error bound (15). Until the SFT is able to capture enough useful information in $\hat{a}$, the $\|a - a^s\|_{L^\infty}$ in the error bound dominates. Eventually, this factor is reduced far enough that the stamping term becomes apparent.

We provide another example, where sparsity is fixed at $s = 16$, and dimension and stamping level are increased. Again we solve (18) with the diffusion coefficient replaced by the two-term Gaussian series $a = c_0 + G^{\mathcal{I}}_{\mathbf{c},\mathbf{r}}$, where

$$\mathcal{I} \sim \text{Unif}\left(\left([-249, 250]^d \cap \mathbb{Z}^d\right)^2\right), \quad \mathbf{c} \sim \text{Unif}\left([-1, 1]^2\right),$$
$$\mathbf{r} = 1.1^d \mathbf{1}, \quad c_0 = 10\left\lceil \|\mathbf{c}\|_2 \right\rceil,$$

and $\mathbf{c}$ and c_0 are not redrawn across test cases. The bandwidth of the SFT is set to 1 000 to again account for the potentially widened Fourier transform of a. With a 1 000 point Monte Carlo approximation of the proxy error, the results are given in Figure 12.

Here we observe much the same behavior as the previous test case. This is due to the fact that the dimension additionally drives the sparsity of the Gaussian Fourier transforms based on the choice of dilation $\mathbf{r} = 1.1^d\mathbf{1}$. In additional experiments performed at higher dimensions (not pictured here), this factor results in numerical instability and the approximation error blows up. We also see that the $d = 2$ and $d = 4$ examples are swapped from their assumed positions (and the $d = 2$ case even mildly benefits from increased

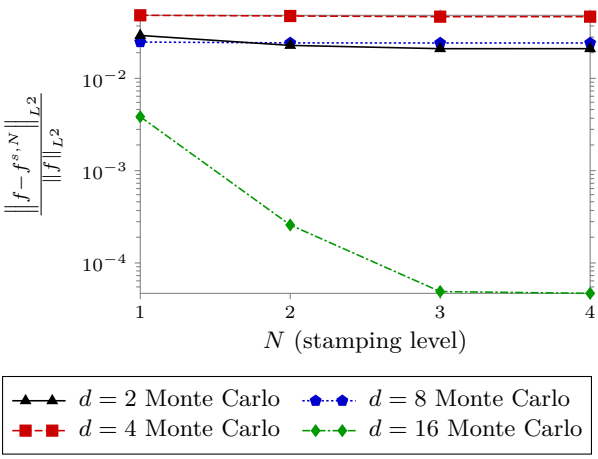


Fig. 12: Approximate proxy error solving (18) with Gaussian series diffusion coefficient with $d = 2$, 4, 8, 16 and $N = 1, \ldots, 5$.

stamping level). This is attributed to the random draw of the frequency locations affecting the proxy error as well as the SFT algorithm performing better in lower dimensions when all parameters are fixed.

Returning to the $d = 2$ example from Figure 11, we take this opportunity to investigate the conditions (13) and (14) ensuring that the terms related to stamping in the error bounds given by Corollaries 4 and 5 decay. Dividing both sides of these inequalities by the right hand side gives a ratio that should be less than one to ensure geometric decay in the stamping term.

In order to check these conditions, we use a $1\,000 \times 1\,000$ two-dimensional FFT of the a described by (20) as the "ground truth" $\hat{a}$. We then compute the two ratios for various sparse approximations as depicted in Figure 13.

We first observe that condition (13) deteriorates as sparsity increases while (14) improves. This is expected, due to (13) corresponding to (7) for the diffusion coefficient a restricted to $\text{supp}(\hat{\mathbf{a}}^s)$ in frequency. As the sparsity increases, this condition gets closer and closer to (7) on the true a.

Additionally, we see that for this diffusion coefficient, only sparsity values $s = 2, 4, 8$ satisfy a condition necessary for decay in the stamping term. This clearly contrasts with the fact that the $s = 16, 32, 64$ cases in Figure 11 are the only examples which do actually decay with the stamping level. However, as stated there, it is possible there is stamping decay in the $s = 2, 4, 8$ terms, but this is being overpowered by the error from a poor sparse approximation. The fact that there is still decay in the larger sparsity cases shows that conditions (13) and (14) are too pessimistic. It remains an open problem as to whether this gap is in producing the more "user friendly" condition (14) or whether the condition (7) can be improved, reducing (13) and (14) as a result.

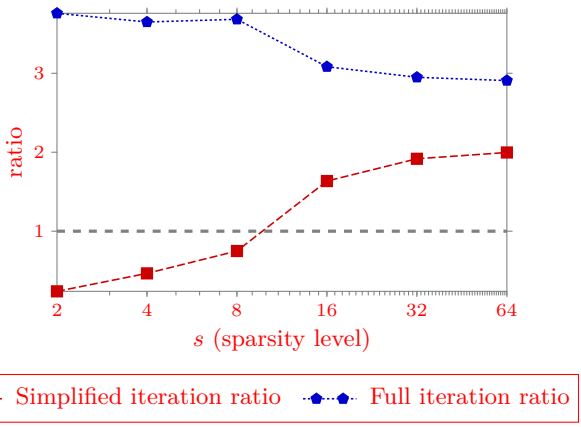


Fig. 13: Comparison of the "simplified iteration ratio" corresponding to (13) and the "full iteration ratio" corresponding to (14) versus the sparsity used in the SFT approximation.

10.5 Three-dimensional exactly sparse advection-diffusion-reaction equation

We now extend our numerical experiments to the situation of a three-dimensional advection-diffusion-reaction equation. See Remark 6 for the PDE setup and necessary algorithmic modifications.

Numerically, we work with the following exactly sparse data:

$$\begin{aligned} a(\mathbf{x}) &= \hat{a}_{\mathbf{0}} + \sum_{\mathbf{k}\in\mathcal{I}_a^{\text{sine}}} c_{a,\mathbf{k}}^{\text{sine}} \sin(2\pi\mathbf{k}\cdot\mathbf{x}) + \sum_{\mathbf{k}\in\mathcal{I}_a^{\text{cosine}}} c_{a,\mathbf{k}}^{\text{cosine}} \cos(2\pi\mathbf{k}\cdot\mathbf{x}) \\ b_j(\mathbf{x}) &= \sum_{\mathbf{k}\in\mathcal{I}_{b_j}^{\text{sine}}} c_{b_j,\mathbf{k}}^{\text{sine}} \sin(2\pi\mathbf{k}\cdot\mathbf{x}) + \sum_{\mathbf{k}\in\mathcal{I}_{b_j}^{\text{cosine}}} c_{b_j,\mathbf{k}}^{\text{cosine}} \cos(2\pi\mathbf{k}\cdot\mathbf{x}) \text{ for all } j=1,2,3 \\ c(\mathbf{x}) &= \hat{c}_{\mathbf{0}} + \sum_{\mathbf{k}\in\mathcal{I}_c^{\text{sine}}} c_{c,\mathbf{k}}^{\text{sine}} \sin(2\pi\mathbf{k}\cdot\mathbf{x}) + \sum_{\mathbf{k}\in\mathcal{I}_c^{\text{cosine}}} c_{c,\mathbf{k}}^{\text{cosine}} \cos(2\pi\mathbf{k}\cdot\mathbf{x}) \\ f(\mathbf{x}) &= \sum_{\mathbf{k}\in\mathcal{I}_f^{\text{sine}}} c_{f,\mathbf{k}}^{\text{sine}} \sin(2\pi\mathbf{k}\cdot\mathbf{x}) + \sum_{\mathbf{k}\in\mathcal{I}_f^{\text{cosine}}} c_{f,\mathbf{k}}^{\text{cosine}} \cos(2\pi\mathbf{k}\cdot\mathbf{x}), \end{aligned} \tag{21}$$

where

$$\begin{gathered} \left|\mathcal{I}_a^{\text{sine}}\right| = \left|\mathcal{I}_a^{\text{cosine}}\right| = 2 \\ \left|\mathcal{I}_{b_j}^{\text{sine}}\right| = \left|\mathcal{I}_{b_j}^{\text{cosine}}\right| = \left|\mathcal{I}_c^{\text{sine}}\right| = \left|\mathcal{I}_{\mathbf{c}}^{\text{cosine}}\right| = 5 \text{ for all } j=1,2,3 \\ \left|\mathcal{I}_f^{\text{sine}}\right| = 2, \text{ and } \left|\mathcal{I}_f^{\text{cosine}}\right| = 3. \end{gathered}$$

In total, there are 46 terms composing the differential operator, and 5 terms composing the forcing function. Each frequency is randomly drawn from

$\mathrm{Unif}([-49, 50]^3 \cap \mathbb{Z}^3)$ and each coefficient for a and f from $\mathrm{Unif}([-1, 1])$. The coefficients for $\mathbf{b}$ and c are drawn from $\mathrm{Unif}([0, 1])$. To ensure well-posedness, $\hat{a}_{\mathbf{0}} = 4\left\lceil \sqrt{\|c_a^{\mathrm{sine}}\|_2^2 + \|c_a^{\mathrm{cosine}}\|_2^2} \right\rceil$, and $\hat{c}_{\mathbf{0}} = 4\left\lceil \sqrt{\|c_c^{\mathrm{sine}}\|_2^2 + \|c_c^{\mathrm{cosine}}\|_2^2} \right\rceil$. The bandwidth of the SFT is set to $K = 100$ and we consider sparsity levels $s = 2, 5$ and 10. Due to the large size of the stamp, we only consider stamping levels $N = 1, 2$.

		$\|f - f^{s,N}\|_{L^2} / \|f\|_{L^2}$	
s	N	exact	Monte Carlo
2	1	0.518	0.518
	2	0.518	0.518
5	1	0.054	0.054
	2	0.031	0.031
10	1	0.047	0.047
	2	0.012	0.012

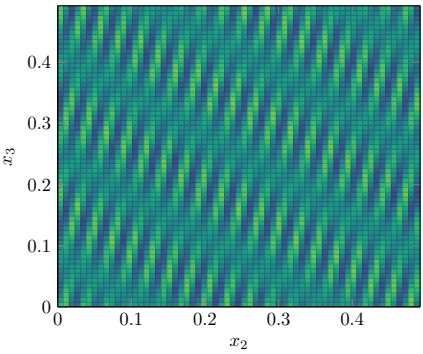
Table 1: Error in approximating solution to ADR equation (16).

The resulting true and Monte Carlo proxy error (sampled over 1 000 points) is given in Table 1. Additionally, Figure 14 shows a portion of a slice through f as well as $f^{2,1}$ and $f^{10,2}$ which are computed by passing $u^{2,1}$ and $u^{10,2}$ through the differential operator.

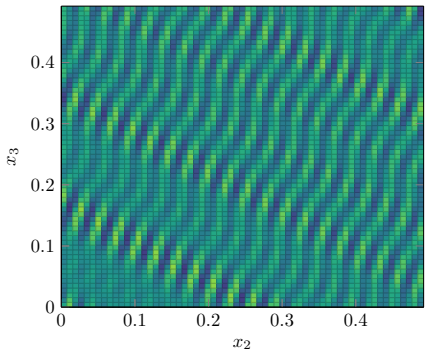
We note that $f^{10,2}$ and f appear qualitatively indistinguishable. However, since the sparsity level, $s = 2$, used to compute $u^{2,1}$ is lower than the sparsity of any term in (21), $f^{2,1}$ loses some of characteristics of the original source term. Though it captures some of the true behavior in both larger scales (e.g., the oscillations moving in the northeast direction) and finer scales (e.g., the oscillations moving in the southeast direction), some interfering modes which produce the "wavy" effect are left out. This is supported by the relative errors reported in Table 1. Note also that the stamping level affects the convergence in the $s = 5$ and $s = 10$ cases, but not in the $s = 2$ case. This is due to the sparsity related errors in (15) overwhelming the stamping term until the SFT approximations of the data are accurate enough.

11 Conclusion

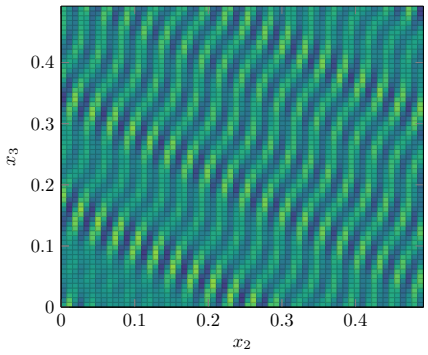
In this paper, we have presented a new way to join the theory of spectral methods and sparse Fourier transforms for solving diffusion equations. The key contribution is the idea of stamping sets, which allow us to find the provably most important frequencies of the PDE solution by running efficient,



(a) Slice through $f^{2,1}$.



(b) Slice through $f^{10,2}$.



(c) Slice through f.

Fig. 14: Samples of $f^{10,2}$ and f on the $x_1 = 63/128$ plane.

high-dimensional SFTs on the PDE data. This recharacterizes the Galerkin solution step into a procedure which is no longer directly dependent on either the spatial dimension of the problem or the solution's effective bandwidth. We have demonstrated the method's applicability in solving problems with both bandwidths and dimensions over 1 000, with sparse or compressible data, and in generalizations to ADR equations. We also provide a full H^1 convergence guarantee in Corollary 5, with a condition on the PDE data and sparse approximation that can be checked to ensure convergence.

However, as demonstrated in Section 10.4, we believe that condition (14) which ensures that the estimate in Corollary 5 holds is too pessimistic. In future work, it would be useful to better understand the relationship between condition (14) and conditions (13) and (7) from which it is derived. Since all are tightly coupled with decay in their respective upper bounds, it is likely that any improvement in these conditions would also result in improved convergence guarantees.

Another area of future investigation would be the extension of these methods to time-dependent problems, especially those exhibiting spatially multiscale behavior. Rather than needing to use a very fine grid to resolve or eventually adapt to any high-frequency behavior, we anticipate that SFTs and stamping sets can be used to directly find the solution's most important frequencies at various time steps. This would ideally be able to accelerate the time-stepping procedure while also potentially allowing for a stamp-adapted stability condition to relax time-step sizes.

Finally, the numerical experiments in this paper have demonstrated that our sparse spectral method can achieve high accuracy at low-sparsity levels, and moderate accuracy for moderate sparsity levels (≈ 50 complex Fourier coefficients). Thus, we do not expect that this technique will be easily applicable to high-dimensional problems without a strong sparsity assumption. However, we do see its viability in accelerating existing sparse grid or other high-dimensional PDE solvers. In particular, stamping sets can be used to better understand and adapt to anisotropic behavior in higher dimensions. Applying a quick SFT on the diffusion coefficient and observing the shape of the output's support can help show the dimensions in which a solution will have important high-frequency information. Additionally, one could also consider restricting attention to a stamping set within a fixed frequency truncation (e.g., a hyperbolic cross), to quickly create an adaptive spectral basis. An especially interesting application of this approach could be in weighted ℓ^1 minimization in compressed sensing [1, 40, 41], where polynomial basis coefficient indices are penalized in recovery based on their likelihood to include important coefficients in the recovered solution. The stamping level in which these frequencies lie could, e.g., be used as an additional penalization based on the truncation analysis in Lemma 3.

Acknowledgments. This work was supported in part by the National Science Foundation Award Numbers DMS 2106472 and 1912706.

This work was also supported in part through computational resources and services provided by the Institute for Cyber-Enabled Research at Michigan State University.

We thank Lutz Kämmerer for helpful discussions related to random rank-1 lattice construction and Ben Adcock and Simone Brugiapaglia for motivating discussions related to compressive sensing and high-dimensional PDEs. We also thank the reviewers of this manuscript for their insightful comments and suggestions.

Appendix A Stamp set cardinality bound

We begin by proving the following combinatorial upper bound for the cardinality of a stamp set.

Lemma 8. *Suppose that* $\mathbf{0} \in \operatorname{supp}(\hat{a})$, $\operatorname{supp}(\hat{a}) = -\operatorname{supp}(\hat{a})$, *and* $|\operatorname{supp}(\hat{a})| = s$. *Then*

$$\left|\mathcal{S}^N[\hat{a}](\operatorname{supp}(\hat{f}))\right| \leq \left|\operatorname{supp}(\hat{f})\right| \sum_{n=0}^{N} \sum_{t=0}^{\min(n,(s-1)/2)} 2^t \binom{(s-1)/2}{t} \binom{n-1}{t-1}. \quad \text{(A1)}$$

Proof We begin by separating $\mathcal{S}^N$ into the disjoint pieces

$$\mathcal{S}^N = \bigsqcup_{n=0}^{N} \left(\mathcal{S}^n \setminus \left(\bigcup_{i=0}^{n-1} \mathcal{S}^i\right)\right)$$

and computing the cardinality of each of these sets (where we take $S^{-1} = \emptyset$). If $\mathbf{k} \in \mathcal{S}^n \setminus \left(\cup_{i=0}^{n-1} \mathcal{S}^i\right)$, then we are able to write $\mathbf{k}$ as

$$\mathbf{k} = \mathbf{k}_f + \sum_{m=1}^{n} \mathbf{k}_a^m \quad \text{(A2)}$$

where $\mathbf{k}_f \in \operatorname{supp}(\hat{f})$ and $\mathbf{k}_a^m \in \operatorname{supp}(\hat{a}) \setminus \{0\}$ for all $m = 1, \ldots, n$. Additionally, since $\mathbf{k}$ is not in any earlier stamping sets, this is the smallest n for which this is possible. In particular, it is not possible for any two frequencies in the sum to be negatives of each other resulting in pairs of cancelled terms.

With this summation in mind, arbitrarily split $\operatorname{supp}(\hat{a}) \setminus \{\mathbf{0}\}$ into $A \sqcup -A$ (i.e., place all frequencies which do not negate each other into A and their negatives in $-A$). By collecting like frequencies that occur as a $\mathbf{k}_a^m$ term in (A2), we can rewrite this sum as

$$\mathbf{k} = \mathbf{k}_f + \sum_{\mathbf{k}_a \in A} s(\mathbf{k}, \mathbf{k}_a) m(\mathbf{k}, \mathbf{k}_a) \mathbf{k}_a, \quad \text{(A3)}$$

where the sign function $s(\mathbf{k}, \mathbf{k}_a)$ is given by

$$s(\mathbf{k}, \mathbf{k}_a) := \begin{cases} 1 & \text{if } \mathbf{k}_a \text{ is a term in the summation (A2)} \\ -1 & \text{if } -\mathbf{k}_a \text{ is a term in the summation (A2)} \\ 0 & \text{otherwise} \end{cases}$$

and the multiplicity function $m(\mathbf{k}, \mathbf{k}_a)$ is defined as the number of times that $\mathbf{k}_a$ or $-\mathbf{k}_a$ appears as a $\mathbf{k}_a^m$ term in (A2). Letting $\mathbf{s}(\mathbf{k}) := (s(\mathbf{k}, \mathbf{k}_a))_{\mathbf{k}_a \in A}$ and $\mathbf{m}(\mathbf{k}) := (m(\mathbf{k}, \mathbf{k}_a))_{\mathbf{k}_a \in A}$, we can then identify any $\mathbf{k} \in \mathcal{S}^n \setminus \left(\cup_{i=0}^{n-1} \mathcal{S}^i\right)$ with the tuple

$$(\mathbf{k}_f, \mathbf{s}(\mathbf{k}), \mathbf{m}(\mathbf{k})) \in \operatorname{supp}(\mathbf{f}) \times \{-1, 0, 1\}^A \times \{0, \ldots, n\}^A.$$

Upper bounding the number of these tuples that can correspond to a value of $\mathbf{k} \in \mathcal{S}^n \setminus \left(\cup_{i=0}^{n-1} \mathcal{S}^i\right)$ will then upper bound the cardinality of this set.

Since any $\mathbf{k}_f \in \operatorname{supp}(\hat{f})$ can result in a valid $\mathbf{k}$ value, we will focus on the pairs of sign and multiplicity vectors. Define by $T_n \subseteq \{-1, 0, 1\}^A \times \{0, \ldots, n\}^A$ the set of valid sign and multiplicity pairs that can correspond to a $\mathbf{k} \in \mathcal{S}^n \setminus \left(\cup_{i=0}^{n-1} \mathcal{S}^i\right)$. In particular, for $(\mathbf{s}, \mathbf{m}) \in T_n$, $\|\mathbf{m}\|_1 = n$ and $\operatorname{supp}(\mathbf{s}) = \operatorname{supp}(\mathbf{m})$. Thus, we can write

$$T_n \subseteq \bigsqcup_{t=0}^{\min(n,|A|)} \Big\{(\mathbf{s}, \mathbf{m}) \in \{-1, 0, 1\}^A \times \{0, \ldots, n\}^A \\ \mid \|\mathbf{m}\|_1 = n \text{ and } |\operatorname{supp}(\mathbf{s})| = |\operatorname{supp}(\mathbf{m})| = t\Big\}.$$

This inner set then corresponds to the t-partitions of the integer n spread over the $|A|$ entries of $\mathbf{m}$ where each non-zero term is assigned a sign -1 or 1. The cardinality is therefore $2^t\binom{|A|}{t}\binom{n-1}{t-1}$: the first factor is from the possible sign options, the second is the number of ways to choose the entries of $\mathbf{m}$ which are nonzero, and the last is the number of t-partitions of n which will fill the nonzero entries of $\mathbf{m}$. Noting that $|A| = \frac{s-1}{2}$, our final cardinality estimate is

$$\begin{aligned}\left|\mathcal{S}^N\right| &= \sum_{n=0}^{N}\left|\mathcal{S}^n \setminus \left(\bigcup_{i=0}^{n-1}\mathcal{S}^i\right)\right| \\ &\leq \sum_{n=0}^{N}\left|\operatorname{supp}(\hat{f})\right||T_n| \\ &\leq \left|\operatorname{supp}(\hat{f})\right|\sum_{n=0}^{N}\sum_{t=0}^{\min(n,(s-1)/2)} 2^t\binom{(s-1)/2}{t}\binom{n-1}{t-1}\end{aligned}$$

as desired. □

Though this upper bound is much tighter than the one given in the main text, it is harder to parse. As such, we simplify it to the bound presented in Lemma 2, restated here for convenience.

Lemma 2. *Suppose that* $\mathbf{0} \in \operatorname{supp}(\hat{a})$, $\operatorname{supp}(\hat{a}) = -\operatorname{supp}(\hat{a})$, *and* $|\operatorname{supp}(\hat{a})| = s$. *Then*

$$\left|\mathcal{S}^N[\hat{a}](\operatorname{supp}(\hat{f}))\right| \leq 6\left|\operatorname{supp}(\hat{f})\right|\left(\frac{\max(s,2N)}{\sqrt{2}}\right)^{\min(s,2N)}.$$

Proof Let $r = (s-1)/2$. We consider two cases:

Case 1: ($s \geq 2N \implies r \geq N$) We estimate the innermost sum of (A1). Since $r \geq N \geq n$, $\min(n,(s-1)/2) = n$. By upper bounding the binomial coefficients with powers of r, we obtain

$$\begin{aligned}\sum_{t=0}^{n} 2^t\binom{r}{t}\binom{n-1}{t-1} &\leq \sum_{t=0}^{n} 2^t(r^t)^2 \\ &\leq 2(2r^2)^n\end{aligned}$$

where the second estimate follows from approximating the geometric sum. Again, bounding the next geometric sum by double the largest term, we have

$$\left|\mathcal{S}^N\right| \leq \left|\operatorname{supp}(\hat{f})\right|\sum_{n=0}^{N} 2(2r^2)^n \leq \left|\operatorname{supp}(\hat{f})\right|4(2r^2)^N \leq 4\left|\operatorname{supp}(\hat{f})\right|\left(\frac{s}{\sqrt{2}}\right)^{2N}.$$

Case 2: ($s < 2N \implies r < N$) Bounding the innermost sum of (A1) proceeds much the same way as Case 1, but we must first split the outermost sum into the first $r+1$ terms and last $N-r$ terms. Working with the first terms, we find

$$\sum_{n=0}^{r}\sum_{t=0}^{n} 2^t\binom{r}{t}\binom{n-1}{t-1} \leq 4(2r^2)^r$$

using the argument in Case 1. Now, we bound

$$\begin{aligned}\sum_{n=r+1}^{N}\sum_{t=0}^{r} 2^t\binom{r}{t}\binom{n-1}{t-1} &\leq \sum_{n=r+1}^{N} 2(2(n-1)^2)^r \\ &\leq 2N(2(N-1)^2)^r \\ &\leq \sqrt{2}(\sqrt{2}N)^{2r+1}.\end{aligned}$$

Thus,

$$\left|\mathcal{S}^N\right| \leq \left|\mathrm{supp}(\hat{f})\right|\left[4(2r^2)^r + \sqrt{2}(\sqrt{2}N)^{2r+1}\right] \leq (4+\sqrt{2})\left|\mathrm{supp}(\hat{f})\right|\left(\frac{2N}{\sqrt{2}}\right)^s.$$

Combining the two cases gives the desired upper bound.

□

Appendix B Proof of SFT recovery guarantees

We restate the theorem for convenience.

Theorem 2 ([26], Corollary 2)**.** *Let $\mathcal{I} \subseteq \mathbb{Z}^d$ be a frequency set of interest with expansion defined as $K := \max_{j\in\{1,\ldots,d\}}(\max_{\mathbf{k}\in\mathcal{I}} k_j - \min_{\mathbf{l}\in\mathcal{I}} l_j) + 1$ (i.e., the sidelength of the smallest hypercube containing $\mathcal{I}$), and $\Lambda(\mathbf{z}, M)$ be a reconstructing rank-1 lattice for $\mathcal{I}$.*

There exists a fast, randomized SFT which, given $\Lambda(\mathbf{z}, M)$, sampling access to $g \in L^2$, and a failure probability $\sigma \in (0, 1]$, will produce a $2s$-sparse approximation $\hat{\mathbf{g}}^s$ of $\hat{g}$ and function $g^s := \sum_{\mathbf{k}\in\mathrm{supp}(\hat{\mathbf{g}}^s)} \hat{g}^s_{\mathbf{k}} e_{\mathbf{k}}$ approximating g satisfying

$$\|g - g^s\|_{L^2} \leq \|\hat{g} - \hat{\mathbf{g}}^s\|_{\ell^2} \leq (25+3K)\left[\frac{\|\hat{g}|_{\mathcal{I}} - (\hat{g}|_{\mathcal{I}})^{\mathrm{opt}}_s\|_1}{\sqrt{s}} + \sqrt{s}\|\hat{g} - \hat{g}|_{\mathcal{I}}\|_1\right]$$

with probability exceeding $1-\sigma$. If $g \in L^\infty$, then we additionally have

$$\|g - g^s\|_{L^\infty} \leq \|\hat{g} - \hat{\mathbf{g}}^s\|_{\ell^1} \leq (33+4K)\left[\left\|\hat{g}|_{\mathcal{I}} - (\hat{g}|_{\mathcal{I}})^{\mathrm{opt}}_s\right\|_1 + \|\hat{g} - \hat{g}|_{\mathcal{I}}\|_1\right]$$

with the same probability estimate. The total number of samples of g and computational complexity of the algorithm can be bounded above by

$$\mathcal{O}\left(ds\log^3(dKM)\log\left(\frac{dKM}{\sigma}\right)\right).$$

Proof The L^2 upper bound is mostly the same as the original result. We are not considering noisy measurements here which removes the $\sqrt{s}e_\infty$ term from that result (though, this could be added back in if desired). Additionally, we have upper bounded $\|\hat{g} - \hat{g}|_{\mathcal{I}}\|_2$ by $\sqrt{s}\|\hat{g} - \hat{g}|_{\mathcal{I}}\|_1$ adding one to the constant.

The L^∞ / ℓ^1 bound was not given in the original paper, but can be proven using the same techniques. In particular, replacing the ℓ^2 norm by the ℓ^1 norm in [26,

Lemma 4] has the effect of replacing all ℓ^2 norms with ℓ^1 norms and replacing $\sqrt{2s}$ by $2s$. This small change cascades through the proof of Property 3 in [26, Theorem 2] (again, with ℓ^2 norms replaced by ℓ^1 norms) to produce the univariate ℓ^1 upper bound (in the language of the original paper)

$$\|\hat{\mathbf{a}} - \mathbf{v}\|_1 \leq \left\|\hat{\mathbf{a}} - \hat{\mathbf{a}}_{2s}^{\text{opt}}\right\|_1 + (16 + 6\sqrt{2})\left(\left\|\hat{\mathbf{a}} - \hat{\mathbf{a}}_{s}^{\text{opt}}\right\|_1 + s(\|\hat{a} - \hat{\mathbf{a}}\|_1 + \|\mu\|_\infty)\right) =: \eta_1.$$

A similar logic applies to revising the proof of [26, Lemma 1]. Equation (4) with all ℓ^2 norms replaced by ℓ^1 norms is derived the same way, and the first term is upper bounded by the maximal entry of the vector multiplied by the number of elements without the square root. The remainder of the proof carries through without change which leads to a final error estimate of

$$\|\mathbf{b} - c\|_{\ell^2} \leq (\beta + \eta_\infty)\max(s - |\mathcal{S}_\beta|, 0) + \eta_1 + \left\|c|_{\mathcal{I}} - c|_{\mathcal{S}_\beta}\right\|_1 + \|c - c|_{\mathcal{I}}\|_1.$$

Finally, the proof of [26, Corollary 2] follows using the same logic as the original substituting these revised upper bounds. □

References

[1] Ben Adcock, Simone Brugiapaglia, and Clayton G. Webster, *Sparse polynomial approximation of high-dimensional functions*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2022.

[2] Sina Bittens, Ruochuan Zhang, and Mark A Iwen, *A deterministic sparse FFT for functions with structured Fourier sparsity*, Advances in Computational Mathematics **45** (2019), no. 2, 519–561.

[3] John P. Boyd, *Chebyshev and Fourier spectral methods*, second, rev ed., Dover Publications, Mineola, N.Y, 2001.

[4] S Brugiapaglia, S Micheletti, F Nobile, and S Perotto, *Wavelet-Fourier CORSING techniques for multidimensional advection-diffusion-reaction equations*, IMA Journal of Numerical Analysis (2020), no. draa036.

[5] S. Brugiapaglia, S. Micheletti, and S. Perotto, *Compressed solving: A numerical approximation technique for elliptic PDEs based on compressed sensing*, Computers & Mathematics with Applications **70** (2015), no. 6, 1306–1335 (en).

[6] Simone Brugiapaglia, *COmpRessed SolvING: Sparse Approximation of PDEs based on compressed sensing*, Ph.D. thesis, Polytecnico Di Milano, Milan, Italy, January 2016.

[7] Simone Brugiapaglia, *A compressive spectral collocation method for the diffusion equation under the restricted isometry property*, Quantification of Uncertainty: Improving Efficiency and Technology: QUIET selected contributions (Marta D'Elia, Max Gunzburger, and Gianluigi Rozza, eds.), Lecture Notes in Computational Science and Engineering, Springer International Publishing, Cham, 2020, pp. 15–40 (en).

[8] Simone Brugiapaglia, Sjoerd Dirksen, Hans Christian Jung, and Holger Rauhut, *Sparse recovery in bounded Riesz systems with applications to numerical methods for PDEs*, Applied and Computational Harmonic Analysis **53** (2021), 231–269 (en).

[9] Simone Brugiapaglia, Fabio Nobile, Stefano Micheletti, and Simona Perotto, *A theoretical study of COmpRessed SolvING for advection-diffusion-reaction problems*, Mathematics of Computation **87** (2018), no. 309, 1–38 (en).

[10] Hans-Joachim Bungartz and Michael Griebel, *Sparse grids*, Acta Numerica **13** (2004), 147–269 (en).

[11] Claudio Canuto, M. Yousuff Hussaini, Alfio Quarteroni, and Thomas A. Zang, *Spectral methods: Fundamentals in single domains*, Scientific Computation, Springer-Verlag, Berlin Heidelberg, 2006 (en).

[12] H. Cho, D. Venturi, and G.E. Karniadakis, *Numerical methods for high-dimensional probability density function equations*, Journal of Computational Physics **305** (2016), 817–837 (en).

[13] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore, *Compressed sensing and best k-term approximation*, Journal of the American Mathematical Society **22** (2009), no. 1, 211–231 (en).

[14] Dinh Dũng, Vladimir Temlyakov, and Tino Ullrich, *Hyperbolic cross approximation*, Advanced Courses in Mathematics - CRM Barcelona, Springer International Publishing, Cham, 2018 (en).

[15] Ingrid Daubechies, Olof Runborg, and Jing Zou, *A sparse spectral method for homogenization multiscale problems*, Multiscale Modeling & Simulation **6** (2007), no. 3, 711–740.

[16] Michael Döhler, Stefan Kunis, and Daniel Potts, *Nonequispaced hyperbolic cross fast Fourier transform*, SIAM Journal on Numerical Analysis **47** (2010), no. 6, 4415–4428.

[17] Weinan E, Jiequn Han, and Arnulf Jentzen, *Algorithms for solving high dimensional PDEs: from nonlinear Monte Carlo to machine learning*, Nonlinearity **35** (2021), no. 1, 278 (en), Publisher: IOP Publishing.

[18] Lawrence C. Evans, *Partial differential equations*, second ed., Graduate studies in mathematics, no. v. 19, American Mathematical Society, Providence, R.I, 2010.

[19] Anna C Gilbert, Sudipto Guha, Piotr Indyk, Shanmugavelayutham Muthukrishnan, and Martin Strauss, *Near-optimal sparse Fourier representations via sampling*, Proceedings of the thiry-fourth annual ACM symposium on Theory of computing, 2002, pp. 152–161.

[20] Anna C Gilbert, Piotr Indyk, Mark Iwen, and Ludwig Schmidt, *Recent developments in the sparse Fourier transform: A compressed Fourier transform for big data*, IEEE Signal Processing Magazine **31** (2014), no. 5, 91–100.

[21] Gene H. Golub and Charles F. Van Loan, *Matrix computations*, fourth ed., Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 2013.

[22] V Gradinaru, *Fourier transform on sparse grids: Code design and the time dependent Schrödinger equation*, Computing (Wien. Print) **80** (2007), no. 1, 1–22.

[23] Michael Griebel and Jan Hamaekers, *Sparse grids for the Schrödinger equation*, Special issue on molecular modelling **41** (2007), no. 2, 215–247.

[24] Michael Griebel and Jan Hamaekers, *Fast discrete Fourier transform on generalized sparse grids*, Sparse Grids and Applications - Munich 2012 (Jochen Garcke and Dirk Pflüger, eds.), vol. 97, Springer International Publishing, Cham, 2014, pp. 75–107 (en).

[25] Craig Gross, *Sparsity in the spectrum: sparse Fourier transforms and spectral methods for functions of many dimensions*, Ph.D., Michigan State University, East Lansing, Michigan, USA, May 2023.

[26] Craig Gross, Mark Iwen, Lutz Kämmerer, and Toni Volkmer, *Sparse Fourier transforms on rank-1 lattices for the rapid and low-memory approximation of functions of many variables*, Sampling Theory, Signal Processing, and Data Analysis **20** (2021), no. 1, 1.

[27] Craig Gross, Mark A Iwen, Lutz Kämmerer, and Toni Volkmer, *A deterministic algorithm for constructing multiple rank-1 lattices of near-optimal size*, Advances in Computational Mathematics **47** (2021), no. 6, 1–24.

[28] Tristan Guillaume, *On the multidimensional Black-Scholes partial differential equation*, Annals of Operations Research **281** (2019), no. 1, 229–251 (en).

[29] Haitham Hassanieh, Piotr Indyk, Dina Katabi, and Eric Price, *Simple and practical algorithm for sparse Fourier transform*, Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms, SIAM, 2012, pp. 1183–1194.

[30] Mark A Iwen, *Combinatorial sublinear-time Fourier algorithms*, Foundations of Computational Mathematics **10** (2010), no. 3, 303–338.

[31] Dante Kalise and Karl Kunisch, *Polynomial approximation of high-dimensional Hamilton–Jacobi–Bellman equations and applications to feedback control of semilinear parabolic PDEs*, SIAM Journal on Scientific Computing **40** (2018), no. 2, A629–A652, Publisher: Society for Industrial and Applied Mathematics.

[32] Frances Kuo, Giovanni Migliorati, Fabio Nobile, and Dirk Nuyens, *Function integration, reconstruction and approximation using rank-1 lattices*, Mathematics of Computation **90** (2021), no. 330, 1861–1897 (en).

[33] Friedrich Kupka, *Sparse grid spectral methods for the numerical solution of partial differential equations with periodic boundary conditions*, Ph.D., Universität Wien, Vienna, Austria, November 1997.

[34] Lutz Kämmerer, Stefan Kunis, and Daniel Potts, *Interpolation lattices for hyperbolic cross trigonometric polynomials*, Journal of Complexity **28** (2012), no. 1, 76–92 (en).

[35] Lutz Kämmerer, Daniel Potts, and Toni Volkmer, *Approximation of multivariate periodic functions by trigonometric polynomials based on rank-1 lattice sampling*, Journal of Complexity **31** (2015), no. 4, 543–576 (en).

[36] Dong Li and Fred J. Hickernell, *Trigonometric spectral collocation methods on lattices*, Recent advances in scientific computing and partial differential equations (Hong Kong, 2002), Contemp. Math., vol. 330, Amer. Math. Soc., Providence, RI, 2003, pp. 121–132. MR 2011715

[37] Sami Merhi, Ruochuan Zhang, Mark A. Iwen, and Andrew Christlieb, *A new class of fully discrete sparse Fourier transforms: Faster stable implementations with guarantees*, Journal of Fourier Analysis and Applications **25** (2019), no. 3, 751–784 (en).

[38] Hans Munthe-Kaas and Tor Sørevik, *Multidimensional pseudo-spectral methods on lattice grids*, Applied Numerical Mathematics **62** (2012), no. 3, 155–165 (en).

[39] Gerlind Plonka, Daniel Potts, Gabriele Steidl, and Manfred Tasche, *Numerical Fourier analysis*, Applied and Numerical Harmonic Analysis, Springer International Publishing, Cham, 2018 (en).

[40] Holger Rauhut and Christoph Schwab, *Compressive sensing Petrov-Galerkin approximation of high-dimensional parametric operator equations*, Mathematics of Computation **86** (2017), no. 304, 661–700 (en).

[41] Holger Rauhut and Rachel Ward, *Interpolation via weighted ℓ_1 minimization*, Applied and Computational Harmonic Analysis **40** (2016), no. 2, 321–351 (en).

[42] A.D. Rubio, A. Zalts, and C.D. El Hasi, *Numerical solution of the advection-reaction-diffusion equation at different scales*, Environmental Modelling & Software **23** (2008), no. 1, 90–95 (en).

[43] Jie Shen and Li-Lian Wang, *Sparse spectral approximations of high-dimensional problems based on hyperbolic cross*, SIAM Journal on Numerical Analysis **48** (2010), no. 3, 1087–1109.

[44] Weiqi Wang and Simone Brugiapaglia, *Compressive fourier collocation methods for high-dimensional diffusion equations with periodic boundary conditions*, 2022.

[45] H. Yserentant, *Sparse grid spaces for the numerical solution of the electronic Schrödinger equation*, Numerische Mathematik **101** (2005), no. 2, 381–389 (en).

[46] Harry Yserentant, *On the regularity of the electronic Schrödinger equation in Hilbert spaces of mixed derivatives*, Numerische Mathematik **98** (2004), no. 4, 731–759 (en).