MULTISCALE MODEL. SIMUL.
Vol. 6, No. 3, pp. 711–740

A SPARSE SPECTRAL METHOD FOR HOMOGENIZATION MULTISCALE PROBLEMS*

INGRID DAUBECHIES†, OLOF RUNBORG‡, AND JING ZOU§

Abstract. We develop a new sparse spectral method, in which the fast Fourier transform (FFT) is replaced by RAℓSFA (randomized algorithm of sparse Fourier analysis); this is a sublinear randomized algorithm that takes time $O(B \log N)$ to recover a B-term Fourier representation for a signal of length N, where we assume $B \ll N$. To illustrate its potential, we consider the parabolic homogenization problem with a characteristic fine scale size ε. For fixed tolerance the sparse method has a computational cost of $O(|\log \varepsilon|)$ per time step, whereas standard methods cost at least $O(\varepsilon^{-1})$. We present a theoretical analysis as well as numerical results; they show the advantage of the new method in speed over the traditional spectral methods when ε is very small. We also show some ways to extend the methods to hyperbolic and elliptic problems.

Key words. multiscale methods, sublinear algorithm, homogenization, spectral methods, sparse Fourier representation

AMS subject classifications. 65M70, 35K05, 65T50, 68W20

DOI. 10.1137/060676258

1. Introduction.

Multiscale modeling and computation have attracted a huge amount of attention in recent years, with the interest stemming mainly from multiscale problems in applied fields such as materials science, chemistry, complex fluids, and biology. Multiscale problems involve phenomena taking place on vastly different time and/or spatial scales. The influence of the small scales is important for the large scale behavior, but they are very expensive to simulate directly with numerical methods. For many practical problems, traditional computational methods are prohibitively expensive.

The goal of multiscale methods is to find an efficient way to incorporate the fine scales' effect in the numerical solution of the coarse dynamics. One way to do this is to analytically derive "effective" equations, which model the fine scale effects. This is done, for instance, in averaging [2], homogenization [4], and boundary-layer analysis [22]. These techniques are very useful when they are applicable. For general problems, however, there is typically no simple way to derive closed effective models. Another approach is taken by a more recent class of numerical methods, which model the fine scale effect numerically; in a sense they replace the manual derivation of effective equations by direct numerical simulation of the fine scale equations in small domains over a short time. Some early examples of this type of method are the Car–Parrinello method in quantum chemistry [8], the kinetic-hydrodynamic models of complex fluids [7], and the quasi-continuum method in solid mechanics [28]. Recently, a more comprehensive view of these approaches has been taken and put in

*Received by the editors November 29, 2006; accepted for publication (in revised form) April 4, 2007; published electronically August 29, 2007. This work was partially supported by NSF grant DMS-0219233 and the SSF programme "Advanced Multiscale Techniques for New Discoveries in Science and Engineering."
http://www.siam.org/journals/mms/6-3/67625.html
†Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544 (ingrid@math.princeton.edu).
‡Kungl Tekniska Högskolan, 10044 Stockholm, Sweden (olofr@nada.kth.se).
§Center for Scientific Computation and Mathematical Modeling (CSCAMM), University of Maryland, College Park, MD 20742 (jzou@cscamm.umd.edu).

frameworks, which more systematically exploits this idea. The "equation-free" methods of Kevrekidis et al. [23, 29] and the heterogeneous multiscale method [12, 13] are examples of this. An overview listing more multiscale approaches is given in [11].

In this paper we will study the use of sublinear Fourier algorithms in the context of multiscale problems. This is a recently developed type of discrete Fourier transform method with a time complexity significantly smaller than $O(N)$ for an N-length signal, in particular much faster than the standard fast Fourier transform (FFT); in the sublinear methods, not all modes are computed, however. We focus here on the RAℓSFA (randomized algorithm of sparse Fourier analysis) algorithm [16, 32, 17, 33]. Given an error level α, RAℓSFA computes a (near-)optimal B-term Fourier representation R in time and space $\mathrm{poly}(B, |\log\delta|, \log N, \log M, 1/\alpha)$, such that $||S - R||_2^2 \le (1+\alpha)\|S - \mathbf{P}_B(S)\|_2^2$, with success probability at least $1-\delta$, where M is related to the machine precision of the computer, and $\mathbf{P}_B(S)$ is the optimal B-term Fourier representation of S (obtained by retaining only the B frequency modes of S that have the largest amplitudes). The algorithm contains some random elements (which do not depend on the signal); the approach guarantees that the error of estimation is of order $\alpha\|S\|_2^2$ with probability exceeding $1-\delta$. The empirical experiments in [32, 33] present a practical (and improved) implementation of the algorithm, showing that it is of interest; i.e., it outperforms the FFT for reasonably large N. It convincingly beats the highly optimized FFTW implementation [15] of FFT when the number of grid points N is reasonably large. For an eight-mode signal ($B = 8$), the crossover point lies at $N \simeq 70{,}000$ in one dimension and at $N \simeq 900$ for data on an $N \times N$ grid in two dimensions. When $B = 64$, RAℓSFA surpasses the FFTW, in one dimension, at 3×10^7.

Our study will focus on a model multiscale problem in the form of the parabolic PDE

$$\partial_t u - \partial_x a^\varepsilon(t,x)\partial_x u = 0, \qquad u(0,x) = f(x), \tag{1}$$

with periodic boundary conditions, $x \in [0, 2\pi)$ and $a^\varepsilon(t, x + 2\pi) = a^\varepsilon(t,x)$. The coefficient a^ε is bounded and uniformly positive,

$$0 < a_{\min} \le a^\varepsilon(t,x) \le a_{\max} \qquad \forall t, x, \tag{2}$$

where $a_{\min}$ and $a_{\max}$ are the minimum and maximum values of a^ε, respectively. It is also assumed to have a fine scale structure of characteristic length proportional to ε, or, more precisely, $\partial_x^p a^\varepsilon \sim \varepsilon^{-p}$. The typical example will be coefficients of the type $a^\varepsilon = a(x, x/\varepsilon)$ or $a^\varepsilon = a(x/\varepsilon)$, where a is periodic in all arguments and $1/\varepsilon \in \mathbb{N}$. The solutions in this case have a smooth profile on which rapid oscillations are superimposed; their period is proportional to ε. This problem has been widely studied in the context of multiscale problems; indeed, the solution's behavior when $\varepsilon \to 0$ is well understood through homogenization theory [4]. Numerically, the difficulty is related to the smallness of ε; direct methods must resolve the ε-scale to be accurate, and for a fixed tolerance the computational cost is at least $O(\varepsilon^{-d})$ in d dimensions. A number of methods have been proposed and tested for this problem, such as finite element methods with special multiscale basis functions [3, 19, 20, 21, 25], heterogeneous multiscale methods [1, 9, 26], equation-free methods [27], and wavelet-based numerical homogenization [6, 5, 14].

In this paper we use a spectral method based on RAℓSFA to solve (1). The main difference from earlier methods is that a randomized sampling algorithm is used to

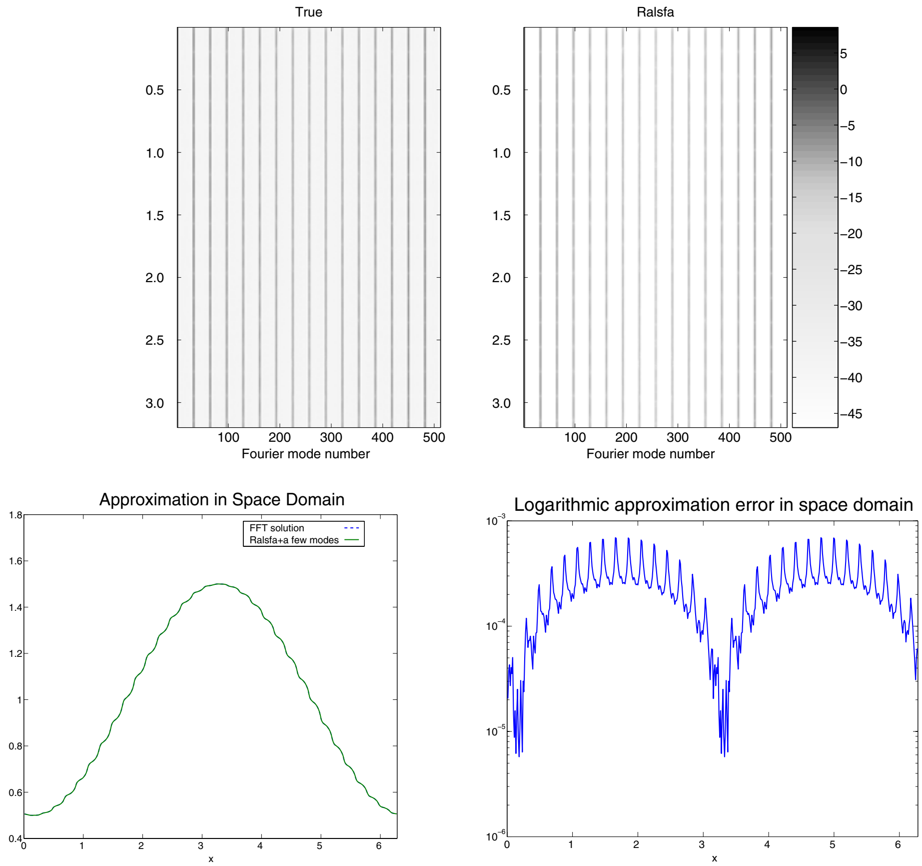

FIG. 1. *Comparison of a numerical solution to* (1) *with the traditional and sparse spectral method: size of Fourier modes for* $t \in [0, 3.2]$ *(top), solution at* $t = 3.2$ *(bottom left), and approximation error at* $t = 3.2$ *(bottom right).*

identify an optimal representation of the full solution; other methods either compute the representation explicitly or keep only a representation of the coarse scales.

The classical spectral method [18, 31] approximates the solution with the N lowest Fourier modes and computes the spatial derivatives by the FFT and the inverse FFT at each time step. This gives a high accuracy compared to conventional difference methods. However, to capture the microstructure of length ε, the smallest length scale in the representation of the solution must be at least of the same order, and we must take $N \sim 1/\varepsilon$ (see Corollary 3.3). It follows that for a fixed tolerance, the computational cost of the spectral method would be $O(\varepsilon^{-d}|\log \varepsilon|)$ per time step in d dimensions. Hence, it is still very expensive to seek the solution of problem (1) when $\varepsilon \ll 1$.

A simple example will explain our motivation to replace FFT in the spectral method by RAℓSFA. We consider (1) with coefficient $a^\varepsilon(x) = [1 + 0.5\sin(\frac{x}{\varepsilon})]^{-1}$, $\varepsilon = 1/32$, and initial data $f(x) = 1 + 0.5\cos(x)$. We first solve it using a traditional spectral method and $N = 512$. The bottom left subfigure in Figure 1 shows a snapshot of this reference solution at $t = 3.2$. The top left subfigure shows the strength of the Fourier modes of the solution as a function of time; the gray scale level indicates their absolute value (in log scale) from $t = 0$ (top) to $t = 3.2$ (bottom). It is clear that not only the diffusion coefficient a^ε and the initial condition f but also the full the solution u are well approximated by a sparse Fourier representation. We then make a second experiment, where these functions are instead represented by only the 17 largest modes; the identity of these modes may change over time. In each time step we approximate the time derivative using RAℓSFA. Details are given in section 2.2. The corresponding results are shown in the top right and bottom left subfigures. At the scale of the plot, the results coincide. The precise difference between the RAℓSFA solution and the FFT solution is shown in the bottom right subfigure in log scale.

For this test we are thus able to get a solution very close to the reference solution with only $B \ll N$ modes if they are chosen as the largest modes of the solution. The numerical results are qualitatively the same when we take $a^\varepsilon = a(x, x/\varepsilon)$, and we also observe that, unlike in the traditional spectral method, the number of modes needed for a certain tolerance does not seem to increase when ε becomes smaller. These numerical results stimulate us to explore the sparse spectral method further. With just $B \ll N$ modes in the representation of the solution, the time complexity of the sparse spectral method is only $\mathrm{poly}(B, |\log \varepsilon|, 1/\alpha, |\log \delta|)$ per time step, or the time required to find a near-optimal B-term Fourier representation of the initial condition f and the coefficient function a^ε. For small ε it thus outperforms the traditional spectral method in cost per time step if B depends only mildly on ε.

In this paper we study periodic problems with distinct scale separation of the type (1). We perform numerical experiments and analysis of a simplified setting. The results show that it is indeed possible to solve these problems at a computational cost that is essentially independent of the small scale parameter for fixed accuracy. We note that the gain compared to standard spectral methods comes not so much from the fact that the solution can be well represented by only a few modes but rather from the fact that there are large gaps in the spectrum between the modes, which is a recurring feature of multiscale problems. These gaps do not cost anything extra in the sparse method. We hope that the results in the paper will carry over to more complex problems where there are large gaps between the significant Fourier modes of the solution, including problems with time-dependent coefficients and nonlinear terms. More elaborate schemes will undoubtedly be needed for these problems though.

This paper is organized as follows. In section 2, we discuss the standard spectral method for (1) and introduce a new sparse spectral method based on the RAℓSFA algorithm. In section 3, we analyze the approach in the simplified setting of fixed projections and show that the computational cost to achieve a given tolerance is much lower for the sparse spectral method. Next, in section 4, we give some numerical results for the error analysis. Finally, we show some ways to extend the methods to elliptic and hyperbolic problems in section 5.

2. Sparse spectral methods. In this section we discuss spectral methods for the parabolic PDE

$$\partial_t u - \partial_x a^\varepsilon(x) \partial_x u = 0, \qquad u(0, x) = f(x). \tag{3}$$

See, e.g., [30]. We will compare a standard spectral method with a new sparse spectral method.

We discretize time uniformly with time step Δt and denote $t_n = n\Delta t$. We let $U^n(x)$ be the approximation of the exact solution at $t = t_n$ so that $U^n(x) \approx u(t_n, x)$. This approximate solution will be band-limited with respect to x, uniformly in t; as a result, we can restrict ourselves to a uniform spatial grid $\{x_j\}$ in x, with $N = J/\varepsilon$

points along each dimension and spacing $\Delta x = \varepsilon/J$ for some constant J. We also denote the 1-dimensional sphere (i.e., $[0, 2\pi]$, periodized) by $\mathbb{S}$.

We frequently consider L^2 functions on $\mathbb{S}$, and for such a function $f(x)$ we generically denote its Fourier coefficients by $\hat{f}_k$:

$$f(x) = \sum_k \hat{f}_k e^{ikx}.$$

The norm $||\cdot||_m$ will denote the H^m Sobolev norm, and without subscript, the norm $\|\cdot\|$ always means the L^2 norm.

2.1. Standard spectral method. There are many versions of spectral schemes for (3). We take a very simple representative. The spatial approximation will be made with the low frequency projection

$$(P_N f)(x) = \sum_{|k|<N/2} \hat{f}_k e^{ikx}, \qquad f(x) = \sum_k \hat{f}_k e^{ikx}. \tag{4}$$

Hence, P_N constructs a Fourier representation by simply taking the N Fourier modes with the lowest frequencies. This gives a very high ("spectral") accuracy in space. We combine the low-frequency projection with a forward Euler discretization in time to get our standard spectral scheme:

$$U^{n+1} = P_N \left[U^n + \Delta t \partial_x \tilde{a}^\varepsilon \partial_x U^n\right], \qquad U^0 = P_N f, \quad \tilde{a}^\varepsilon = P_N a^\varepsilon.$$

The solution can be represented by N Fourier modes. By using the FFT, spatial derivatives, the projection P_N, and the multiplication $\tilde{a}^\varepsilon \partial_x U_n$ can all be computed in $O(N \log N)$ time. As we shall see below in Corollary 3.3, one typically needs to take $N \sim \varepsilon^{-1}$ to maintain fixed accuracy when $\varepsilon \to 0$.

2.2. Sparse spectral method. In this case we replace the projection P_N by a RAℓSFA based projection. Put simply, we will project on the *largest* modes instead of the lowest modes as in the standard spectral method. This implies that the set of significant modes may vary in each time step, which is attractive in the sense that the approximation adapts to the solution. If the functions f, a^ε and the solution u^ε can be well represented by a few judicially chosen modes, the sparse method will have a small local truncation error. Since the scheme is also stable (see section 2.3), we expect a convergent method. We are, however, not able to conclude this rigorously, as the scheme is nonlinear and difficult to analyze, but a simplified analysis is given in section 3.

We introduce the projection operator $\mathbf{P}_B$, which finds the best B-term Fourier representation $R(x)$ for some fixed B. In precise notation a linear projection operator $\bar{\mathbf{P}}_B$ is constructed from an L^2 function f and is subsequently applied to a, possibly different, function g as follows:

$$(\bar{\mathbf{P}}_B(f)g)(x) = R(x) = \sum_{\ell=1}^{B} \hat{g}_{k_\ell} e^{ik_\ell x},$$

where $k_1, \ldots, k_B$ are the B largest modes of f:

$$\{k_1, \ldots, k_B\} = \mathcal{M}_B(f) := \underset{\substack{\Lambda \subset \mathbb{Z} \\ \#\Lambda = B}}{\operatorname{argmax}} \sum_{k \in \Lambda} |\hat{f}_k|^2. \tag{5}$$

The operator $\mathbf{P}_B$ is then defined as $\mathbf{P}_B(f) := \bar{\mathbf{P}}_B(f)f$.

Remark. Note that $\mathbf{P}_B$ is indeed a projection since $\mathcal{M}_B(\bar{\mathbf{P}}_B(f)f) = \mathcal{M}_B(f)$, and therefore $\mathbf{P}_B \circ \mathbf{P}_B(f) = \bar{\mathbf{P}}_B(\bar{\mathbf{P}}_B(f)f)f = \mathbf{P}_B(f)$. The operator $\mathbf{P}_B$ is, however, *not* linear: the number B of modes is fixed, but not the identity of the modes, so that $\mathbf{P}_B(f+g) \neq \mathbf{P}_B(f) + \mathbf{P}_B(g)$ in general. (For a fixed f the operator $\bar{\mathbf{P}}_B(f)$ is clearly linear though.) Moreover, if B is kept fixed, multiplication and addition operations may lead to undesirable growth of errors. For example, let $B = 2$. Suppose $f = \phi_3 + 0.9\phi_7$ and $g = \phi_1 + 0.8\phi_4$. Then $\mathbf{P}_2(f) = f$ and $\mathbf{P}_2(g) = g$. However, $\mathbf{P}_2(f+g) = \phi_1 + \phi_3 \neq \mathbf{P}_2(f) + \mathbf{P}_2(g)$, and the relative error $||f + g - \mathbf{P}_2(f+g)||/\|f+g\| = 0.42$ is quite large.

The sparse scheme then reads

$$U^{n+1} = \mathbf{P}_B\left(P_N\left[U^n + \Delta t \partial_x \tilde{a}^\varepsilon \partial_x U^n\right]\right), \qquad U^0 = \mathbf{P}_B(P_N f), \quad \tilde{a}^\varepsilon = \mathbf{P}_B(P_N a^\varepsilon). \tag{6}$$

In each step the solution is represented by $B \ll N$ modes, whose identity may change over time. We assume that we have N-term approximations of a^ε and f. By applying RAℓSFA to these N-term approximations, we can then obtain (near-optimal) B-term Fourier representations. In other words, RAℓSFA can be used to approximate the $\mathbf{P}_B(P_N\,\cdot\,)$ projection operator. The complexity for this is $O(B \log N)$. In subsequent steps U^n and $\tilde{a}^\varepsilon$ are both supported on a maximum of B modes. Consequently, $U^n + \Delta t \partial_x \tilde{a}^\varepsilon \partial_x U^n$ is supported on a maximum of $B + B^2$ modes. Applying $\mathbf{P}_B$ to such a short signal is easier, and the RAℓSFA algorithm is not necessary: in every time step, we compute the coefficients of the $B + B^2$ modes by simple convolution, we sort them, and we retain only the largest B; this can be done in $O(B^2 \log B)$ time. Hence, the complexity is $O(B \log N)$ for the initial data and $O(B^2 \log B)$ for every subsequent time step.

Alternatively, if the samples of a^ε are given, then these values can be used whenever we need to estimate, e.g., samples of $a^\varepsilon(x)\partial_x U^n(x)$, without explicitly decomposing $a^\varepsilon(x)$ into its most important modes. In this case we can use RAℓSFA in each time step. Since $U^n(x)$ is supported on at most B modes, the sampling cost is $O(B)$, and we get a $O(B^2 \log N)$ cost in each time step. This more expensive strategy may be necessary when a^ε is time-dependent.

Remark. We may get good approximation of a^ε with many fewer modes than needed for the U^n. In this case, we could introduce two numbers, B_a and B_U, giving the number of modes we retain for a^ε and U^n, respectively. Then the first method would require $O(B_a B_U \log B_U)$ cost for all time steps after the first (assuming $B_a \ll B_U$) and the second method $O(B_U^2 \log N)$.

Remark. The sparse scheme may be seen as an adaptive Galerkin method, where the approximation subspace is spanned by a set of Fourier modes that can change in every time step. RAℓSFA provides the adaptation algorithm, driving the method to use subspaces corresponding to the largest Fourier modes of the solution.

One can also improve the projection and use a projection which also takes into account the size of the time derivative of the solution. Let $\bar{\mathbf{Q}}_B$ be defined as

$$\bar{\mathbf{Q}}_B(f)g = \sum_{\ell=1}^{B'} \hat{g}_{k_\ell} e^{ik_\ell x}, \qquad (k_1, \dots, k_{B'}) = \mathcal{M}_B(f) \cup \mathcal{M}_B(\partial_x a^\varepsilon \partial_x f),$$

with $\mathcal{M}_B$ defined in (5). Then we set $\mathbf{Q}_B(f) := \bar{\mathbf{Q}}_B(f)f$. Hence, when v solves (1), then $\mathbf{Q}_B(v)$ projects v on the largest modes of v and v_t. By replacing $\mathbf{P}_B$ with $\mathbf{Q}_B$

in the adaptive scheme, we get the improved sparse spectral scheme

$$U^{n+1} = \mathbf{Q}_B\left(P_N\left[U^n + \Delta t \partial_x \tilde{a}^\varepsilon \partial_x U^n\right]\right), \qquad U^0 = \mathbf{Q}_B(P_N f). \tag{7}$$

The complexity for this scheme is similar to the one above.

2.3. Stability. The numerical stability of the sparse scheme can be studied as follows. For any function $u(x)$ one can easily derive that, as long as $a^\varepsilon > 0$,

$$\langle u, \partial_x a^\varepsilon \partial_x u\rangle = -\langle u_x, a^\varepsilon u_x\rangle = -\left\langle \frac{1}{a^\varepsilon} a^\varepsilon u_x, a^\varepsilon u_x\right\rangle \le -\frac{1}{a_{\max}}||a^\varepsilon u_x||^2$$

and

$$||P_N \partial_x u||^2 = \sum_{|k|<N/2} k^2 \hat{u}_k^2 \le \frac{N^2}{4} \sum_{|k|<N/2} \hat{u}_k^2 = \frac{N^2}{4}||P_N u||^2.$$

All the spectral schemes discussed above can be written in the form

$$U^{n+1} = P^{n+1}W^{n+1}, \qquad W^{n+1} = U^n + \Delta t \partial_x a^\varepsilon \partial_x U^n, \quad P_N P^n = P^n.$$

The standard spectral scheme uses simply $P^n = P_N$, while the sparse scheme has $P^n = \bar{\mathbf{P}}_B(P_N W^n)$ or $P^n = \bar{\mathbf{Q}}_B(P_N W^n)$. Clearly, $P_N P^n = P^n$ implies $P_N U^n = U^n$. Therefore, as long as $\tilde{a}^\varepsilon$ keeps the positivity imposed on the exact a^ε, we will have

$$\begin{aligned}
||U^{n+1}||^2 &= ||P^{n+1}W^{n+1}||^2 \le ||W^{n+1}||^2 \\
&= ||U^n||^2 + 2\Delta t\langle U^n, \partial_x \tilde{a}^\varepsilon \partial_x U^n\rangle + (\Delta t)^2 ||\partial_x \tilde{a}^\varepsilon U_x^n||^2 \\
&\le ||U^n||^2 - 2\Delta t \frac{1}{\tilde{a}_{\max}}||\tilde{a}^\varepsilon \partial_x U^n||^2 + \left(\frac{N\Delta t}{2}\right)^2 ||\tilde{a}^\varepsilon \partial_x U^n||^2 \\
&= ||U^n||^2 - \Delta t||\tilde{a}^\varepsilon \partial_x U^n||^2 \left(\frac{2}{\tilde{a}_{\max}} - \frac{N^2 \Delta t}{4}\right).
\end{aligned}$$

This shows that all the spectral schemes are stable as long as the CFL condition

$$\Delta t \le \frac{8}{N^2 \tilde{a}_{\max}}$$

holds and the approximated coefficient stays positive, i.e., $\tilde{a}^\varepsilon(x) > 0$ for all x.

Remark. As we see this puts a severe constraint on the time step also for the sparse spectral scheme. In fact, the time step must be taken proportional to N^{-2}, which, as we will see below, amounts to $\Delta t \sim \varepsilon^2$ if we shall maintain a fixed accuracy when $\varepsilon \to 0$. In this paper, however, we are not concerned with the trade-off between complexity and accuracy in the time-stepping, just in the spatial approximation.

In principle there are ways to deal with the small scale also in the time-stepping. Implicit schemes will alleviate the stability constraint, but in general we will still need to take $\Delta t \sim \varepsilon$ to maintain accuracy since $u_t^\varepsilon \sim \varepsilon^{-1}$. When a^ε varies slowly in time one can, however, do better. Then it is not necessary to update the projection in every time step, and one can consider schemes of the following type: Initial data is approximated as

$$U^0 = \mathbf{Q}_B(f), \qquad Q^0 = \bar{\mathbf{Q}}_B(f).$$

For $n > 0$, we compute recursively

$$\partial_t v^n - Q^n \partial_x a^\varepsilon(t_n, x) \partial_x v^n = 0, \qquad v^n(0,x) = U^n(x) \tag{8}$$

and

$$U^{n+1} = \mathbf{Q}_B(v^n(\Delta t, \cdot)), \qquad Q^{n+1} = \bar{\mathbf{Q}}_B(v^n(\Delta t, \cdot)).$$

This is accurate if we take $\Delta t \sim 1/|a_t^\varepsilon|$, which is assumed independent of ε. Moreover, since $Q^n U^n = U^n$,

$$\frac{1}{2}\frac{d}{dt}||v^n(t,\cdot)||^2 = \langle v^n, v_t^n\rangle = -\langle Q^n v_x^n, a^\varepsilon Q^n v_x^n\rangle \le 0,$$

and therefore

$$||U^{n+1}||^2 \le ||v^n(\Delta t, \cdot)||^2 = ||v^n(0,\cdot)||^2 = ||U^n||^2.$$

Thus, the scheme is unconditionally stable and we can take Δt as large as we like. As for complexity, (8) can in principle be solved exactly at ε-independent cost. It reduces to a linear ODE for the B modes in Q^n, with a $B \times B$ system matrix. One could also potentially use adaptive implicit ODE methods or projective integration techniques to solve (8) at a cost independent of ε. We will, however, not pursue these possibilities in the present paper.

Remark. The RAℓSFA algorithm gives only an approximation of $\mathbf{P}_B(P_N\,\cdot\,)$ and is reliable only with some fixed probability. This introduces additional errors in the computations, in particular if RAℓSFA is used in every time step. The error level α can be set in the algorithm, and also the success probability $1-\delta$. The latter will cause $O(1)$ errors in the approximation on average $\delta/\Delta t$ times when RAℓSFA is called upon in every time step. We will not analyze these errors in detail, but note that if the scheme we use is stable, their effect should be limited if we take, e.g., $\alpha, \delta \sim \Delta t^2$. With some adaptation of RAℓSFA to the present case, one could probably use larger α, δ.

3. Analysis for fixed projections. Numerical tests indicate that our sparse spectral scheme very quickly settles on a projection which does not change much when a^ε is time-independent. (See, e.g., Figure 1.) As a first step it therefore makes sense to study the case when we have a *fixed* (time-independent) projection but not necessarily just a low-frequency projection. In this section we shall study this case. Since the projection is fixed, we can consider semidiscrete schemes where time is continuous. More precisely, we shall be interested in approximating (1) by spectral schemes of the form

$$\partial_t v - P\partial_x a^\varepsilon(t,x)\partial_x v = 0, \qquad v(0,x) = v_0(x), \quad Pv_0 = v_0, \tag{9}$$

where P is a time-independent projection of $L^2(\mathbb{S})$ to a finite-dimensional subspace, and as before, $0 < a_{\min} \le a^\varepsilon(t,x) \le a_{\max}$. The typical situation below is that P is a projection on a certain set of Fourier modes. If P projects on all modes smaller than a number N, this is the classical (semidiscrete) spectral scheme.

3.1. General error analysis. Let us first state a general theorem on the approximation quality of $v(x)$ in (9) compared to the solution $u(x)$ of (1). The statement is similar to Céa's lemma in finite element analysis [10], in the sense that they both relate the approximation error of the exact solution in the chosen subspace to the error in the numerical solution.

THEOREM 3.1. *Suppose P is a time-independent, linear, and orthogonal projection operator on $L^2(\mathbb{S})$. Let $e = u - v$ be the difference between the spectral and the exact solution u of* (1). *Let $\delta = u - Pu$ be the approximation error of the exact solution. Then*

$$||e(t,\cdot)||^2 + \int_0^t ||e_x(s,\cdot)||^2\, ds \le c||Pf - v_0||^2 + c||\delta(t,\cdot)||^2 + c' \int_0^t ||\delta_x(s,\cdot)||^2\, ds.$$

The constants depend only on $a_{\min}$ and $a_{\max}$, not on derivatives of a^ε.

Proof. Divide e into two parts

$$e = \delta + \eta, \qquad \delta = u - Pu, \quad \eta = Pu - v,$$

and set $\eta_0(x) := \eta(0,x)$. From the equation for u (see (1)) and v (see (9)), which differ only by the presence of P in (9), we get

$$\begin{aligned} \partial_t\eta - P\partial_x a^\varepsilon \partial_x \eta &= \partial_t Pu - \partial_t v - P\partial_x a^\varepsilon \partial_x Pu + P\partial_x a^\varepsilon \partial_x v \\ &= P\partial_t u - P\partial_x a^\varepsilon \partial_x (u - \delta) = P\partial_x a^\varepsilon \partial_x \delta, \end{aligned} \tag{10}$$

where we also used that P and ∂_t commute. Let us define the weighted (time-dependent) norm

$$||u||_a^2 := \int_0^{2\pi} u(x)^2 a^\varepsilon(t,x)\, dx.$$

Since $\eta_0 = Pf - v_0 = P\eta_0$, we will clearly have $P\eta(t,x) = \eta(t,x)$ for all $t \ge 0$. Using integration by parts, we then get

$$\begin{aligned} \frac{d}{dt}\left(||\eta||^2 + \int_0^t ||\eta_x||_a^2\, dt\right) &= 2\langle \eta, \eta_t\rangle + ||\eta_x||_a^2 \\ &= 2\langle \eta, P\partial_x a^\varepsilon \partial_x \eta\rangle + 2\langle \eta, P\partial_x a^\varepsilon \partial_x \delta\rangle + ||\eta_x||_a^2 \\ &= 2\langle \eta, \partial_x a^\varepsilon \partial_x \eta\rangle + 2\langle \eta, \partial_x a^\varepsilon \partial_x \delta\rangle + ||\eta_x||_a^2 \\ &= -\langle \eta_x, a^\varepsilon \eta_x\rangle - 2\langle \eta_x, a^\varepsilon \delta_x\rangle \\ &\le -||\eta_x||_a^2 + 2||\eta_x||_a \cdot ||\delta_x||_a \le ||\delta_x||_a^2. \end{aligned}$$

Consequently,

$$||\eta||^2 + \int_0^t ||\eta_x||_a^2\, ds \le ||\eta_0||^2 + \int_0^t ||\delta_x||_a^2\, ds,$$

and therefore

$$||e||^2 + \int_0^t ||e_x||_a^2\, ds \le ||\eta_0||^2 + ||\delta||^2 + 2\int_0^t ||\delta_x||_a^2\, ds.$$

The final result follows from the equivalence of the $||\cdot||_a$ norm and the usual L^2 norm:

$$\frac{1}{a_{\max}}||u||_a^2 \le ||u||^2 \le \frac{1}{a_{\min}}||u||_a^2. \qquad \square$$

Remark. This theorem tells us that if we can find a projection that is good for the *exact* solution and its derivative, then the corresponding spectral scheme will also give a good approximation of the solution and the derivative. More precisely, if P approximates u well in H^1, then v approximates u well in H^1 for $t > 0$. This means we have a *pointwise* correct solution. We note also that it is not enough for P to approximate u well in just L^2. (An example is to take the homogenization problem with $a^\varepsilon(x) = a(x/\varepsilon)$, where $a(y)$ is 1-periodic. If a projection on the lowest, say $\varepsilon^{-1}/4$, modes is used, the L^2 error of the solution goes to zero as $\varepsilon \to 0$, but the spectral approximation does not converge to the exact solution.)

Of course, the big problem here is how to find that good projection, preferably on a very low-dimensional subspace. This motivates using RAℓSFA for finding the right frequencies.

Remark. We have not assumed that $P\partial_x = \partial_x P$. This means that, so far, the analysis holds also for, e.g., wavelet projections and not just Fourier projections.

3.2. Standard spectral scheme. In this case we have $P = P_N$, the projection on the lowest N modes, and we take $v_0 = P_N f$. We can then use the following result on spectral accuracy: for any $u \in H_m$,

$$||u - P_N u|| \leq \frac{1}{N^m} ||\partial_x^m u||. \tag{11}$$

For the solution u^ε to (1) we thus have

$$||P_N u^\varepsilon - u^\varepsilon|| \leq \frac{1}{N^m} ||\partial_x^m u^\varepsilon||, \qquad ||P_N u_x^\varepsilon - u_x^\varepsilon|| \leq \frac{1}{N^{m-1}} ||\partial_x^m u^\varepsilon||. \tag{12}$$

We next use a theorem estimating u^ε in terms of initial data f.

Let us first define the set $\mathcal{E}$ as all functions $v(t, x, \varepsilon)$ which are infinitely differentiable in t, x and for which there are constants C_{pq} independent of ε such that

$$|\partial_t^p \partial_x^q v|_\infty \leq C_{pq}\, \varepsilon^{-q} \qquad \forall p, q, \varepsilon \geq 0.$$

A typical member of $\mathcal{E}$ would be $v(t, x, \varepsilon) := a(x, x/\varepsilon)$, where $a(x, y) \in C^\infty$.

Remark. The variables x, t, and ε play very different roles; for this reason, we shall, with a slight abuse of notation, write $v^\varepsilon(t, x)$ instead of $v(t, x, \varepsilon)$ for $v \in \mathcal{E}$. Typically, we would consider a sequence ε_m of values for ε, with $\varepsilon_m \underset{m\to\infty}{\longrightarrow} 0$, and be interested in the asymptotic behavior, for $m \to \infty$, of the sequence of functions $v^m(t, x) := v(t, x, \varepsilon_m)$, or, with our new notation, $v^m(t, x) := v^{\varepsilon_m}(t, x)$.

We then have the following theorem.

THEOREM 3.2. *Suppose that $a^\varepsilon(t, x) \in \mathcal{E}$ and that u^ε is the solution to* (1) *with initial data $f \in H^M$. Then for all $1 \leq p \leq M$ and $t > 0$ there are constants $C(p, T)$ independent of ε such that*

$$||\partial_x^p u^\varepsilon|| \leq \frac{C(p, t)}{\varepsilon^{p-1}} ||f||_p, \qquad ||u^\varepsilon|| \leq ||f||. \tag{13}$$

This bound on the solution by a constant times initial data is standard for parabolic equations, but we include a proof in Appendix A to show precisely how the constant in the estimate depends on ε. The proof uses energy estimates along well-known lines.

Together with (12) we get

$$||P_N u^\varepsilon - u^\varepsilon|| \leq \frac{C_1(m, t)}{N^m \varepsilon^{m-1}}, \qquad ||P_N u_x^\varepsilon - u_x^\varepsilon|| \leq \frac{C_2(m, t)}{(\varepsilon N)^m} \tag{14}$$

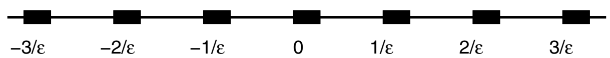

FIG. 2. *The R_b^ε projection. Solid rectangles have width* $2b$ *and correspond to the pass region in frequency space.*

when $f \in H^{m+1}$. Since $\delta = u^\varepsilon - P_N u^\varepsilon$ in Theorem 3.1, we furthermore obtain

$$||e(t,\cdot)||^2 + \int_0^t ||e_x(s,\cdot)||^2\, ds \le \frac{C_1'(m,t)}{N^{2m}\varepsilon^{2m-2}} + \frac{1}{\varepsilon^{2m}N^{2m}}\int_0^t C_2'(m,s)\, ds \le \frac{C(m,t)}{(\varepsilon N)^{2m}}$$

if $v_0 = P_N f$. We have thus proved the following corollary.

COROLLARY 3.3. *Suppose* $f \in H^{m+1}$*. If* $P = P_N$ *and* $v_0 = P_N f$ *in* (9)*, then*

$$||e(t,\cdot)||^2 + \int_0^t ||e_x(s,\cdot)||^2\, ds \le \frac{C(m,t)}{(\varepsilon N)^{2m}}, \qquad e = u - v.$$

Thus for $t > 0$ fixed, the L^2 error is of the order $O(1/(N\varepsilon)^m)$, and for a fixed tolerance, we need to take "a fixed number of modes per wavelength": $N \sim \varepsilon^{-1} = J$. Note that for any $T > 0$ this estimate is equivalent to an $L^2(H^1)$ estimate and thus controls the error pointwise at least for almost all $0 < t < T$.

3.3. Error analysis and complexity for a sparse spectral scheme. It is clear from Figure 1 that in practice the significant modes cluster around multiples of $1/\varepsilon$ in the homogenization problems where a^ε is either of the form $a(x/\varepsilon)$ or $a(x, x/\varepsilon)$. We will here show that it is in fact also enough to track only these modes to still maintain an accurate solution. We thus consider the fixed projection $R_b^\varepsilon P_N$, where

$$(R_b^\varepsilon u)(x) = \sum_{|j|\le b}\sum_{\ell\in\mathbb{Z}} \hat{u}_{\ell n + j} e^{i(\ell n + j)x}, \qquad n = 1/\varepsilon \in \mathbb{N}.$$

The projection is on modes in a b-wide band around multiples of $1/\varepsilon$; see Figure 2. Clearly, R_b^ε commutes with both P_N and ∂_x so that $P_N R_b^\varepsilon = R_b^\varepsilon P_N$ and $\partial_x R_b^\varepsilon = R_b^\varepsilon \partial_x$.

We then study the semidiscrete scheme

$$\partial_t v - R_b^\varepsilon P_N \partial_x a^\varepsilon(x)\partial_x v = 0, \qquad v(0,x) = R_b^\varepsilon P_N f(x). \tag{15}$$

We look at two kinds of coefficients: $a^\varepsilon(x) = a(x/\varepsilon)$ and $a^\varepsilon(x) = a(x, x/\varepsilon)$. As we will see, for these cases, even the *fixed* projection R_b^ε will allow us to remove the ε-dependence for the complexity at fixed accuracy. In terms of accuracy the adaptive sparse scheme should in principle be able to do at least as well.

3.3.1. The case when $a^\varepsilon = a(x/\varepsilon)$. We consider $a^\varepsilon = a(x/\varepsilon)$ in (15), where $\varepsilon = 1/n$ for some $n \in \mathbb{N}$. For this case, we note that R_b^ε also commutes with multiplication by $a(x/\varepsilon)$. In fact, we have the following proposition.

PROPOSITION 3.4. *If* $v \in L^2(\mathbb{S})$ *and* $1/\varepsilon \in \mathbb{N}$*, then* $R_b^\varepsilon v(x/\varepsilon) = v(x/\varepsilon)R_b^\varepsilon$ *on* $L^2(\mathbb{S})$.

Proof. Let $n = 1/\varepsilon$ and suppose $u(x) \in L^2(\mathbb{S})$ has the Fourier coefficients $\hat{u}_\ell$. Then, after writing $\ell = p + nq$, where $-n/2 \le p < n/2$, we get

$$v(x/\varepsilon)u(x) = \sum_{k\in\mathbb{Z}} \hat{v}_k e^{iknx} \sum_{|p|\le n/2}\sum_{q\in\mathbb{Z}} \hat{u}_{p+nq} e^{i(p+nq)x} = \sum_{|p|\le n/2}\sum_{k,q\in\mathbb{Z}} \hat{v}_k \hat{u}_{p+nq} e^{i(p+n(q+k))x}.$$

Therefore,

$$R_b^\varepsilon v(x/\varepsilon)u(x) = \sum_{|p|\le b}\sum_{k,q\in\mathbb{Z}} \hat{v}_k \hat{u}_{p+nq} e^{i(p+n(q+k))x} = \sum_{k\in\mathbb{Z}} \hat{v}_k e^{iknx} \sum_{|p|\le b}\sum_{q\in\mathbb{Z}} \hat{u}_{p+nq} e^{i(p+nq)x},$$

which shows that $R_b^\varepsilon v(x/\varepsilon)u(x) = v(x/\varepsilon)R_b^\varepsilon u(x)$. $\square$

Since (15) implies that $R_b^\varepsilon v = v$, we get from Proposition 3.4

$$0 = \partial_t v - P_N \partial_x a(x/\varepsilon)\partial_x R_b^\varepsilon v = \partial_t v - P_N \partial_x a(x/\varepsilon)\partial_x v, \qquad v(0,x) = R_b^\varepsilon P_N f(x),$$

and it is clear that we are in fact doing the same approximation as for a standard spectral scheme; the only difference is in the approximation of initial data. Supposing $f \in H^{m+1}$ we then have

$$||P_N(I - R_b^\varepsilon)f|| \le ||P_b f - f|| \le \frac{||f||_{m+1}}{b^{m+1}}$$

by (11). Therefore, we get

$$||e(t,\cdot)|| + \int_0^t ||e_x(s,\cdot)||\, ds \le C(t)\left[\frac{1}{b^{m+1}} + \frac{1}{(\varepsilon N)^m}\right]$$

in a way similar to the proof of Corollary 3.3 in the previous section.

We can now compute the complexity of the traditional spectral method and our sparse spectral scheme for a preassigned error tolerance τ. If we take $b \sim (N\varepsilon)^{\frac{m}{m+1}} = J^{\frac{m}{m+1}}$, then the error estimate is proportional to $(\varepsilon N)^{-m} = J^{-m}$ for both methods. To achieve an error with tolerance τ, we must thus have $J = O(\tau^{-\frac{1}{m}})$. In the traditional spectral method, we need

$$O(N\log N) = O\left(\frac{J}{\varepsilon}\log\left(\frac{J}{\varepsilon}\right)\right) = O\left(\varepsilon^{-1}\tau^{-\frac{1}{m}}\log(\varepsilon^{-1}\tau^{-\frac{1}{m}})\right)$$

computations per time step to achieve this tolerance. The sparse scheme uses $B = 2bN\varepsilon = 2bJ = O(J^{\frac{2m+1}{m+1}})$ modes. As derived in section 2.2 the complexity in the first step is

$$O(B\log N) = O\left(J^{\frac{2m+1}{m+1}}\log N\right) = O\left(\tau^{-\frac{2m+1}{m(m+1)}}\log N\right) \le O\left(\tau^{-\frac{2}{m}}\log(\varepsilon^{-1}\tau^{-\frac{1}{m}})\right).$$

For subsequent time steps the complexity is

$$O(B^2\log B) = O\left(J^{\frac{4m+2}{m+1}}\log J\right) \le O\left(\tau^{-\frac{4}{m}}\log\tau^{-1}\right).$$

Thus if we fix τ, we have the complexity $O(\varepsilon^{-1}|\log\varepsilon|)$ per time step for the standard spectral scheme, whereas our sparse scheme requires $O(1)$ computations for all but the first time step to achieve the same result. The cost of the first step is $O(|\log\varepsilon|)$. We can conclude that the cost for the sparse spectral scheme is essentially independent of ε.

3.3.2. The case when $a^\varepsilon = a(x, x/\varepsilon)$. For this case we invoke a well-known expansion from homogenization theory and write the solution to (1) as

$$u^\varepsilon(t,x) = u_0(t,x) + \varepsilon u_1(t,x,x/\varepsilon) + \varepsilon^2 u_2(t,x,x/\varepsilon) + \cdots + \varepsilon^r u_r(t,x,x/\varepsilon) + \varepsilon^{r+1}T(t,x), \tag{16}$$

where $T \in \mathcal{E}$. (See, e.g., Theorem 5.1 and Remark 5.2 in [4].) The functions $u_0(t,x)$, $u_1(t,x,y)$, ..., $u_r(t,x,y)$ can be assumed to be uniformly smooth and are periodic in both the x- and y-arguments. By the general error analysis, it is enough to show that $||Pu^\varepsilon - u^\varepsilon||$ and $||Pu_x^\varepsilon - u_x^\varepsilon||$ are small for our projection $P_N R_b^\varepsilon$. We need the following proposition, which shows that the Fourier coefficients of a smooth multiscale function $w(x) = v(x, nx)$ satisfy $\hat{w}_{p+nq} \approx \hat{v}_{pq}$ to very good accuracy and that therefore $R_b^\varepsilon w$ converges rapidly to w when b increases.

PROPOSITION 3.5. *Suppose $v(x,y)$ is 2π-periodic in both x and y and $v \in H^m(\mathbb{S}^2)$ with $m \geq 2$. Let $w(x) = v(x,nx)$ for some $n \in \mathbb{N}$ and denote the Fourier coefficients of $v(x,y)$ and $w(x)$ by $\hat{v}_{k\ell}$ and $\hat{w}_k$, respectively. Let m_1 and m_2 be two nonnegative integers satisfying*

$$m_1 + m_2 \leq m, \qquad m_1 \geq 2, \quad m_2 \leq m-1.$$

Then

$$|\hat{w}_k - \hat{v}_{pq}| \leq \frac{C}{n^{m_1}(1+|q|)^{m_2}}, \qquad |\hat{w}_k| \leq \frac{C'}{(1+|p|)^{m_1}(1+|q|)^{m_2}}, \quad ||R_b^\varepsilon w - w|| \leq \frac{C''}{b^{m-3/2}}, \tag{17}$$

where

$$\varepsilon = 1/n, \qquad k = p + qn, \quad p, q \in \mathbb{Z}, \quad -n/2 \leq p < n/2.$$

The constants depend only on m and on $||v||_m$.

Proof. Since $w(x)$ is real, $|\hat{w}_k| = |\hat{w}_{-k}|$, and so we need only to consider $k \geq 0$ and in particular only $q \geq 0$. Moreover,

$$v(x,nx) = \sum_k \sum_\ell \hat{v}_{k\ell} e^{i(k+n\ell)x} = \sum_k \sum_\ell \hat{v}_{k-n\ell,\ell} e^{ikx}$$

so that

$$\hat{w}_k = \sum_\ell \hat{v}_{k-n\ell,\ell}.$$

We next define the two sums S_1 and S_2 as

$$|\hat{w}_k - \hat{v}_{pq}| \leq \sum_{\ell \neq q} |\hat{v}_{p-n(\ell-q),\ell}| = \sum_{\substack{\ell \neq q \\ |\ell| < \lambda}} |\hat{v}_{p-n(\ell-q),\ell}| + \sum_{\substack{\ell \neq q \\ |\ell| \geq \lambda}} |\hat{v}_{p-n(\ell-q),\ell}| =: S_1 + S_2,$$

where $\lambda = k/2n$. To estimate these sums, we use the fact that for any $f(x,y) \in H^m(\mathbb{S}^2)$ we can bound its Fourier coefficients $\hat{f}_{k\ell}$ by

$$|\hat{f}_{k\ell}| \leq C \frac{||f||_m}{(1+|k|)^{m_1}(1+|\ell|)^{m_2}}, \qquad m_1 + m_2 \leq m, \tag{18}$$

where C is a constant independent of f, k, ℓ, m_1, m_2. Applied to S_1 we get with $m_1 = m$ and $m_2 = 0$

$$S_1 \leq \sum_{\substack{\ell \neq q \\ |\ell| < \lambda}} \frac{C_m}{(1+|k-n\ell|)^m} \leq C_m \frac{2\lambda}{(1+k/2)^m} \leq \frac{C'_m}{n(1+k)^{m-1}},$$

since here $|k - n\ell| \geq k - n|\ell| \geq k - n\lambda = k/2$. Moreover, $q \geq 1$ since $q = 0$ is incompatible with the restrictions $|\ell| < \lambda = (p/n + q)/2$ and $q \neq \ell$. Therefore, $q - 1/2 \geq (q+1)/4$ and

$$S_1 \leq \frac{C}{n(1+p+nq)^{m-1}} \leq \frac{C}{n(1+n(q-\frac{1}{2}))^{m-1}} \leq \frac{C4^{m-1}}{n^m(1+q)^{m-1}}. \tag{19}$$

To estimate S_2 we use (18) again:

$$\begin{aligned} S_2 &\leq \sum_{\substack{\ell\neq q\\ |\ell|\geq\lambda}} \frac{C}{(1+|k-n\ell|)^{m_1}(1+|\ell|)^{m_2}} \leq \frac{C}{(1+\lambda)^{m_2}} \sum_{\ell\neq q} \frac{1}{(1+|k-n\ell|)^{m_1}} \\ &= \frac{C}{(1+\lambda)^{m_2}} \sum_{\ell\neq 0} \frac{1}{(1+|p-n\ell|)^{m_1}} \leq \frac{4^{m_1}C}{n^{m_1}(1+\lambda)^{m_2}} \sum_{\ell\neq 0} \frac{1}{(|\ell|+1/2)^{m_1}}. \end{aligned}$$

Here we used the fact that when $|\ell| \geq 1$ we have as before $|p - n\ell| \geq n(|\ell| - 1/2) \geq n(|\ell|+1)/4$. Next, noting that the series converges for $m_1 \geq 2$ and that also $1 + \lambda \geq 1/2 + q$, we get

$$S_2 \leq \frac{C}{n^{m_1}(1+\lambda)^{m_2}} \leq \frac{C'}{n^{m_1}(1+q)^{m_2}}. \tag{20}$$

By combining (19) and (20), we get the first inequality in (17) since $m_1 \leq m$ and $m_2 \leq m - 1$. Then after applying (18) to $v(x, y)$ we have

$$|\hat{w}_k| \leq |\hat{v}_{pq}| + |\hat{w}_k - \hat{v}_{pq}| \leq \frac{||v||_m}{(1+|p|)^{m_1}(1+|q|)^{m_2}} + \frac{C}{n^{m_1}(1+q)^{m_2}},$$

and the second inequality in (17) follows upon noting that $n \geq 1 + |p|$.

Finally, we use this last estimate with $m_1 = m - 1$ and $m_2 = 1$ to get

$$\begin{aligned} ||R_b^\varepsilon w - w||^2 &= \sum_{|p|=b+1}^{n/2} \sum_{q\in\mathbb{Z}} |\hat{w}_{p+nq}|^2 \leq \sum_{p=b+1}^{n/2} \sum_{q\in\mathbb{Z}} \frac{2C}{(1+|p|)^{2m-2}(1+|q|)^2} \\ &= \sum_{p=b+1}^{n/2} \frac{C'}{(1+p)^{2m-2}} \leq C'_m \int_b^\infty \frac{dx}{(1+x)^{2m-2}} = \frac{C''_m(2m-3)}{(1+b)^{2m-3}}. \end{aligned}$$

This shows the remaining inequality in (17). ◻

Let us now show that the projections are accurate when applied to (16).

Corollary 3.6. *Suppose u^ε has the expansion* (16) *up to r terms and that $1/\varepsilon$ is an integer. If the functions $u^\varepsilon, u_0, u_1, \ldots, u_r \in H^m$ with $m \geq 3$, then*

$$||P_N R_b^\varepsilon u^\varepsilon - u^\varepsilon|| \leq C(m)\left[\frac{1}{b^{m-3/2}} + \frac{\varepsilon}{(N\varepsilon)^m} + \varepsilon^{r+1}\right], \tag{21}$$

$$||P_N R_b^\varepsilon u_x^\varepsilon - u_x^\varepsilon|| \leq C'(m)\left[\frac{1}{b^{m-5/2}} + \frac{1}{(N\varepsilon)^m} + \varepsilon^{r}\right]. \tag{22}$$

Proof. We have

$$||P_N R_b^\varepsilon u^\varepsilon - u^\varepsilon|| \leq ||P_N u^\varepsilon - u^\varepsilon|| + ||P_N(R_b^\varepsilon u^\varepsilon - u^\varepsilon)|| \leq C(t)\frac{\varepsilon}{(N\varepsilon)^m} + ||R_b^\varepsilon u^\varepsilon - u^\varepsilon||$$

by (14). For the R_b^ε term we first note that since $u_j \in H^m$ we get from (11) and Proposition 3.5

$$||R_b^\varepsilon u_0 - u_0|| \le ||P_b u_0 - u_0|| \le \frac{||u_0||_m}{b^m}, \qquad ||R_b^\varepsilon u_j - u_j|| \le \frac{C_m}{b^{m-3/2}}, \quad j > 0.$$

Hence, by (16)

$$||R_b^\varepsilon u^\varepsilon - u^\varepsilon|| \le \sum_{j=0}^{r} \varepsilon^j ||R_b^\varepsilon u_j - u_j|| + O(\varepsilon^{r+1}) \le C\left(\frac{1}{b^{m-3/2}} + \varepsilon^{r+1}\right).$$

This shows (21).

Next, we differentiate (16) to get

$$\partial_x u^\varepsilon(t,x) = w_0(t,x) + \varepsilon w_1(t,x,x/\varepsilon) + \cdots + \varepsilon^{r-1} w_r(t,x,x/\varepsilon) + O(\varepsilon^r),$$

where $w_j := \partial_x u_j + \partial_y u_{j+1}$. As before, this gives us

$$||P_N R_b^\varepsilon u_x^\varepsilon - u_x^\varepsilon|| \le ||P_N u_x^\varepsilon - u_x^\varepsilon|| + ||R_b^\varepsilon u_x^\varepsilon - u_x^\varepsilon|| \le \frac{C}{(N\varepsilon)^m} + \sum_{j=0}^{r-1} \varepsilon^j ||R_b^\varepsilon w_j - w_j|| + O(\varepsilon^r).$$

But since $w_j(x) \in H^{m-1}$, we have

$$||R_b^\varepsilon w_j - w_j|| \le \frac{C}{b^{m-5/2}},$$

showing (22). $\square$

We conclude from (21), (22), and Theorem 3.1 that, under the assumptions of Corollary 3.6,

$$||e(t,\cdot)|| + \int_0^t ||e_x(s,\cdot)||\, ds \le C(t)\left[\frac{1}{b^{m-5/2}} + \frac{1}{(N\varepsilon)^m} + \varepsilon^r\right]. \tag{23}$$

For $b = (N\varepsilon)^{m/(m-2.5)} = J^{m/(m-2.5)}$, the error bound is thus proportional to $J^{-m} + \varepsilon^r$. We assume that we are in the asymptotic regime where ε^r is much smaller than the tolerance τ. Then, if we write $\tau = (\alpha+1)\varepsilon^r$ with $\alpha \gg 1$, this implies that we need J^{-m} to be proportional to $\alpha\varepsilon^r = \frac{\alpha}{\alpha+1}\tau$ or J to be of order $O(\tau^{-1/m})$ as in section 3.3.1. The standard spectral scheme still needs $O(N \log N) = O(\frac{J}{\varepsilon}\log\frac{J}{\varepsilon}) = O(\varepsilon^{-1}\tau^{-1/m}\log(\varepsilon^{-1}\tau^{-1/m}))$ operations. Here $B = 2bJ = O(J^{(2m-2.5)/(m-2.5)})$ and the complexity is

$$O\left(B^2 \log N\right) = O\left(J^{\frac{4m-5}{m-2.5}} \log N\right) \le O\left(\tau^{-\frac{4}{m-2.5}} \log(\varepsilon^{-1}\tau^{-1/m})\right)$$

for the sparse spectral method. (The logarithmic contribution can again be omitted after the first time step.) We note thus that we have a similar complexity for this case as when $a^\varepsilon = a(x/\varepsilon)$. In particular the sparse scheme with fixed tolerance has a cost of $O(|\log \varepsilon|)$ in the first time step and $O(1)$ in later time steps.

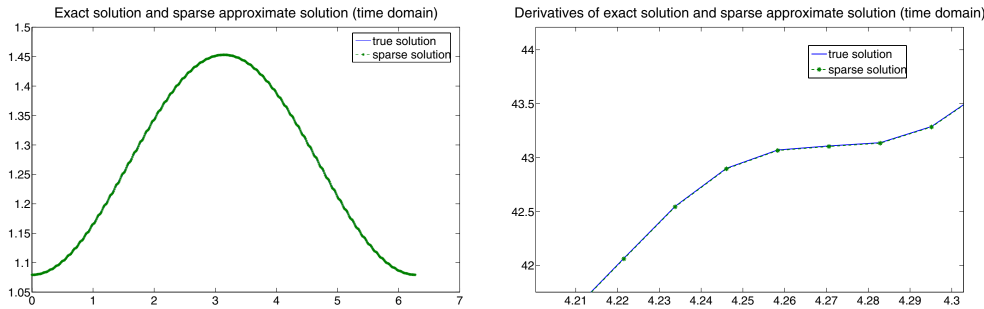

FIG. 3. *Solution to Testcase* 1 *in the time domain. Left: approximate and true solution; they coincide, and it is hard to distinguish them. Right: the magnified comparison of the derivative of the approximate solution and the true solution. Since the two solutions are again very close, we have to zoom in to see their difference.*

4. Numerical experiments. In this section, we investigate the performance of the sparse spectral method in a few numerical experiments. In [32, 33], the advantage in speed of RAℓSFA over the FFTW, for processing sparse signals of large size, has already been extensively documented; we shall not repeat this here. We concentrate on the accuracy issues of the approximation solution obtained by the sparse spectral method.

We compute approximate solutions to

$$u_t = \partial_x a^\varepsilon(x)\partial_x u, \qquad u(0,x) = f(x), \quad x \in [0, 2\pi), \tag{24}$$

for various $a^\varepsilon(x)$. We compare the numerical solutions with a solution obtained from the standard spectral method with a high resolution (N large) applied to (24).

4.1. Testcase 1: $a^\varepsilon = a(x/\varepsilon)$. We begin with experiments for solving the PDE with coefficient function a^ε that is dependent only on x/ε. We use

$$a^\varepsilon(x) = \frac{3}{5 + 3\sin(\frac{x}{\varepsilon})}, \qquad f(x) = \exp(-\cos(x))$$

as coefficient function and initial data. The solution is computed using (6) with $\tilde{a}^\varepsilon$ and U^0 approximated by RAℓSFA. For subsequent steps, direct computation of B^2 modes plus sorting is used to evaluate the $\mathbf{P}_B(P_N\,\cdot\,)$ projection, as discussed in section 2.2. For the initial RAℓSFA step, we set the failure probability (remember that RAℓSFA is a randomized algorithm!) at $\delta = 0.05$, and the accuracy required for the truncated approximation is set at $\alpha = 10^{-8}$. The problem is solved with $\varepsilon = 1/64$, $N = 512$, and $B = 15$ modes. Figure 3 provides a comparison at time $t = 3$ of the sparse spectral method with the exact solution in the time domain. The solutions and its derivatives are very close, and their difference can be distinguished only by magnifying the graph. In Figure 4 a corresponding plot for the frequency domain is given. It shows that the sparse spectral solution accurately captures the largest 15 Fourier modes of the true solution.

4.2. Testcase 2: $a^\varepsilon = a(x, x/\varepsilon)$. In this case we use a more complicated coefficient function that also depends on x, namely

$$a\left(x, \frac{x}{\varepsilon}\right) = \frac{1}{10}\exp\left(\frac{0.6 + 0.2\cos(x)}{1 + 0.7\sin\left(\frac{x}{\varepsilon}\right)}\right), \qquad f(x) = \exp(-\cos(x)). \tag{25}$$

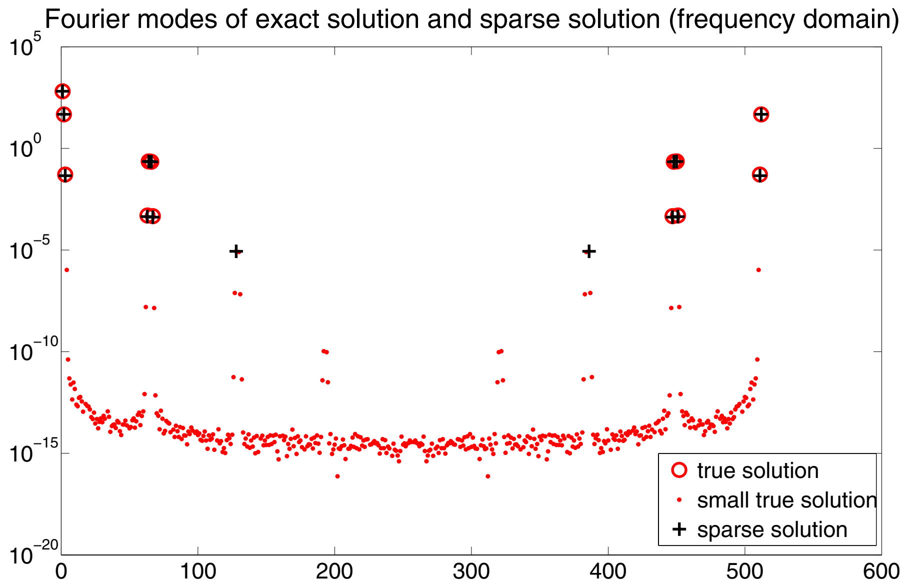

FIG. 4. *Solution to Testcase* 1 *in the frequency domain. The Fourier modes of the approximate solution are the largest* 15 *among all the* $N = 512$ *modes; their amplitudes for the* 15 *largest modes are almost the same as those of the traditional spectral solution.*

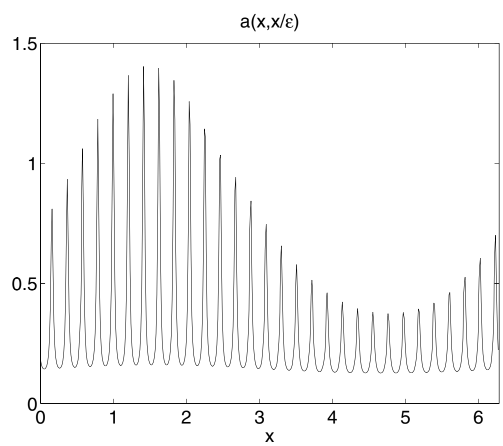

FIG. 5. *The coefficient* $a^\varepsilon = a(x, x/\varepsilon)$.

The coefficient $a(x, x/\varepsilon)$ is plotted in Figure 5. We use the same method as in Testcase 1 to compute the solution, again with $\varepsilon = 1/64$ and $N = 512$, but this time we need slightly more modes, $B = 30$, to accurately capture the solution. Comparisons of the computed and true solutions at $t = 3$ in the time and frequency domain are given in Figures 6 and 7. The same general conclusions as in Testcase 1 hold. A good pointwise approximation of the true solution is obtained.

4.3. Testcase 3: $a^\varepsilon = a(x, x/\varepsilon)$. In this testcase we use the same coefficient and initial data as in Testcase 2; see (25). We use, however, the improved sparse

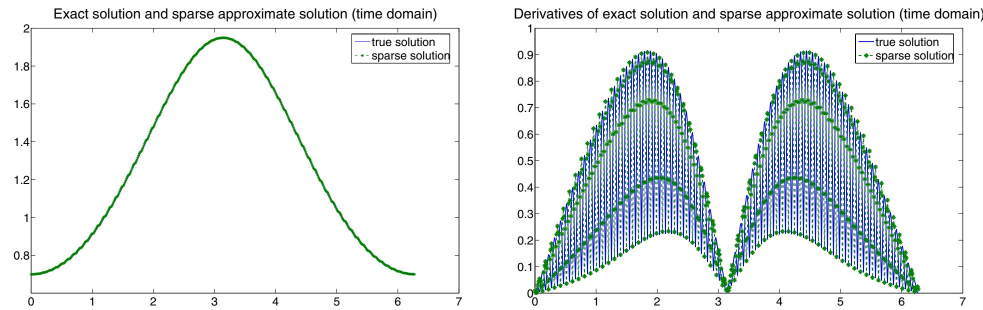

FIG. 6. *Solution to Testcase* 2 *in the time domain. Left: approximate and true solution. Right: absolute value of the derivatives of the true and the approximate solution.*

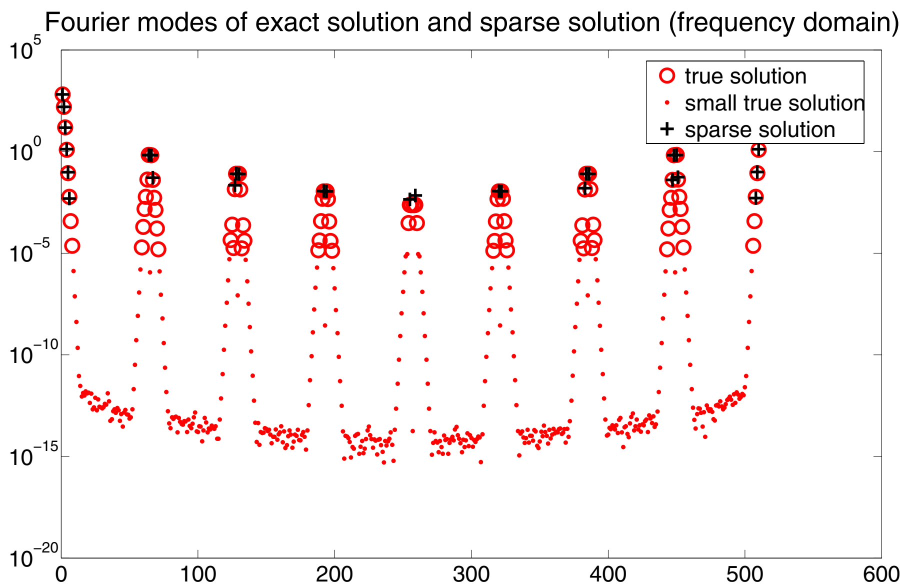

FIG. 7. *Solution to Testcase* 2 *in the frequency domain. The* 30 *largest Fourier modes of the true solution agree well with the modes of the approximate solution. Since $a^\varepsilon(x)$ is more complicated, the number of significant modes in this example is larger than in Testcase* 1.

scheme in (7) rather than the one in (6).

We do not approximate a^ε, i.e., $\tilde{a}^\varepsilon = a^\varepsilon$, but we assume samples of a^ε are available so that samples of $a^\varepsilon \partial_x U^n(x)$ can be fed directly into the algorithm for computing $\mathbf{Q}_B(P_N\,\cdot)$. For simplicity, we do not use RAℓSFA to approximate $\mathbf{Q}_B(P_N\,\cdot)$ but instead compute it exactly by simply using standard FFT plus sorting. This way we avoid the extra approximation errors introduced by RAℓSFA; the code is, of course, asymptotically slower, but we are mainly interested in studying the accuracy, not the speed.

The solution is computed with $B = 11$ and $N = 512$. Since $\mathbf{Q}_B$ in general projects on $2B$ modes, we compare with a solution using the lowest 22 modes, i.e., the standard spectral scheme with $N = 22$. In Figure 8 the sparse solution at $t = 5$ is plotted when

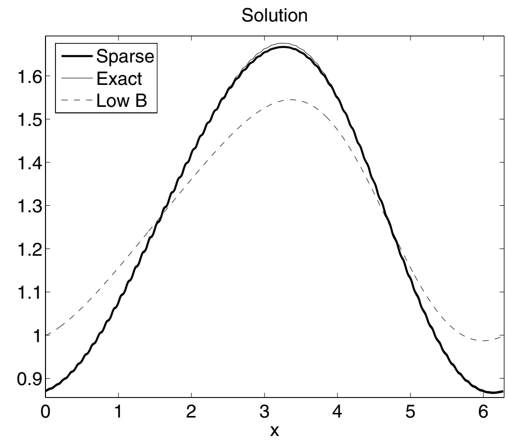

FIG. 8. *Here we used 22 modes for the sparse scheme and the low mode scheme.*

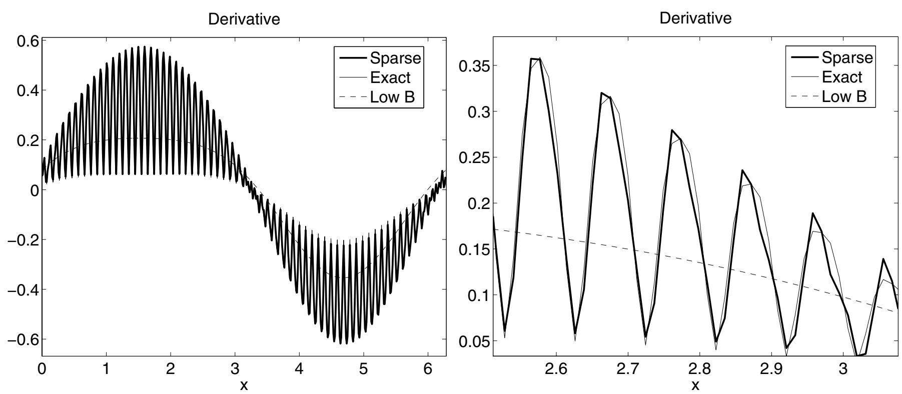

FIG. 9. *The derivative converges pointwise. The right figure is a zoom of the left figure.*

we used $\varepsilon = 1/64$. It agrees very well with the exact solution, while the corresponding solution with the 22 lowest modes is completely wrong. The corresponding results for the derivative of the solution are given in Figure 9. Convergence diagrams in B for the solution and its derivative are shown in Figure 10. The error and convergence rate is essentially independent of ε. Since we have set up the problem such that the $O(1/b^{m-2.5})$ term dominates in (23), this result is as predicted for the fixed projection scheme.

4.4. Testcase 4: $\boldsymbol{a^\varepsilon = a(t, x, x/\varepsilon)}$. Here we use a modified version of the improved scheme (7) to allow for time-dependent coefficients. It reads

$$U^{n+1} = \mathbf{Q}_B\left(P_N\left[U^n + \Delta t \partial_x a_n^\varepsilon \partial_x U^n\right]\right), \qquad U^0 = \mathbf{Q}_B(P_N f), \quad a_n^\varepsilon = a^\varepsilon(t_n, \cdot).$$

We compute $\mathbf{Q}_B$ exactly as in Testcase 3. We use $B = 11$ (corresponding to roughly 22 modes) and $N = 512$. The coefficient is

$$a^\varepsilon(t, x) = \alpha(t)\frac{2 + 1.6\sin(\omega(t)x)}{3 + \cos(x)} + (1 - \alpha(t))\left[1.3 + 0.5\sin\left(\frac{\omega(t)x}{4}\right)\right],$$

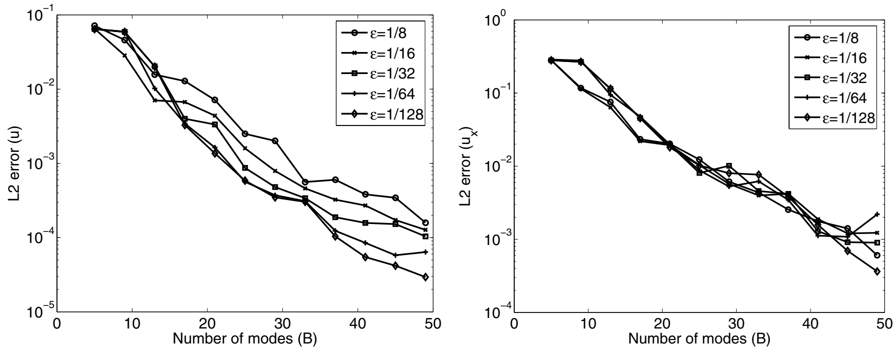

FIG. 10. *Convergence in B for the solution and its derivative.*

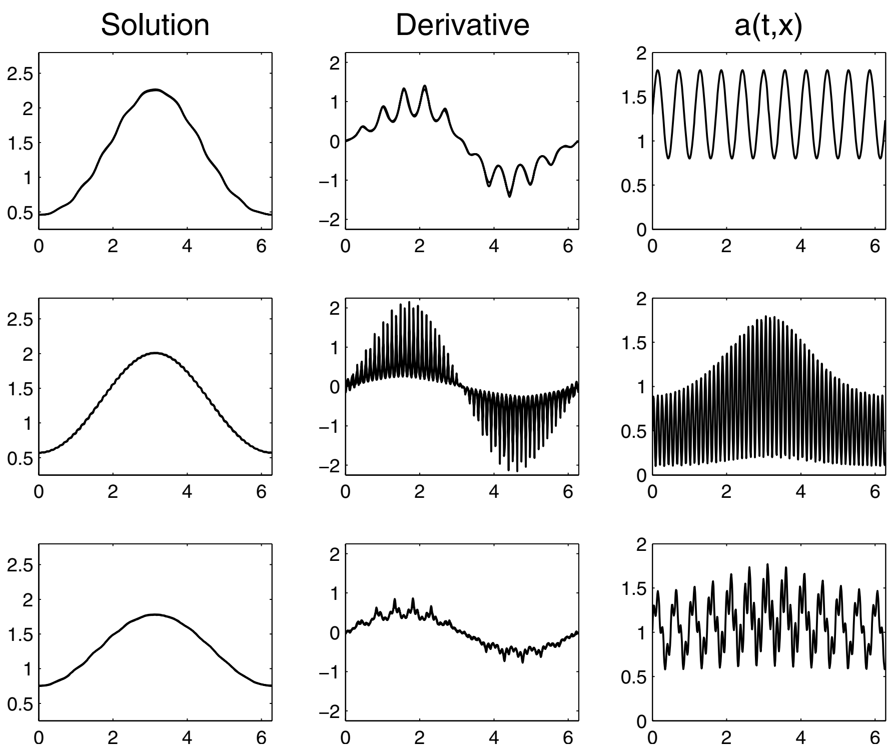

FIG. 11. *Solution u^ε, derivative $\partial_x u^\varepsilon$, and coefficient a^ε at times $t = 0.25$, $t = 0.5$, and $t = 0.85$ (from top to bottom). There is no discernible difference between the sparse numerical solution and the exact solution.*

where

$$\alpha(t) = \cos^2(2\pi t), \qquad \omega(t) = \frac{1}{\varepsilon} - 10(1 - \sin(6\pi t)).$$

The two dominant frequencies, $\omega(t)$ and $\omega(t)/4$, hence change slowly, with the smaller of them appearing and disappearing as time progresses. In the computations we use $\varepsilon = 1/64$. The sparse and exact solutions, their derivatives, and a^ε are plotted in Figure 11 at three different times. The sparse solution cannot be distinguished from

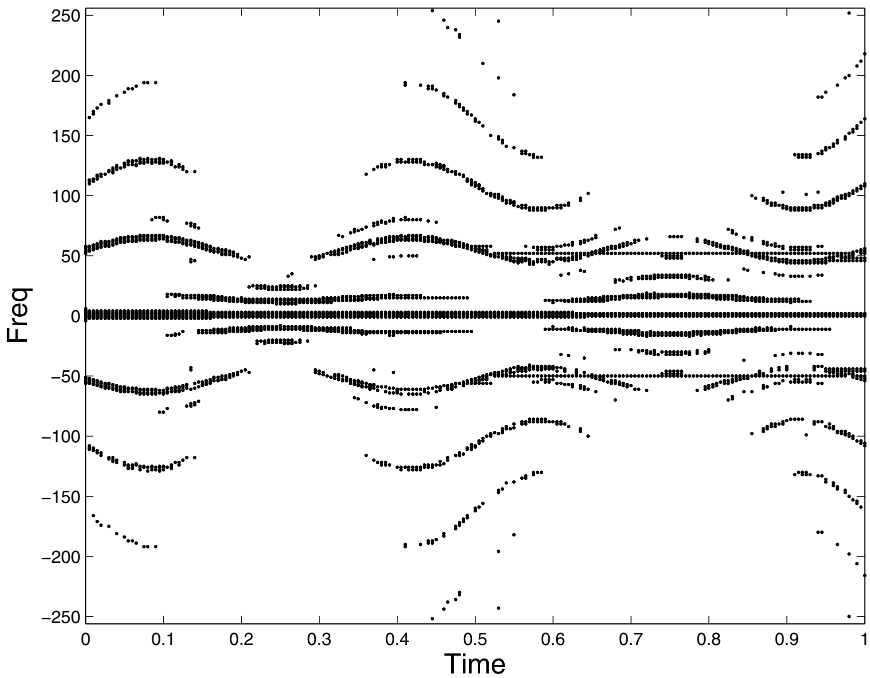

FIG. 12. *Frequencies picked out by the sparse spectral scheme for the case with time-dependent coefficients. Significant modes are indicated by dots.*

the exact solution in the plots. In Figure 12 we show which modes that the sparse scheme picks out. Modes appear and disappear roughly according to the changes of $\omega(t)$. They can appear far from modes in previous time steps, which would be difficult for a standard adaptive scheme.

5. Extensions to elliptic and hyperbolic problems.

We consider here briefly some possible ways of extending the methods described above also to elliptic and hyperbolic problems. We also show some numerical results for these extensions.

5.1. Elliptic problems.

In the elliptic case we would like to solve the model problem

$$-\partial_x a^\varepsilon(x)\partial_x u = f. \tag{26}$$

We assume periodic boundary conditions and therefore also impose that the mean value of f is zero, in order to have a well-posed problem. To solve this we propose to first apply the improved sparse parabolic scheme to

$$\partial_t u - \partial_x a^\varepsilon(x)\partial_x u = f, \qquad u(0,x) = f. \tag{27}$$

We make just a few steps, $n = 1, \ldots, M$,

$$U^{n+1} = \mathbf{Q}_B(P_N[U^n + \Delta t\partial_x a^\varepsilon \partial_x U^n + \Delta t f]), \qquad U^0 = \mathbf{Q}(P_N f),$$

with the same computational strategy as in Testcases 3 and 4 above. Then we define the projection $Q^M = \bar{\mathbf{Q}}_B(U^M)$ and use it to solve the elliptic problem

$$-Q^M \partial_x a^\varepsilon(x)\partial_x Q^M v = Q^M f.$$

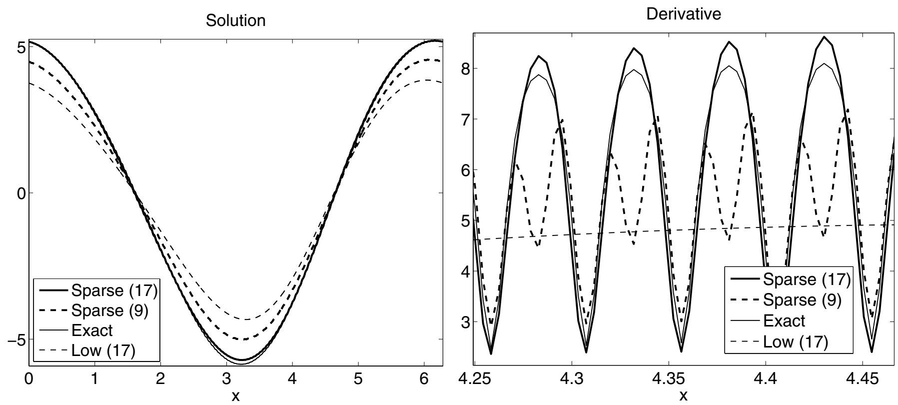

FIG. 13. *Solution for the elliptic case with* 9 *and* 17 *modes (left) and zoom of the solution's derivative (right).*

This is thus a Galerkin approximation of (26) with a particular approximation subspace chosen to correspond to the largest modes in the solution to (27). Since the projection quickly settles on a fixed set of modes for the parabolic case, when a^ε is time-independent, we thus assume the projection obtained in this way agrees well with the corresponding projection after a long time, i.e., the steady, elliptic case. Once the Q^M projection is found, the problem reduces to a linear system of equations with at most $2B$ unknowns that can be solved at an ε-independent cost. We note that for the elliptic case, Céa's lemma,

$$||u - v||_1 \le c||Q^M u - u||_1$$

is a direct analogue of Theorem 3.1 in the parabolic case, and the analysis for fixed projections would be similar.

We show numerical results when a^ε is as given in (25) and $f(x) = \exp(-\cos(x)) - c$, where c is chosen so that f has zero mean. In Figure 13 we show results when $\varepsilon = 1/128$ for $B = 5, 9$ and $N = 1536$. This corresponds to 9 and 17 modes, respectively. We took $M = B$ steps in the parabolic scheme. When using only 17 modes the solution is very close to the exact solution, while the solution when Q^M projects on the lowest 17 modes is very bad. Convergence in B for the solution and its derivative is shown in Figure 14. As in the parabolic case the error and convergence rate are essentially independent of ε.

5.2. Hyperbolic problems. Here we consider the simple hyperbolic model equation

$$u_t + a^\varepsilon(x)u_x = 0, \qquad u(0,x) = f(x),$$

with periodic boundary conditions. This problem is more sensitive than the parabolic and elliptic problems. We can, however, still solve the problem with a sparse method, but we need to use a different time stepping strategy, where the projection is changed more seldom to avoid inducing instabilities. We define a new projection with the same philosophy as we defined $\mathbf{Q}_B$: that it should project on the largest modes of both u and u_t. In this case we get

$$\bar{\mathbf{Q}}^a_B(f)g = \sum_{\ell=1}^{B'} \hat{g}_{k_\ell} e^{ik_\ell x}, \qquad (k_1, \dots, k_{B'}) = \mathcal{M}_B(f) \cup \mathcal{M}_B(a^\varepsilon \partial_x f),$$

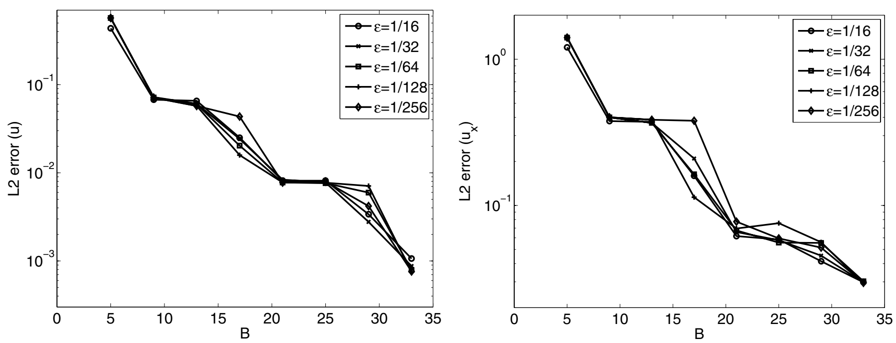

FIG. 14. *Convergence in B for the solution and its derivative.*

with $\mathcal{M}_B$ defined in (5). We then construct the following leapfrog scheme:

$$U^{n+1} = Q^n(P_N[U^{n-1} - 2\Delta t a^\varepsilon \partial_x U^n]), \qquad U^0(x) = Q^0 f,$$

where the projection Q^n is updated every Mth time step:

$$Q^n = \begin{cases} \bar{\mathbf{Q}}^a_B(U^n), & n = kM, \ k \in \mathbb{Z}, \\ Q^{n-1} & \text{otherwise.} \end{cases}$$

In the numerical computations we use a^ε and f as given in (25). Figure 15 shows results at $t = 5$ when $\varepsilon = 1/128$, $B = 35$ (circa 50–60 modes[1]), $\Delta t = 0.0002$, $M = 500$, and $N = 2048$. The sparse solution agrees well with the exact solution. In contrast, a solution computed with the standard spectral scheme using a comparable number of modes (the lowest 70 modes) has the wrong speed of propagation and is far from correct. Convergence in B for the solution and its derivative is shown in Figure 16. The error and rate of convergence for the solution itself are still practically independent of ε. Unlike in the parabolic and elliptic cases, however, there is no convergence for the derivative. This is another manifestation of the more sensitive nature of the hyperbolic case.

6. Conclusions. We provide a new sparse spectral method. Its speed is significantly faster than the traditional spectral method in solving some multiscale PDE problems, while retaining good accuracy.

Appendix A. Proofs.

A.1. Utility results. In the proofs we will denote the binomial coefficients by c_{jk}:

$$c_{jk} := \binom{j}{k} = \frac{j!}{(j-k)!k!}.$$

To prove the theorem we first need a couple of lemmas, starting with one about the set $\mathcal{E}$ defined in section 3.2.

LEMMA A.1. *The set $\mathcal{E}$ is closed under addition and multiplication,*

$$u^\varepsilon, v^\varepsilon \in \mathcal{E} \quad \Rightarrow \quad u^\varepsilon + v^\varepsilon, u^\varepsilon v^\varepsilon \in \mathcal{E},$$

[1]There is more overlap between the largest modes of u and u_t in the hyperbolic case than in the parabolic case.

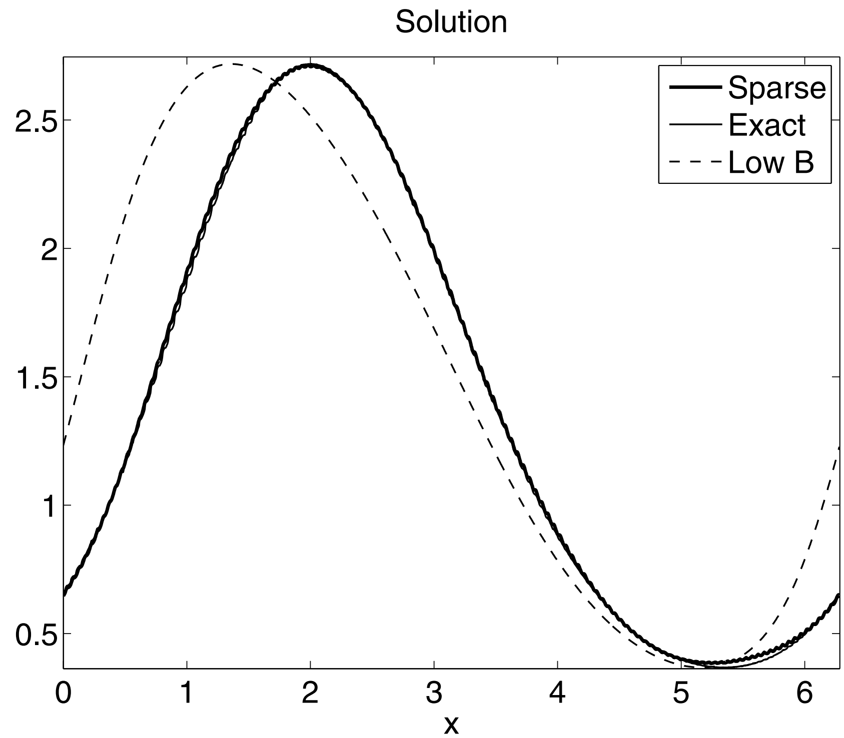

FIG. 15. *Solution for the hyperbolic case with $B = 35$ or circa 50–60 modes.*

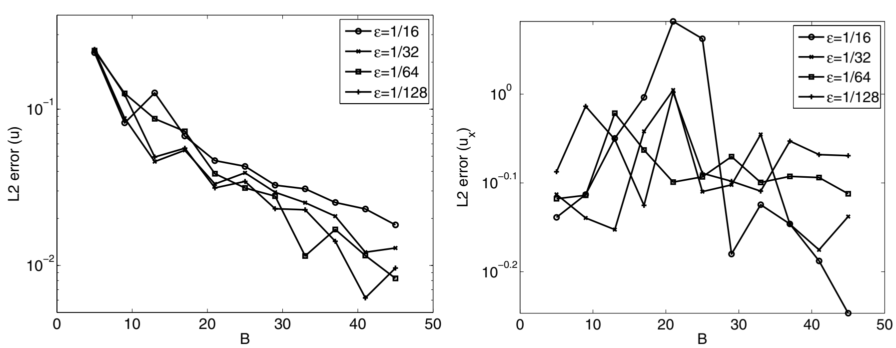

FIG. 16. *Convergence in B for the solution and its derivative.*

and invariant under the operation of $\varepsilon\partial_x$:

$$u^\varepsilon \in \mathcal{E} \quad \Rightarrow \quad \varepsilon\partial_x u^\varepsilon \in \mathcal{E}.$$

Proof. The addition part is obvious. If $u^\varepsilon, v^\varepsilon \in \mathcal{E}$, then the product satisfies

$$|\partial_t^p \partial_x^q u^\varepsilon v^\varepsilon|_\infty = \left| \sum_{j=0}^{p} \sum_{k=0}^{q} c_{pj} c_{qk} (\partial_t^j \partial_x^k u^\varepsilon)(\partial_t^{p-j} \partial_x^{q-k} v^\varepsilon) \right|_\infty \leq \sum_{j=0}^{p} \sum_{k=0}^{q} d_{jk} \varepsilon^{-k} \varepsilon^{k-q} \leq C\varepsilon^{-q},$$

where d_{jk} are some constants. Moreover, if $u^\varepsilon \in \mathcal{E}$, then

$$|\partial_t^p \partial_x^q \partial_x u^\varepsilon|_\infty = \left|\partial_t^p \partial_x^{q+1} u^\varepsilon\right|_\infty \leq C_{pq}\, \varepsilon^{-q-1}, \qquad q = 0, \dots, m-1.$$

Hence, $\varepsilon \partial_x u^\varepsilon \in \mathcal{E}$. ∎

LEMMA A.2. *Suppose that $a^\varepsilon(t,x) \in \mathcal{E}$ and that $u^\varepsilon \in H^{2p+q}$ is a solution to* (1) *with $p \geq 1$. Then*

$$\partial_t^p \partial_x^q u^\varepsilon = \sum_{j=1}^{2p+q} \varepsilon^{j-2p-q} r_{j,p,q}^\varepsilon \partial_x^j u^\varepsilon, \qquad r_{j,p,q}^\varepsilon \in \mathcal{E}, \quad r_{2p+q,p,q}^\varepsilon = (a^\varepsilon)^p. \tag{28}$$

Proof. We show this by induction. For $p = 1$ and $q = 0$ we have $u_t^\varepsilon = a_x^\varepsilon u_x^\varepsilon + a^\varepsilon u_{xx}^\varepsilon$, and by Lemma A.1

$$r_{1,1,0}^\varepsilon = \varepsilon a_x^\varepsilon \in \mathcal{E}, \qquad r_{2,1,0}^\varepsilon = a^\varepsilon \in \mathcal{E}.$$

Suppose the claim holds up to $p = n$ when $q = 0$. After temporarily dropping the last two indices for legibility ($r_{j,n,0}^\varepsilon \to r_j^\varepsilon$), we get

$$\begin{aligned}
\partial_t^{n+1} u^\varepsilon &= \sum_{j=1}^{2n} \varepsilon^{j-2n} \partial_t r_j^\varepsilon \partial_x^j u^\varepsilon = \sum_{j=1}^{2n} \varepsilon^{j-2n} \left[(\partial_t r_j^\varepsilon)\partial_x^j u^\varepsilon + r_j^\varepsilon \partial_x^{j+1} a^\varepsilon \partial_x u^\varepsilon\right] \\
&= \sum_{j=1}^{2n} \varepsilon^{j-2n} \left[(\partial_t r_j^\varepsilon)\partial_x^j u^\varepsilon + r_j^\varepsilon \sum_{k=0}^{j+1} c_{j+1,k} (\partial_x^{j+1-k} a^\varepsilon) \partial_x^{k+1} u^\varepsilon\right] \\
&= \sum_{j=1}^{2n} \varepsilon^{j-2n} (\partial_t r_j^\varepsilon)\partial_x^j u^\varepsilon + \sum_{k=0}^{2n+1} \sum_{j=\max(1,k-1)}^{2n} c_{j+1,k} \varepsilon^{j-2n} r_j^\varepsilon (\partial_x^{j+1-k} a^\varepsilon) \partial_x^{k+1} u^\varepsilon \\
&= \sum_{j=1}^{2n} \varepsilon^{j-2n} (\partial_t r_j^\varepsilon)\partial_x^j u^\varepsilon + \sum_{j=1}^{2n+2} \sum_{k=\max(1,j-2)}^{2n} c_{k+1,j-1} \varepsilon^{k-2n} r_k^\varepsilon (\partial_x^{k+2-j} a^\varepsilon) \partial_x^j u^\varepsilon.
\end{aligned}$$

Thus,

$$\partial_t^{n+1} u = \sum_{j=1}^{2n+2} \varepsilon^{j-2n-2} r_{j,n+1,0}^\varepsilon \partial_x^j u,$$

where

$$r_{j,n+1,0}^\varepsilon = \varepsilon^2 (\partial_t r_{j,n,0}^\varepsilon) + \sum_{k=\max(1,j-2)}^{2n} c_{k+1,j-1} r_{k,n,0}^\varepsilon (\varepsilon \partial_x)^{k+2-j} a^\varepsilon$$

with the convention that $r_{j,p,0}^\varepsilon \equiv 0$ for $j < 1$ and $j > 2p$. It follows from Lemma A.1 that $r_{j,n+1,0}^\varepsilon \in \mathcal{E}$. Moreover, $r_{2n+2,n+1,0}^\varepsilon = c_{2n+1,2n+1} r_{2n,n,0}^\varepsilon a^\varepsilon = (a^\varepsilon)^{n+1}$ since $c_{n,n} = 1$. We have thus proved (28) for $q = 0$.

When $q > 0$ we differentiate (28),

$$\begin{aligned}
\partial_t^p \partial_x^q u^\varepsilon &= \sum_{j=1}^{2p} \varepsilon^{j-2p} \partial_x^q r_{j,p,0}^\varepsilon \partial_x^j u^\varepsilon = \sum_{j=1}^{2p} \sum_{\ell=0}^{q} \varepsilon^{j-2p} c_{q,\ell} (\partial_x^{q-\ell} r_{j,p,0}^\varepsilon) \partial_x^{j+\ell} u^\varepsilon \\
&= \sum_{j=1}^{2p+q} \sum_{\ell=0}^{\min(q,j-1)} \varepsilon^{j-\ell-2p} c_{q,\ell} (\partial_x^{q-\ell} r_{j-\ell,p,0}^\varepsilon) \partial_x^j u^\varepsilon,
\end{aligned}$$

which agrees with (28) when we identify

$$r^{\varepsilon}_{j,p,q} = \sum_{\ell=0}^{\min(q,j-1)} c_{q,\ell}(\varepsilon\partial_x)^{q-\ell} r^{\varepsilon}_{j-\ell,p,0}.$$

By Lemma A.1 these functions all belong to $\mathcal{E}$. Finally, since $r^{\varepsilon}_{j,p,0} = 0$ when $j > 2p$,

$$r^{\varepsilon}_{2p+q,p,q} = \sum_{\ell=0}^{q} c_{q,\ell}(\varepsilon\partial_x)^{q-\ell} r^{\varepsilon}_{2p+q-\ell,p,0} = c_{q,q} r^{\varepsilon}_{2p,p,0} = (a^{\varepsilon})^p,$$

and we have shown (28) for all $q \geq 0$. ∎

LEMMA A.3. *Suppose $a^{\varepsilon}(t,x) \in \mathcal{E}$ satisfies* (2) *and*

$$u_t = (a^{\varepsilon} u_x)_x + W_x, \qquad v_t = (a^{\varepsilon} v)_{xx} + W_x, \quad t \geq 0.$$

Then

$$||u(t,\cdot)|| \leq ||u(0,\cdot)|| + C(t) \sup_{0\leq s\leq t} ||W(s,\cdot)||, \tag{29}$$

$$||v(t,\cdot)|| \leq C(t) \left(||v(0,\cdot)|| + \sup_{0\leq s\leq t} ||W(s,\cdot)|| \right). \tag{30}$$

Proof. For $u(t,x)$ we get

$$\begin{aligned}\frac{1}{2}\partial_t ||u||^2 &= \langle u, u_t\rangle = -\langle u_x, a^{\varepsilon} u_x\rangle - \langle u_x, W(x)\rangle \\ &\leq -a_{\min}||u_x||^2 + ||u_x||\,||W|| \leq \frac{1}{4a_{\min}}||W||^2.\end{aligned}$$

Consequently,

$$||u(t,\cdot)||^2 \leq ||u(0,\cdot)||^2 + \frac{1}{2a_{\min}} \int_0^t ||W(s,\cdot)||^2\, ds,$$

from which (29) follows. Furthermore,

$$\begin{aligned}\frac{1}{2}\partial_t \langle v, a^{\varepsilon} v\rangle &= \frac{1}{2}\langle v, a^{\varepsilon}_t v\rangle + \langle a^{\varepsilon} v, v_t\rangle = \frac{1}{2}\langle v, a^{\varepsilon}_t v\rangle - ||(a^{\varepsilon} v)_x||^2 - \langle (a^{\varepsilon} v)_x, W(x)\rangle \\ &\leq \frac{|a^{\varepsilon}_t|_{\infty}}{2a_{\min}}\langle v, a^{\varepsilon} v\rangle + \frac{1}{4}||W||^2.\end{aligned}$$

By Grönwall's lemma (see, e.g., [24]),

$$||v||^2 \leq \frac{1}{a_{\min}}\langle v, a^{\varepsilon} v\rangle \leq \frac{C(t)}{a_{\min}} \left(||v(0,\cdot)||^2 + \int_0^t ||W(s,\cdot)||^2\, ds \right).$$

This shows (30). ∎

LEMMA A.4. *Suppose that $a^{\varepsilon}(t,x) \in \mathcal{E}$ and that u^{ε} is the solution to* (1) *with initial data $f \in H^{2M+1}$. Then for all $1 \leq n \leq M$ and $t > 0$ there are constants $C(n,t)$, independent of ε, such that*

$$||\partial_t^n u^{\varepsilon}(t,\cdot)|| \leq C(n,t) \left(\varepsilon^{1-2n}||f||_{2n} + \sup_{0\leq s\leq t} \sum_{j=1}^{2n-1} \varepsilon^{j-2n+1} ||\partial_x^j u^{\varepsilon}(t,\cdot)|| \right), \tag{31}$$

$$||\partial_t^n u_x^{\varepsilon}(t,\cdot)|| \leq C(n,t) \left(\varepsilon^{-2n}||f||_{2n+1} + \sup_{0\leq s\leq t} \sum_{j=1}^{2n} \varepsilon^{j-2n} ||\partial_x^j u^{\varepsilon}(t,\cdot)|| \right). \tag{32}$$

Proof. We define $W(x)$ as

$$\partial_t^{n+1} u^\varepsilon = \partial_t^n \partial_x a^\varepsilon \partial_x u^\varepsilon = \sum_{j=0}^{n} c_{nj} \partial_x (\partial_t^{n-j} a^\varepsilon) \partial_t^j \partial_x u^\varepsilon =: \partial_x a^\varepsilon \partial_x \partial_t^n u^\varepsilon + \partial_x W(x). \tag{33}$$

Then, by Lemma A.3,

$$||\partial_t^n u^\varepsilon(t,\cdot)|| \le ||\partial_t^n u^\varepsilon(0,\cdot)|| + C(t) \sup_{0\le s\le t} ||W(s,\cdot)||. \tag{34}$$

For $W(t,x)$ we have by Lemma A.2 with $q=1$

$$W(t,x) = \sum_{j=0}^{n-1} c_{nj} (\partial_t^{n-j} a^\varepsilon) \partial_t^j u_x^\varepsilon = \sum_{j=0}^{n-1} c_{nj} (\partial_t^{n-j} a^\varepsilon) \sum_{\ell=1}^{2j+1} \varepsilon^{\ell-2j-1} r_{\ell,j,1}^\varepsilon \partial_x^\ell u^\varepsilon.$$

Hence, since $a^\varepsilon, r_{\ell,j,1}^\varepsilon \in \mathcal{E}$,

$$||W(t,\cdot)|| \le C \sum_{j=0}^{n-1} \sum_{\ell=1}^{2j+1} \varepsilon^{\ell-2j-1} ||\partial_x^\ell u^\varepsilon(t,\cdot)|| \le C \sum_{\ell=1}^{2n-1} \varepsilon^{\ell-2n+1} ||\partial_x^\ell u^\varepsilon(t,\cdot)||. \tag{35}$$

Lemma A.2 with $q=0$ also shows us that

$$||\partial_t^n u^\varepsilon(0,\cdot)|| = \left\| \sum_{j=1}^{2n} \varepsilon^{j-2n} r_{j,n,0}^\varepsilon \partial_x^j f \right\| \le C(n) \varepsilon^{1-2n} ||f||_{2n}.$$

Together with (34) and (35) this shows (31). For (32), we differentiate (33) with respect to x,

$$\partial_t^{n+1} u_x^\varepsilon = \partial_{xx} a^\varepsilon \partial_t^n u_x^\varepsilon + \partial_{xx} W,$$

and by Lemma A.3,

$$||\partial_t^n u_x^\varepsilon(t,\cdot)|| \le C(t) \left(||\partial_t^n u_x^\varepsilon(0,\cdot)|| + \sup_{0\le s\le t} ||W_x(s,\cdot)|| \right). \tag{36}$$

Letting $s_{\ell,j}^\varepsilon := c_{nj} (\partial_t^{n-j} a^\varepsilon) r_{\ell,j,1}^\varepsilon \in \mathcal{E}$, we get from Lemma A.2 with $q=1$

$$\begin{aligned} W_x(t,x) &= \partial_x \sum_{j=0}^{n-1} \sum_{\ell=1}^{2j+1} \varepsilon^{\ell-2j-1} s_{\ell,j}^\varepsilon \partial_x^\ell u^\varepsilon \\ &= \sum_{j=0}^{n-1} \sum_{\ell=1}^{2j+1} \varepsilon^{\ell-2j-2} (\varepsilon \partial_x s_{\ell,j}^\varepsilon) \partial_x^\ell u^\varepsilon + \sum_{j=0}^{n-1} \sum_{\ell=2}^{2j+2} \varepsilon^{\ell-2j-2} s_{\ell-1,j}^\varepsilon \partial_x^\ell u^\varepsilon. \end{aligned}$$

Hence, since $s_{\ell,j}^\varepsilon \in \mathcal{E}$,

$$||W_x(t,\cdot)|| \le C \sum_{j=0}^{n-1} \sum_{\ell=1}^{2j+2} \varepsilon^{\ell-2j-2} ||\partial_x^\ell u^\varepsilon(t,\cdot)|| \le C \sum_{\ell=1}^{2n} \varepsilon^{\ell-2n} ||\partial_x^\ell u^\varepsilon(t,\cdot)||. \tag{37}$$

Lemma A.2 with $q=1$ also shows us that

$$||\partial_t^n u_x^\varepsilon(0,\cdot)|| = \left\| \sum_{j=1}^{2n+1} \varepsilon^{j-2n-1} r_{j,n,0}^\varepsilon \partial_x^j f \right\| \le C(n) \varepsilon^{-2n} ||f||_{2n+1}.$$

Together with (36) and (37) this shows (32). $\square$

A.2. Proof of Theorem 3.2. We now proceed to show Theorem 3.2 by induction. The right inequality in (13) and the case $p = 1$ follow directly from Lemma A.3 with $W \equiv 0$. Suppose that the statement is true up to an odd number, $p = 2n-1 < M$. By Lemma A.2 with $q = 0$,

$$\partial_t^n u = \sum_{j=1}^{2n} \varepsilon^{j-2n} r^{\varepsilon}_{j,n,0} \partial_x^j u = (a^\varepsilon)^n \partial_x^{2n} u^\varepsilon + \sum_{j=1}^{2n-1} \varepsilon^{j-2n} r^{\varepsilon}_{j,n,0} \partial_x^j u,$$

where $|r^{\varepsilon}_{j,n,0}|_\infty \leq C$. Therefore, using Lemma A.4

$$\begin{aligned} ||\partial_x^{2n} u|| &\leq (a^\varepsilon)^{-n} ||\partial_t^n u|| + C \sum_{j=1}^{2n-1} \varepsilon^{j-2n} ||\partial_x^j u|| \\ &\leq C(n,t) \left(\varepsilon^{1-2n} ||f||_{2n} + \sup_{0 \leq s \leq t} \sum_{j=1}^{2n-1} \varepsilon^{j-2n+1} ||\partial_x^j u^\varepsilon(t,\cdot)|| \right) + C \sum_{j=1}^{2n-1} \varepsilon^{j-2n} ||\partial_x^j u||, \end{aligned}$$

and by the induction hypothesis we get $||\partial_x^{2n} u|| \leq C\varepsilon^{1-2n} ||f||_{2n}$. On the other hand, if it is true up to an even number, $p = 2n < M$, then we get from Lemma A.2 with $q = 1$

$$\partial_t^n u_x = \sum_{j=1}^{2n+1} \varepsilon^{j-2n-1} r^{\varepsilon}_{j,n,1} \partial_x^j u = (a^\varepsilon)^n \partial_x^{2n+1} u^\varepsilon + \sum_{j=1}^{2n} \varepsilon^{j-2n-1} r^{\varepsilon}_{j,n,1} \partial_x^j u,$$

where as before $|r^{\varepsilon}_{j,n,1}|_\infty \leq C$, and by Lemma A.4,

$$\begin{aligned} ||\partial_x^{2n+1} u|| &\leq (a^\varepsilon)^{-n} ||\partial_t^n u_x|| + C \sum_{j=1}^{2n} \varepsilon^{j-2n-1} ||\partial_x^j u|| \\ &\leq C(n,t) \left(\varepsilon^{-2n} ||f||_{2n+1} + \sup_{0 \leq s \leq t} \sum_{j=1}^{2n} \varepsilon^{j-2n} ||\partial_x^j u^\varepsilon(t,\cdot)|| \right) + C \sum_{j=1}^{2n} \varepsilon^{j-2n-1} ||\partial_x^j u||. \end{aligned}$$

From the induction hypothesis we conclude that $||\partial_x^{2n+1} u|| \leq C\varepsilon^{-2n} ||f||_{2n+1}$, which shows the theorem.

Acknowledgments. For many helpful discussions we thank Björn Engquist, who initially suggested using RAℓSFA to solve PDEs, Eitan Tadmor, and Weinan E. We would also like thank Anna Gilbert and Martin Strauss for their RAℓSFA code.

REFERENCES

[1] A. Abdulle and W. E, *Finite difference heterogeneous multiscale method for homogenization problems*, J. Comput. Phys., 191 (2003), pp. 18–39.

[2] V. Arnold, *Mathematical Methods in Classical Mechanics*, Springer-Verlag, Berlin, New York, 1978.

[3] I. Babuška, G. Caloz, and J. E. Osborn, *Special finite element methods for a class of second order elliptic problems with rough coefficients*, SIAM J. Numer. Anal., 31 (1994), pp. 945–981.

[4] A. Bensoussan, J. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North–Holland, Amsterdam, New York, 1978.

[5] G. Beylkin, M. E. Brewster, and A. C. Gilbert, *A multiresolution strategy for numerical homogenization of nonlinear ODEs*, Appl. Comput. Harmon. Anal., 5 (1998), pp. 450–486.

[6] G. Beylkin and M. Brewster, *A multiresolution strategy for numerical homogenization*, Appl. Comput. Harmon. Anal., 2 (1995), pp. 327–349.

[7] R. B. Bird, C. F. Curtiss, R. C. Armstrong, and O. Hassager, *Dynamics of Polymeric Liquids, Volume* 2: *Kinetic Theory*, Wiley, New York, 1987.

[8] R. Car and M. Parrinello, *Unified approach for molecular dynamics and density-functional theory*, Phys. Rev. Lett., 55 (1985), pp. 2471–2474.

[9] S. Chen, W. E, and C.-W. Shu, *The heterogeneous multiscale method based on the discontinuous Galerkin method for hyperbolic and parabolic problems*, Multiscale Model. Simul., 3 (2005), pp. 871–894.

[10] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, New York, 1978.

[11] W. E and B. Engquist, *Multiscale modeling and computation*, Notices Amer. Math. Soc., 50 (2003), pp. 1062–1070.

[12] W. E and B. Engquist, *The heterogeneous multiscale methods*, Commun. Math. Sci., 1 (2003), pp. 87–133.

[13] W. E, B. Engquist, and Z. Huang, *Heterogeneous multiscale method: A general methodology for multiscale modeling*, Phys. Rev. B, 67 (2003), 092101.

[14] B. Engquist and O. Runborg, *Wavelet-based numerical homogenization with applications*, in Multiscale and Multiresolution Methods: Theory and Applications, Lect. Notes Comput. Sci. Eng. 20, T. J. Barth, T. F. Chan, and R. Haimes, eds., Springer-Verlag, Berlin, 2001, pp. 97–148.

[15] M. Frigo and S. Johnson, *The design and implementation of FFTW*3, Proc. IEEE, 93 (2005), pp. 216–231.

[16] A. C. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. Strauss, *Near-optimal sparse Fourier representations via sampling*, in Proceedings of the 34th Annual ACM Symposium on Theory of Computing, 2002, pp. 152–161.

[17] A. C. Gilbert, S. Muthukrishnan, and M. Strauss, *Improved time bounds for near-optimal sparse Fourier representation*, in Wavelets XI, Proc. SPIE 5914, M. Papadakis, A. F. Laine, and M. A. Unser, eds., SPIE, Bellingham, WA, 2005, pp. 398–412.

[18] D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, Philadelphia, 1977.

[19] T. Y. Hou and X.-H. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.

[20] T. Y. Hou, X.-H. Wu, and Z. Cai, *Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients*, Math. Comp., 68 (1999), pp. 913–943.

[21] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J.-B. Quincy, *The variational multiscale method—a paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 3–24.

[22] J. Kevorkian and J. D. Cole, *Multiple Scale and Singular Perturbation Methods*, Springer-Verlag, New York, 1996.

[23] I. G. Kevrekidis, C. W. Gear, J. M. Hyman, P. G. Kevrekidis, O. Runborg, and C. Theodoropoulos, *Equation-free, coarse-grained multiscale computation: Enabling microscopic simulators to perform system-level analysis*, Commun. Math. Sci., 1 (2003), pp. 715–762.

[24] H.-O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier–Stokes Equations*, SIAM, Philadelphia, 2004.

[25] A.-M. Matache and C. Schwab, *Homogenization via p-FEM for problems with microstructure*, Appl. Numer. Math., 33 (2000), pp. 43–59.

[26] P. Ming and P. Zhang, *Analysis of the heterogeneous multiscale method for parabolic homogenization problems*, Math. Comp., 76 (2007), pp. 153–177.

[27] G. Samaey, D. Roose, and I. G. Kevrekidis, *The gap-tooth scheme for homogenization problems*, Multiscale Model. Simul., 4 (2005), pp. 278–306.

[28] E. B. Tadmor, M. Ortiz, and R. Phillips, *Quasicontinuum analysis of defects in crystals*, Phil. Mag. A, 73 (1996), pp. 1529–1563.

[29] C. Theodoropoulos, Y.-H. Qian, and I. G. Kevrekidis, *"Coarse" stability and bifurcation analysis using time-steppers: A reaction-diffusion example*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 9840–9843.

[30] L. N. Trefethen, *Spectral Methods in MATLAB*, SIAM, Philadelphia, 2000.

[31] R. Voigt, D. Gottlieb, and M. Hussaini, eds., *Spectral Methods for Partial Differential Equations*, SIAM, Philadelphia, 1984.

[32] J. Zou, A.C. Gilbert, M. Strauss, and I. Daubechies, *Theoretical and experimental analysis of a randomized algorithm for sparse Fourier transform analysis*, J. Comput. Phys., 211 (2006), pp. 572–595.
[33] J. Zou, *Sublinear Algorithms for the Fourier Transform of Signals with Very Few Fourier Modes: Theory, Implementations and Applications*, Ph.D. dissertation, Princeton University, Princeton, NJ, 2005.