SCALABLE MATRIX-FREE ADAPTIVE PRODUCT-CONVOLUTION APPROXIMATION FOR LOCALLY TRANSLATION-INVARIANT OPERATORS∗

NICK ALGER†, VISHWAS RAO‡, AARON MYERS†, TAN BUI-THANH§, AND OMAR GHATTAS¶

Abstract. We present an adaptive grid matrix-free operator approximation scheme based on a “product-convolution” interpolation of convolution operators. This scheme is appropriate for operators that are locally approximately translation-invariant, even if these operators are high-rank or full-rank. Such operators arise in Schur complement methods for solving partial differential equations (PDEs), as Hessians in PDE-constrained optimization and inverse problems, as integral operators, as covariance operators, and as Dirichlet-to-Neumann maps. Constructing the approximation requires computing the impulse responses of the operator to delta function sources centered on nodes in an adaptively refined grid of sample points. A randomized a-posteriori error estimator drives the adaptivity. Once constructed, the approximation can be efficiently applied to vectors using the fast Fourier transform (FFT), and used in place of the operator. The approximation can also be efficiently converted to hierarchical matrix (H-matrix) format, then inverted or factorized using scalable H-matrix arithmetic. The quality of the approximation degrades gracefully as fewer sample points are used, allowing cheap lower quality approximations to be used as preconditioners. We directly address issues related to boundaries and prove that our scheme eliminates boundary artifacts. We test the scheme on the non-local component of a boundary Schur complement for the Poisson operator, and on the data misfit Hessian for an advection dominated advection-diffusion inverse problem. Numerical results show that the number of sample points required is independent of the mesh size for the non-local component of the Poisson Schur complement, and independent of the informativeness of the data for the advection-diffusion inverse problem data misfit Hessian.

Key words. convolution, operator approximation, hierarchical matrix, H-matrix, PDE-constrained optimization, inverse problems, data scalability, matrix-free, preconditioning, Hessian, Schur complement

AMS subject classifications. 41A05, 41A35, 42A61, 42A85, 47A58, 49K20, 65F08, 65F22, 65N21, 65T50, 94A12

1. Introduction. In this paper we present a product-convolution approximation scheme for locally approximately translation-invariant operators. That is, operators $A : l^2(\Omega) \rightarrow l^2(\Omega)$ satisfying

$$A [y, x] \approx A [y - x + p, p]$$

whenever $x$ is not too far from $p$ (see Figure 1). Our scheme is well-suited for approximating or preconditioning operators that arise in Schur complement methods for solving partial differential equations (PDEs), reduced Hessians in PDE-constrained

∗This work was partially supported by the following grants: AFOSR FA9550-17-1-0190, NSF ACI-1550593 and CBET-1508713, and DOE DE-SC0010518 and DE-SC0009286. Vishwas Rao was partially supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research Program under contract DE-AC02-06CH11357.

†Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, USA (nalger@ices.utexas.edu, aaron@ices.utexas.edu, tanbui@ices.utexas.edu, omar@ices.utexas.edu).

‡Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL, USA (vhebbur@anl.gov).

§Department of Aerospace Engineering and Engineering Mechanics, and Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, USA.

¶Departments of Geological Sciences and Mechanical Engineering, The University of Texas at Austin, Austin, TX, USA.
optimization and inverse problems, integral operators, covariance operators with spatially varying kernels, and Dirichlet-to-Neumann maps or other Poincaré–Steklov operators in multiphysics problems. These operators are often dense, implicitly defined, and high-rank, making them difficult to approximate with standard techniques (see Section 1.3 for an overview of existing techniques).

Let $\varphi_p$ be the impulse response of $A$ at $p$, that is, the function created by applying $A$ to a delta function source centered at point $p$, then translating the result to recenter it at 0:

$$\varphi_p[z] = (A\delta_p)[z + p], \quad z \in \Omega - p.$$  \hspace{1cm} (2)

If $A$ were translation-invariant (i.e., if (1) held with equality for all $x, y$), then $A$ would be the convolution operator $A : f \mapsto \varphi_p * f$. To approximate operators that are only locally approximately translation-invariant, we patch together a collection of convolution operators, each of which well-approximates $A$ locally. Our approximation of $A$, denoted $\tilde{A}$, takes the following form:

$$Af \approx \tilde{A}f := \sum_{k=1}^{r} \varphi_{E_k}^k * (w_k \cdot f),$$ \hspace{1cm} (3)

where the $w_k$ are locally supported weighting functions that overlap and form a partition of unity, ‘$\cdot$’ denotes pointwise multiplication of functions, ‘$*$’ denotes convolution (see Section 1.4 for more details on notation), and the functions $\varphi_{E_k}^k$ are modified\(^{1}\) versions of the (translated, recentered) impulse responses $\varphi_{p_k}$ associated with a collection of sample points, $p_k$. Each point $p_k$ is contained within the support of the associated weighting function $w_k$.

The basic form of (3) is known as a product-convolution approximation, and is well-established in the literature (see Section 1.3.2). Here we improve upon existing schemes by:

\(^{1}\)To address issues with boundary artifacts, we construct $\varphi_{E_k}^k$ by extending the function $\varphi_{p_k}$ outside of $\Omega - p_k$ using information from neighboring functions, $\varphi_{p_j}$ (more on this in Section 2.6).
• Adaptively and automatically choosing the sample points $p_k$.
• Addressing issues related to boundaries.

In Section 2 we derive our scheme, explain how we choose $p_k$, and detail the process for constructing $w_k$ and $\varphi_k^E$. In Section 3 we detail how $\tilde{A}$ can be used once constructed, including how to efficiently convert it to hierarchical matrix ($H$-matrix) format. In Section 4 we perform an a-priori error analysis of our scheme, and a statistical analysis of the randomized a-posteriori error estimator used for the adaptivity. We demonstrate our scheme numerically in Section 5 and give concluding remarks in Section 6. In the remainder of this section we summarize our results (Section 1.1), outline limitations and extensions of our scheme (Section 1.2), review existing work (Section 1.3), and define our setting and notation (Section 1.4).

1.1. Overview of results.

Generality. Our scheme is matrix-free in the sense that we do not need access to the matrix representation of $A$. We only need the ability to apply $A$ and its adjoint, $A^*$, to vectors. Specifically, constructing $\tilde{A}$ requires applying $A$ to one delta function vector per sample point, $p_k$, plus applying $A^*$ to a fixed number of Gaussian random vectors (we summarize the cost to construct $\tilde{A}$ in Section 2.9). Our scheme does not require a low-rank assumption on $A$, and does not require any sparsity assumptions on $A$’s matrix representation.

Usage. Once constructed, we can compute individual matrix entries of $\tilde{A}$ in $O(1)$ work (Section 3.1). In contrast, since we can only access $A$ by applying it to vectors, we do not have $O(1)$ access to individual matrix entries of $A$. We can apply $A$ and $A^*$ to vectors in nearly linear work using the fast Fourier transform (FFT) (Section 3.2). Further, blocks of $\tilde{A}$ and $\tilde{A}^*$ can be applied to vectors in work that is nearly linear in the size of the block (Section 3.3). Often the ultimate goal is to solve linear systems with $A$ as the coefficient operator. Since $\tilde{A}$ can be applied to vectors cheaply, linear systems with $\tilde{A}$ as the coefficient operator can solved iteratively with Krylov methods such as conjugate gradient or GMRES [22]. However, the convergence of Krylov methods depends heavily on the spectral structure of the coefficient operator, leading to slow convergence when $\tilde{A}$ is ill-conditioned. To address this, we present an efficient method to convert $\tilde{A}$ to $H$-matrix format (Section 3.4). Once in $H$-matrix format, $\tilde{A}$ can be efficiently factorized or inverted using $H$-matrix arithmetic. Using fewer sample points yields a proportionately less accurate approximation; in this case the resulting $H$-matrix can be used as a preconditioner. Alternatively, after $\tilde{A}$ is constructed one can build circulant preconditioners from it [23, 58]; we do not pursue circulant preconditioning in this paper.

Adaptivity. We choose the sample points, $p_k$, in an adaptive grid (Section 2.3). In regions where the error is large, we refine the grid, and in regions where the error is small, we leave the grid alone (Section 2.4). The effect of this refinement process is to place more sample points in regions where $A$ is less translation-invariant, and fewer sample points in regions where $A$ is more translation-invariant. The adaptivity is performed using a theoretically justified (Theorem 7) randomized a-posteriori error estimator (Section 2.7). In practice, we find that the randomized a-posteriori error estimator performs much better than the theory predicts: in our numerical examples (Section 5), we see that it performs almost as well with 5 random samples as it does with 100.

Boundaries. Boundaries introduce two difficulties for product-convolution schemes:

1. Boundary artifacts: The impulse response associated with $p_k$ is naturally defined on $\Omega - p_k$, but the product-convolution scheme (3) requires it to be...
defined on a larger set. The three standard extension techniques—extending the impulse response by zero, reflecting it across the boundary, or replicating it periodically—all create boundary artifacts wherever artificial data is used in place of undefined data.

2. Boundary effects: The underlying operator may fail to be translation-invariant near boundaries due to boundary conditions or other physically meaningful effects.

To overcome 1., we extend the support of the impulse responses using information from neighboring impulse responses (Section 2.6). To overcome 2., we use anisotropic adaptivity. Our adaptive refinement scheme senses the coordinate direction in which $A$ is least translation-invariant within a cell, and preferentially subdivides the cell in that direction (Section 2.8). This allows the scheme to efficiently approximate operators that are not translation-invariant in directions perpendicular to boundaries, but are translation-invariant in directions parallel to boundaries. Boundary effects due to boundary conditions typically exhibit this direction-dependent form of translation-invariance (regardless of the type of boundary condition).

Theory. In Theorem 6, we prove that the error in our scheme is controlled by the local failure of translation-invariance in $A$. This, together with adaptivity, implies convergence: our scheme will continue to add new sample points until it achieves the desired error tolerance. The more translation-invariant $A$ is, the fewer sample points will be used. Additionally, Theorem 6 implies that our approximation scheme will not introduce boundary artifacts. Without our impulse response extension procedure (Section 2.6), the bound in Theorem 6 would fail near the boundary.

Numerical results. We demonstrate the scheme on the non-local component of a boundary Schur complement for the Poisson operator (Section 5.1), and on the data misfit Hessian for an advection dominated advection-diffusion inverse problem (Section 5.2). Numerical results show that the number of sample points required is independent of the mesh size for the non-local component of the Poisson Schur complement, and independent of the informativeness of the data for the advection-diffusion inverse problem data misfit Hessian.

1.2. Limitations and extensions. Throughout the paper we assume that the domain $\Omega$ is a compactly supported box in $\mathbb{Z}^d$, as might arise in piecewise linear approximation of a function on a box discretized with a regular rectilinear mesh. The scheme we present can be extended to irregular unstructured meshes by converting back and forth between the irregular mesh and an overlaid regular mesh with sampling and interpolation. Extension to the continuum setting—where $\Omega$ is a box in $\mathbb{R}^d$—is also possible. There one replaces $l^2$ spaces with $L^2$ spaces, sums with integrals, and so forth.

Although we assume that the domain and range of $A$ are the same, the scheme and theory extend in a straightforward way to operators that map between $l^2$ spaces defined on different boxes, and even boxes of different dimensions. This includes operators that characterize the interactions between degrees of freedom on a facet of a boundary of a domain and degrees of freedom in the whole domain.

The scheme can also be extended to more complicated domains (non-box) by either:

1. Embedding the domain within a larger box,
2. Mapping the domain to a box, or
3. Covering the domain with a collection of “mapped box” patches, using our scheme on each patch, and combining these patch approximations using par-
tition of unity techniques.
If the domain is embedded within a larger box (1), the theory we present regarding boundaries no longer holds. If the domain is mapped to a box (2), the theory still holds, but geometric distortion may decrease translation-invariance, necessitating more sample points to achieve the desired approximation accuracy. Combining local patch approximations (3) requires one to either ignore long-range interactions between patches, or deal with these interactions using other techniques. In some situations well-chosen geometric distortion may increase translation-invariance. For example, in problems involving fluid flow it may be beneficial to align the grid with streamlines.

Although our scheme will eventually converge to any desired error tolerance, it is most useful for computing moderately accurate approximations (say, 80% to 99% accurate) of "difficult" operators that are poorly approximated by standard techniques. In our numerical tests, we observe that the convergence slows beyond this level of accuracy. Moderately accurate approximations are good enough for many engineering applications; in situations where more accuracy is needed, the approximation can be used as a preconditioner.

1.3. Existing work. The most widely used, robust, and general purpose matrix-free operator approximation schemes are based on low-rank approximation (Section 1.3.1). However, many important operators in PDEs, PDE-constrained optimization and inverse problems, and integral equations are not low-rank. Our scheme fits within a class of operator approximation schemes based on interpolation of convolution operators (Section 1.3.2). Despite the fact that even a single convolution operator can be high-rank or full-rank, there are two senses in which interpolation of convolution operators can be interpreted as a low-rank approximation (Sections 1.3.3 and 1.3.4). We do not use these low-rank interpretations directly in the development of our scheme. However, they provide insight into our scheme, and connect convolution interpolation to other relevant forms of operator approximation including matrix probing (Section 1.3.5). Convolution interpolation is also closely connected to sparsity in wavelet domains (Section 1.3.6). Hierarchical matrices (Section 1.3.7) are another well-established operator approximation format; they are simultaneously a tool we use (Section 3.4), and an alternative to our scheme.

1.3.1. Low-rank approximation. Low-rank approximations—matrix factorizations of the form

\[ A \approx BC, \]

where \( B \) is \( N \times r \) (tall), and \( C \) is \( r \times N \) (wide)—can be efficiently constructed in a matrix-free setting by using Krylov methods (Lanczos or Arnoldi), randomized SVD [45] or CUR decomposition/skeletonization [26, 54, 67]. Although low-rank approximations have been used for Dirichlet-to-Neumann maps [20, 21], full-rank or high-rank operators typically still retain a high rank after being restricted to a boundary as a Schur complement. Likewise, although low-rank approximations have been used to approximate the (prior preconditioned) Hessian of the data misfit term in PDE-constrained inverse problems [19, 29, 37, 61, 64], the numerical rank of this term grows as the informativeness of the data in the inverse problem grows [4], making low-rank approximation inefficient for highly informative data. Moreover, even when the operator is low-rank in the sense that \( r << N \), the cost of computing the low-rank approximation may be prohibitive. For example, a low-rank approximation of the Hessian in a PDE constrained optimization or inverse problem requires \( O(r) \)
linearized forward/adjoint PDE solves, so that for large-scale problems with e.g. \( N \) of order \( 10^6 \), even a compression of 0.1\% still means that thousands of forward solves are needed, which is often an expensive proposition [18, 25, 50]. Our scheme is motivated by a desire to go beyond low-rank approximation in these applications.

1.3.2. Convolution interpolation. Since the linear operator that performs a convolution may be numerically full-rank (e.g., convolution with a delta function: the identity operator) or high-rank (e.g., convolution with a Gaussian with a small width), interpolation of convolution operators can, where applicable, be used to approximate dense operators with far fewer terms than the rank of the operator.

Operator approximation schemes based on weighted sums of convolution operators with spatially varying weights (“convolution interpolations”) fall into two categories: product-convolution schemes where one performs element-wise products with weighting functions first and convolutions second, and convolution-product schemes where this order is reversed:

\[
Af \approx \sum_{k=1}^{r} \psi_k \ast (\omega_k \cdot f) \quad \text{vs.} \quad \sum_{k=1}^{r} \omega_k \cdot (\psi_k \ast f).
\]

The terms “product-convolution” and “convolution-product” refer to the general format of the approximations in (4), where \( \psi_k \) and \( \omega_k \) could be any functions. For us, \( \psi_k \) are (modified) impulse response functions and \( \omega_k \) form a partition of unity.

Since the entries of a convolution operator \( M : f \mapsto \psi \ast f \) are \( M[y, x] = \psi[y - x] \), product-convolution and convolution-product approximations have the following the \((y, x)\) matrix entries:

\[
A[y, x] \approx \sum_{k=1}^{r} \omega_k [x] \psi_k [y - x] \quad \text{vs.} \quad \sum_{k=1}^{r} \omega_k [y] \psi_k [y - x].
\]

Both schemes are non-symmetric, but the adjoint of a product-convolution operator is a convolution-product operator, and vice versa. The operators defined by the following actions are adjoints of each other:

\[
\sum_{k=1}^{r} \psi_k \ast (\omega_k \cdot f) \xrightarrow{\text{adjoint}} \sum_{k=1}^{r} \omega_k \cdot (\text{flip} \, (\psi_k) \ast f),
\]

where \( \text{flip} \, (\psi) [x] := \psi [-x] \), and the over-line indicates the complex conjugate. Here we use a product-convolution scheme.

Convolution interpolation schemes have been used in image restoration and deblurring [36, 58, 59, 65] in photography [66], astronomy [1, 44, 63], and microscopy [62], as well as in wireless communication signal processing [49], ultrasound imaging [60], systems biology [40], and Hessian approximation in seismic inversion [69].\(^2\) Aside from the application, convolution interpolation schemes differ based on how they construct the functions \( \omega_k \) and \( \psi_k \). For a comprehensive overview of existing schemes

\(^2\)In many of these applications, the impulse response is known as the point spread function (PSF), as it corresponds to the spreading of a point source of light as it passes through an optical system.
for constructing these functions, we refer the reader to the excellent summaries in [32, 34, 39].

Broadly, existing schemes can be categorized based on whether the span of the functions $\omega_k$ is fixed, or the span of the functions $\psi_k$ is fixed, or both of the spans are fixed, or neither of the spans are fixed. Schemes then attempt to find the remaining (not fixed) functions so that the error in the resulting operator approximation is small. Established choices for the span of the functions $\psi_k$ include the span of impulse responses of $A$ to a collection of delta function sources at fixed locations (we do this), subspaces of this span, and the span of functions with known analytic forms (e.g., Gaussians, spherical harmonics). Established choices for the span of the functions $\omega_k$ include spans of Fourier modes, piecewise polynomials on a regular grid (e.g., piecewise constants, piecewise linear functions, B-splines), wavelets, radial basis functions [12], and functions based on kriging.

On one hand, existing schemes in which the functions $\omega_k$ are not fixed require more access to $A$ than just the ability to apply it to vectors. On the other hand, existing schemes in which the functions $\omega_k$ are fixed do not permit spatial adaptivity, with one exception. This includes existing sectioning approaches that partition the domain into pieces on a regular grid, then use different functions $\psi_k$ for each piece [58]. The exception is [9], which, like this paper, proposes partitioning the domain with an adaptively refined grid. However, [9] only proposes the concept; they do not provide practical algorithms to perform the adaptivity.

1.3.3. Low-rank approximation of the SVIR. Product-convolution schemes are equivalent to low-rank approximations to the spatially varying impulse response function (SVIR):

$$R[y,x] := A[y + x, x].$$

Given a low-rank approximation $R \approx \Psi V^T$, the $k$th column of $\Psi$ is the function $\psi_k$ and the $k$th column of $V$ is the function $\omega_k$ used in a product-convolution scheme [32, 38]. Hence, computing a low-rank approximation of $R$ yields a product-convolution scheme for approximating $A$. However, since the ability to apply $A$ to vectors does not grant the ability to apply $R$ to vectors, matrix-free low-rank approximation algorithms cannot be applied to $R$ when we only have matrix-free access to $A$. Additionally, the domain of $R$ is the non-rectangular set $\{(y, x) : x \in \Omega, y + x \in \Omega\}$. In order to interpret $R$ as a matrix, one must extend the domain of $R$ to an enclosing rectangle, creating boundary artifacts wherever the extension procedure assigns values to matrix entries of $R$ that would otherwise be undefined.

1.3.4. Low-rank approximation of the pseudo-differential symbol. On unbounded or periodic domains, convolution-product schemes are equivalent to low-rank approximation of the pseudo-differential symbol [48, 52] of the operator.\(^4\) Suppose

$$\begin{equation}
(Af)[y] = \frac{1}{c} \sum_{\xi} e^{i(y, \xi)} P[y, \xi] \hat{f}(\xi),
\end{equation}$$

\(^3\)The terminology for this is potentially confusing: in the literature, computed (rather than fixed) functions $\omega_k$ are known as “adaptive” weighting functions, but this is unrelated to our “adaptive grid” weighting functions.

\(^4\)To our knowledge, no other papers have explicitly pointed out this connection between convolution-product schemes and pseudo-differential operators.
where $\hat{f}$ is the (discrete) Fourier transform of $f$, $P$ is the (discrete) pseudo-differential symbol of $A$, and $c$ is a normalizing constant. Given a low-rank approximation of the pseudo-differential symbol, $P \approx V \hat{\Psi}^T$, the convolution theorem implies that the columns of $V$ are the weighting functions $\omega_k$ and the columns of $\hat{\Psi}$ are Fourier transforms of the convolution functions $\psi_k$ in a convolution-product scheme.

**1.3.5. Matrix probing.** In matrix probing [27], one approximates $A$ (or $A^{-1}$) as a weighted sum of other known “basis” operators $B_k$, with unknown coefficients $c_k$:

$$A \approx \sum_{k=1}^{r} c_k B_k.$$ 

The operator $A$ and basis operators $B_k$ are applied to random vectors, then a least squares optimization problem is solved to find the coefficients $c_k$ such that the action of the approximate operator on those random vectors best matches the action of the true operator.

In [30], the inverse of the Hessian in a seismic inverse problem is expressed as a weighted sum of operators $B_k$, each with a rank-one pseudo-differential symbol. The weights $c_k$ are computed with matrix probing. The resulting approximation can be manipulated using fast algorithms for pseudo-differential operators [8, 31]. Since convolution-product schemes are equivalent to low-rank approximation of the pseudo-differential symbol of an operator (see Section 1.3.4), under certain boundary conditions the scheme in [30] could be viewed as a convolution-product scheme (though [30] does not present it in this light).

In [11], matrix probing with basis matrices of the form $B_k [y, x] = \omega_k [x + y] \psi_k [y - x]$ is used to approximate the exterior Dirichlet-to-Neumann map for a forward wave propagation problem. This approximation could be viewed as a middle ground between a product-convolution scheme and a convolution-product scheme, which would correspond to $B_k$ matrices of the form $\omega_k [x] \psi_k [y - x]$ and $\omega_k [y] \psi_k [y - x]$, respectively. After constructing the approximation, [11] proposes converting it to $H$-matrix format for further use. Our approximation is different, but we also propose the same subsequent $H$-matrix conversion.

**1.3.6. Wavelets.** The Fourier transform diagonalizes individual convolution operators defined on unbounded or periodic domains. Sparse approximations in wavelet frames—approximations of the form

$$A \approx W^* S W$$

where $W$ is a wavelet or wavelet-like transformation and $S$ is sparse—naturally generalize this concept. Examples include imaging [35], convolution interpolation [33], and seismic inversion [46].

**1.3.7. Hierarchical matrices.** Hierarchical matrices [41] are matrices that may be full-rank, but the blocks of the matrix associated with clusters of degrees of freedom that are far away from each other (or satisfy some other admissibility condition) are low-rank. This structure allows for compressed storage and fast (nearly linear) matrix arithmetic, including matrix inversion and factorization. Special subclasses of $H$-matrices allow for greater compression and faster matrix arithmetic (e.g., $H^2$-matrices [43] and closely related HSS matrices [24, 68], among others). For an overview of $H$- and $H^2$-matrices, see [14, 42].
Classical \( H \)-matrix construction techniques assume that the original matrix is stored in memory in dense or sparse format, then convert it to \( H \)-matrix format by computing low-rank factorizations of its low-rank blocks using standard numerical linear algebra. Since we can access the operator only by applying it and its adjoint to vectors, we cannot use these classical \( H \)-matrix construction techniques.

If the matrix is not stored in memory, but individual matrix entries can be computed efficiently, then the matrix can be efficiently converted to \( H \)-matrix format by performing CUR factorizations for each of its low-rank blocks [10, 15, 67]. We have only matrix-free access to \( A \), so for us these matrix entries are not efficiently computable. However, matrix entries of our approximation, \( \tilde{A} \), are efficiently computable. After constructing \( \tilde{A} \), one can use CUR approximations to convert it to \( H \)-matrix format (see Section 3.4).

More recently, an asymptotically scalable matrix-free \( H \)-matrix construction technique was proposed based on a recursive “peeling process” that involves computing low-rank factorizations of blocks at the coarsest level of the \( H \)-matrix block cluster hierarchy, then using the low-rank factorizations at the coarsest level to help compute low-rank factorizations at the next coarsest level, and so on all the way to the finest level [53]. The method was improved in [55, 56] for hierarchically off-diagonal low-rank (HODLR) matrices and hierarchically block separable (HBS) [57] matrices (restrictive special cases of \( H \) and \( H^2 \) matrices). However, it has several subtle limitations:

- Although asymptotically scalable in theory, in practice the peeling process must apply the original operator to a large number of vectors.
- Attempting to construct a less accurate approximation by applying the original operator to fewer vectors is not advisable (unlike our scheme where this is fine). Errors at any step of the peeling process compound during subsequent steps.
- The approximation procedure is purely algebraic and does not take into account any translation-invariance properties of the underlying operator. This makes the method more general, at the cost of potentially being less efficient than specialized schemes (like ours) that do take advantage of these properties.

Large-scale matrix-free dense operators of the type we are interested in are often formed through algebraic combinations of other operators that are sparse, with the density properties arising due to matrix inverse operations. A typical example is a Schur complement of the form \( S = K_{11} - K_{12}K_{22}^{-1}K_{21} \), where \( K \) is sparse and \( K_{ij} \) are its \((i,j)\) blocks.\(^5\) In principle one can construct an \( H \)-matrix approximation of \( S \) or other similar operators by converting each constituent matrix into an \( H \)-matrix, then algebraically combining them into the desired form using fast \( H \)-matrix arithmetic. However, this process may be inefficient for the following reasons:

- The intermediate matrices may have a much larger hierarchical rank than the final matrix they combine to form.
- The final matrix may operate on functions defined on boundaries, time slices, or other lower dimensional domains than the domains of functions that the intermediate matrices operate on. This is the case for both numerical examples we present.
- Although \( H \)-matrix arithmetic and inversion is asymptotically scalable, it is

\(^5\)Although building a (dense) matrix representation of \( S \) is computationally infeasible in the large-scale context, we can apply \( S \) to a vector by performing matrix-vector multiplications with \( K_{11} \), \( K_{12} \), and \( K_{21} \), and solving a linear system with \( K_{22} \) as the coefficient matrix.
still computationally expensive. Thus it is desirable to avoid intermediate $H$-matrix operations.

1.4. Setting and notation. We work in $l^2$ spaces on $\mathbb{Z}^d$ or subsets of $\mathbb{Z}^d$; these spaces arise when one discretizes a function on a continuous domain using a regular grid. Norms are denoted with $\| \cdot \|$, or occasionally $\| \cdot \|_{l^2(X)}$, if the domain is not clear from context. For linear operators we always use the Frobenius norm (square root of the sum of squares of all entries of the matrix representation of the linear operator). Hence, mappings $M$ between $l^2$ spaces are dually viewed as linear operators $M : l^2(X) \rightarrow l^2(Y)$, or as functions $M \in l^2(Y \times X)$ on a product space.

Intervals on $\mathbb{Z}$ are denoted with brackets: for $a, b \in \mathbb{Z}$, we write $[a, b] := \{x : x \in \mathbb{Z}, a \leq x \leq b\}$. We routinely encounter Cartesian products of intervals, which we call boxes and denote with a bold letter, as in $C$. Boxes are characterized by their minimum point and maximum point: the points in the box that are component-wise less than or equal to all other points in the box, or greater than or equal to all other points in the box, respectively. We denote the minimum and maximum points of a box with the same letter as the box, but lower-case, and with the subscripts “min” and “max”, respectively. E.g.,

$$
C = \prod_{i=1}^{d} [c_{i \text{min}}^i, c_{i \text{max}}^i],
$$

where $\times$ is the Cartesian product of sets. We write corners$(C)$ to denote the set of corners of $C$:

$$
corners(C) := \prod_{i=1}^{d} \{c_{i \text{min}}^i, c_{i \text{max}}^i\}.
$$

The (approximate) midpoint, $c_{\text{mid}}$, of the box $C$ is the integer vector closest to the real vector $(c_{i \text{max}}^i + c_{i \text{min}}^i)/2$. The linear dimension of a box is the sum of all the dimensions of the box: $\sum_{i=1}^{d} c_{i \text{max}}^i - c_{i \text{min}}^i$.

Minkowski set arithmetic is used for addition and subtraction of one set with another set, negation of a set, and addition and subtraction of a set with a point:

$$
X + Y = \{x + y : x \in X, y \in Y\},
$$

$$
X - Y = \{x - y : x \in X, y \in Y\},
$$

$$
-X = \{-x : x \in X\},
$$

$$
X + p = \{x + p : x \in X\},
$$

$$
X - p = \{x - p : x \in X\}.
$$

The number of elements in a set $X$ is denoted $|X|$. We reserve $N$ for the total number of points in the domain: $N := |\Omega|$.

Brackets indicate both $l^2$ function evaluation, and zero-indexed extraction of “blocks” of a vector or operator. That is, the evaluation of $f$ at $x$ is denoted $f[x]$, and $f[C] \in l^2(C - c_{\text{min}})$, with $(f[C])[x] := f[x + c_{\text{min}}]$. Likewise, $M[y, x]$ is the $(y, x)$ “matrix entry” of $M$, and $M[T, S] \in l^2((T - t_{\text{min}}) \times (S - s_{\text{min}}))$ with $(M[T, S])[y, x] := M[y + t_{\text{min}}, x + s_{\text{min}}]$. A dot within indexing brackets, as in $M[C, \cdot]$ or $M[\cdot, C]$, indicates the matrix of all columns or rows of $M$ corresponding to points in $C$, respectively.
The action of a linear operator $M$ on a vector $f$ is denoted $Mf$:

$$(Mf)[y] = \sum_{x \in X} M[y, x] f[x].$$

We write $M^*$ to denote the adjoint of $M$. That is, $M^*[y, x] = \overline{M[x, y]}$, where the over-line indicates the complex conjugate.

A dot between two functions denotes pointwise multiplication of those functions:

$$(f \cdot g)[x] := f[x] g[x].$$

An asterisk between two functions denotes convolution of those functions:

$$(\psi * f)[y] := \sum_{x \in \mathbb{Z}^d} f[x] \psi[y - x].$$

If the domains of functions $f, \psi$ are only subsets of $\mathbb{Z}^d$, we define their convolution to be the result of extending $f, \psi$ by zero so that they are defined on all of $\mathbb{Z}^d$, then convolving them using formula (8). We use the term “convolution rank” to denote the number of terms in a weighted sum of convolution operators (e.g., $r$ in (3)).

The discrete delta function centered at point $p$ is denoted by $\delta_p$, and the indicator function of a set $X$ by $\mathbbm{1}_X$:

$$\delta_p[x] = \begin{cases} 1, & x = p \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \mathbbm{1}_X[x] := \begin{cases} 1, & x \in X \\ 0, & \text{otherwise} \end{cases}.$$

We denote the support of a function $f$ by $\text{supp}(f)$. By the “support” of a function, we mean the largest set on which the function could, in principle, be non-zero. If the function $f$ happens to be zero at point $z$ for a particular problem, but could be non-zero at $z$ for other problems, we still say $z \in \text{supp}(f)$. We call a function of $N$ nearly linear if it scales as $O(N \log^a N)$ for $N \to \infty$, where $a$ is some small non-negative integer (say $a \in \{0, 1, 2\}$).

2. The adaptive product-convolution approximation. As discussed in Section 1, if $A$ were translation-invariant (i.e., if (1) held with equality for all $x, y \in \mathbb{Z}^d$), then $A$ would be the convolution operator defined by the action $Af = \varphi_p * f$, where $\varphi_p$ is the impulse response of $A$ at $p$, as defined in (2). For example, the solution operator for a homogeneous PDE on an unbounded domain is translation-invariant, and $\varphi_p$ is the Green’s function for the PDE. Of course, translation-invariant operators are rare in practice. It is more common for $A$ to only be approximately translation-invariant (see Figure 1), and for the approximate translation-invariance to be valid only locally. That is,

$$(9) \quad A[p + y - x, p] \approx A[y, x] \quad \text{when } x \in U$$

for some neighborhood $U$ consisting of points “near” $p$. We will provide a rigorous analysis of approximation errors in Section 4; for now we leave the exact nature of this approximate equality ($\approx$) intentionally vague. Just as translation-invariance of $A$ implies that $A$ is a convolution operator, local approximate translation-invariance of $A$ implies that $A$ can be locally approximated by a convolution operator. Specifically, it is straightforward to show that (9) implies

$$(10) \quad Ag \approx \varphi_p * g \quad \text{when } \text{supp}(g) \subset U.$$
In order to approximate the action of $A$ on functions $f$ supported on a larger region of interest, we patch together local convolution operator approximations. Let $\{U_k\}_{k=1}^r$ be a collection of sets covering $\text{supp}(f)$, let $\{w_k\}_{k=1}^r$ be a partition of unity subordinate to this cover, let $p_k \in U_k$ for $k = 1, \ldots, r$, and define $\varphi_k := \varphi_{p_k}$. If the following local approximations hold:

\[
Ag \approx \varphi_k \ast g \quad \text{when supp}(g) \subset U_k, \quad k = 1, \ldots, r,
\]

then $A$ can be globally approximated as follows:

\[
Af = A \sum_{k=1}^r w_k \cdot f
\]

\[
= \sum_{k=1}^r A(w_k \cdot f)
\]

\[
\approx \sum_{k=1}^r \varphi_k \ast (w_k \cdot f).
\]  

(12)

The first line follows from the partition unity property of the functions $w_k$, the second line follows from the linearity of $A$, and the third line follows from the local approximation property (11) and the fact that $\text{supp}(w_k \cdot f) \subset U_k$.

In the preceding discussion we implicitly assumed that $A$ is defined on $\mathbb{Z}^d$, so that the functions $\varphi_k$ are defined everywhere they need to be for the convolutions in (12) to make sense. Extending functions by zero as needed makes these convolutions well-defined for problems defined on bounded domains. However, this leads to boundary artifacts wherever zeros are used in place of undefined impulse response data. These boundary artifacts are purely a side effect of the scheme and are unrelated to real boundary effects present in the underlying operator $A$; they occur even in the case where $A$ is, itself, a convolution operator (see Figure 2). To eliminate such boundary artifacts, we extend the functions $\varphi_k$ outside of their natural support by using information from neighboring functions $\varphi_j$ to create “extended” impulse response functions $\varphi_k^E$. Section 2.6 will detail this extension process.
2.1. Product-convolution format. The derivation in the previous section (Section 2) leads us to approximate $A$ with the following product-convolution approximation:

$$\tilde{A}f := \sum_{k=1}^{r} \phi_k^E \ast (w_k \cdot f),$$

where

- $\{\phi_k^E\}_{k=1}^{r}$ are modified versions of the impulse responses

$$\varphi_k[z] = (A\delta_{p_k})[z + p_k], \quad z \in \Omega - p_k,$$

for a collection of sample points $\{p_k\}_{k=1}^{r}$.
- The sample points $\{p_k\}_{k=1}^{r}$ reside in a collection of overlapping sets $\{U_k\}_{k=1}^{r}$ that cover $\Omega$:

$$p_k \in U_k \quad \text{for} \quad k = 1, \ldots, r \quad \text{and} \quad \Omega \subset \bigcup_{k=1}^{r} U_k.$$

- $\{w_k\}_{k=1}^{r}$ is a partition of unity subordinate to the cover:

$$\text{supp}(w_k) \subset U_k \quad \text{for} \quad k = 1, \ldots, r \quad \text{and} \quad \sum_{k=1}^{r} w_k[x] = 1 \quad \text{for all} \quad x \in \Omega.$$

Our scheme is defined by the points $p_k$, the sets $U_k$, the partition of unity weighting functions $w_k$, and the extended impulse response functions $\varphi_k^E$. We construct the points $p_k$ and sets $U_k$ in an adaptively refined grid, for reasons that will be discussed in Section 2.2. We provide a detailed description of how we construct the adaptive grid and the functions $w_k$ and $\varphi_k^E$ in Sections 2.3, 2.4, 2.5, 2.6, 2.7, and 2.8, and then summarize the cost of constructing the approximation in Section 2.9.

2.2. Motivation for adaptivity. In general, translation-invariance varies spatially. By this, we mean that the size of the neighborhood $U$ on which the error in (9) is sufficiently small depends on the location of $U$. To fix ideas, suppose that $A$ is the solution operator for an inhomogeneous elliptic PDE. In this case, the size of $U$ will typically be small if the coefficient in the PDE varies over short length scales within $U$, and large if the coefficient varies over large length scales within $U$. In order to capture such spatial variations in translation-invariance while minimizing the number of sample points used, we choose $p_k$ and $U_k$ adaptively (see Sections 2.3 and 2.4). A randomized adjoint based a-posteriori error estimator, described in Section 2.7, drives the adaptivity.

Due to boundary effects, translation-invariance typically fails in directions perpendicular to a boundary, but holds in directions parallel to that boundary. For example, let $\varphi_p$ be the Green’s function at $p$ for a homogeneous PDE on an infinite half-space. Although $\varphi_p$ changes as $p$ approaches the boundary, by symmetry it does not change as $p$ moves parallel to the boundary. In order to address this direction-dependent translation-invariance, we refine anisotropically, subdividing preferentially in directions that $\varphi_p$ changes the most as a function of $p$ (see Section 2.8).

The adaptive refinement procedure creates unusually shaped neighborhoods $U_k$. We construct harmonic weighting functions, $w_k$, on these sets by solving local Laplace problems, as described in Section 2.5.
2.3. Adaptive grid structure. We will choose the sample points, \( p_k \), so that they form an adaptively refined rectilinear grid (for example, see Figure 3). This section defines the structure of the adaptive grid; the procedure for constructing it will be explained in Section 2.4.

We organize the domain \( \Omega \) into a binary tree, \( T \), of boxes \( C \subset \Omega \) which we call cells. The root of \( T \) is the whole domain \( \Omega \). Cells may be either refined or not refined; refined cells are internal nodes in \( T \) and unrefined cells are leaves of \( T \). We denote the set of all leaves of the tree by \( \text{leaves}(T) \). Refined cells \( C \) are subdivided in a chosen direction into a set of two child cells that share an internal facet (more about how we choose the subdivision direction in Section 2.8). We denote the set of children of \( C \) by \( \text{children}(C) \). Specifically, if \( C \) is subdivided in the \( j \)th direction, then \( \text{children}(C) = \{ A, B \} \) where

\[
a_{\min} = c_{\min}, \quad a_{\max}^i = \begin{cases} c_{\max}^i, & i \neq j, \\ c_{\mid}^i, & i = j. \end{cases}
\]

and

\[
b_{\min}^j = \begin{cases} c_{\min}^j, & i \neq j, \\ c_{\mid}^i, & i = j. \end{cases}, \quad b_{\max} = c_{\max}.
\]

The corners of all cells form the set of sample points:

\[
\{ p_k \}_{k=1}^r = \bigcup_{C \in T} \text{corners}(C).
\]

Since the cells share facets, typically more than one cell contains a given sample point. We write

\[
\text{cells}(p_k) := \{ C : C \in \text{leaves}(T), p_k \in C \}
\]

to denote the set of all leaf cells containing \( p_k \). We define the blocky neighborhood, \( U_k \), associated with a sample point \( p_k \) as the union of all leaf cells containing \( p_k \):

\[
U_k := \bigcup_{C_i \in \text{cells}(p_k)} C_i.
\]
Sample points $p_k$ and $p_j$ are neighbors if they share a common leaf cell. That is, there exists a leaf cell $C$ such that $p_k \in C$ and $p_j \in C$. Note that under this definition $p_k$ is neighbors with itself. We write $\text{nbrs}(k) \subset \{1, \ldots, r\}$ to denote the set of indices of sample points that are neighbors of $p_k$, including $p_k$ itself. In other words, $j \in \text{nbrs}(k)$ if $p_k$ and $p_j$ are neighbors.

2.4. Adaptive refinement algorithm. Starting with $\Omega$ subdivided once in all directions, we repeatedly estimate the error in all cells in leaves ($T$) using an a-posteriori error estimator, then refine the leaf cell with the largest error. The refinement process continues until either (a) the desired error in the approximation is achieved, or (b) a predetermined maximum number of sample points $p_k$ is reached. At each step of the refinement process we construct or modify the functions $w_k$ and $\varphi E_k$ using methods that will be described in Sections 2.5, 2.6, and 2.9. We perform the a-posteriori error estimation with a randomized method that will be described in Section 2.7. We choose which direction to subdivide cells in using a method that will be described in Section 2.8.

2.5. Harmonic weighting functions. We construct harmonic partition of unity weighting functions, $w_k$, by solving discrete local Laplace (diffusion) problems recursively on subsets of $U_k$. This process is equivalent to the construction of harmonic basis functions in finite element methods [13], and also shares conceptual ties with partition of unity finite element methods [6] and the construction of coarse basis functions in agglomerated element algebraic multigrid [51].

The blocky neighborhood $U_k$ is a union of $d$-dimensional boxes. The boundary of each $d$-dimensional box is a union of $(d-1)$-dimensional facets, each of which is a box. There are $2d$ facets, corresponding to either the front or the back of the box in each coordinate direction. Facets that contain hanging nodes (“broken facets”) are the union of several smaller $(d-1)$-dimensional boxes. Hence the boundary of each $d$-dimensional box can be expressed as the union of $(d-1)$-dimensional boxes, where we exclude broken facets in favor of their constituent smaller boxes. In the same way, the boundary of each $(d-1)$-dimensional box is a union of $(d-2)$-dimensional boxes, the boundary of each $(d-2)$-dimensional box is a union of $(d-3)$-dimensional boxes, and so forth. This continues all the way down until we reach a set of 0-dimensional sample points. We build harmonic weighting functions by solving the Laplace equation $(-\Delta w_k = 0)$ on these boxes recursively in dimension, using the values from lower-dimensional boxes as Dirichlet boundary conditions for higher-dimensional boxes. For
sample points $p_j$ (the lowest level), we assign $w_k$ the values:

$$w_k[p_j] = \begin{cases} 
1, & k = j, \\
0, & k \neq j.
\end{cases}$$

Figure 4 illustrates this process for $d = 2$.

For the discrete Laplace equation we use the (positive definite) discrete graph Laplacian; this is equivalent to discretizing the continuous Laplacian using a standard Kronecker sum finite difference approximation on a regular grid. The local Laplace problems can be solved efficiently (in time proportional to the number of unknowns) with multigrid [7, 16].

**Proposition 1.** The functions $w_k$ form a partition of unity on $\Omega$.

**Proof.** Linearity and a discrete maximum principle [28] imply that a collection of discrete harmonic functions (as discretized here) form a partition of unity on a box if they form a partition of unity on the boundary of the box. Hence the proposition follows from induction on boxes of increasing dimension, starting from the base case that $\sum_k w_k$ forms a partition of unity on the sample points $p_j$. \hfill \Box

### 2.6. Extended impulse response functions.

To construct $\varphi^E_k$, we first compute the impulse responses $\varphi_k$ of $A$ at the points $p_k$ by applying $A$ to delta functions, then translating the results (see (14)). To eliminate boundary artifacts, we create $\varphi^E_k$ by extending the support of $\varphi_k$, using data from neighboring functions $\varphi_j$ to fill in regions outside of $\text{supp}(\varphi_k)$.

1. For $z$ within $\text{supp}(\varphi_k)$, we set $\varphi^E_k[z] := \varphi_k[z]$.
2. For $z$ outside $\text{supp}(\varphi_k)$ but within $\text{supp}(\varphi_j)$ for at least one neighboring $\varphi_j$, we define $\varphi^E_k[z]$ as the average of all neighboring $\varphi_j[z]$ whose support contains $z$.
3. For $z$ outside $\text{supp}(\varphi_k)$ and outside $\text{supp}(\varphi_j)$ for all neighboring $\varphi_j$, we set $\varphi^E_k[z] := 0$.

Figure 5 illustrates this procedure for a 1-dimensional example. Our theory still holds if we use any weighted average of neighboring $\varphi_j[z]$ in Step 2, provided the weights are non-negative and sum to one. We use the average since it simplifies the implementation and the explanation, and since more elaborate schemes are likely to yield only minimal improvements. The fact that we set some entries of $\varphi^E_k[z]$ to zero in Step 3 is irrelevant since our scheme never accesses these entries (this will follow from Proposition 5).

In preparation for the theory we will present in Section 4.1, we now describe the
process of constructing $\varphi^E_k$ more precisely. First, we construct the following counting functions:

$$c_k := 1_{\Omega-p_k} + \sum_{j \in \text{nbrs}(k), j \neq k} 1_{(\Omega-p_j) \setminus (\Omega-p_k)}.$$ 

Since $\text{supp}(\varphi_j) = \Omega - p_j$, $c_k [z]$ counts how many $\varphi_j$ will contribute to $\varphi^E_k [z]$. Next we compute

$$v_k [z] := \begin{cases} 1/c_k [z], & z \in \text{supp}(c_k) \\ 0, & \text{otherwise} \end{cases}$$

and define

$$v^{(j)}_k := \begin{cases} v_k \cdot 1_{\Omega-p_k}, & j = k \\ v_k \cdot 1_{(\Omega-p_j) \setminus (\Omega-p_k)}, & j \in \text{nbrs}(k), j \neq k, \end{cases}$$

The function $v^{(j)}_k [z]$ is the weight given to neighboring impulse response $\varphi_j$ at point $z$ when constructing $\varphi^E_k$. Finally, we construct $\varphi^E_k$:

$$\varphi^E_k := \sum_{j \in \text{nbrs}(k)} v^{(j)}_k \cdot \varphi_j.$$ 

### 2.7. Randomized a-posteriori error estimator.

In order to decide which cells to refine, we wish to compute the Frobenius norm of the error in the approximation,

$$e_C := \left\| \left( \tilde{A} - A \right) \left[ \Omega, C \right] \right\|,$$

for all cells $C \in \text{leaves}(T)$. Computing these norms is prohibitively expensive, so instead we estimate them with a randomized estimator. If $M$ is any matrix with $N$ columns, then the following sample average approximation estimates the square of its Frobenius norm:

$$||M||^2 = \mathbb{E} \left( ||M\zeta||^2 \right) \approx \frac{1}{q} \sum_{i=1}^{q} ||M\zeta_i||^2 = \frac{1}{q} ||MZ||^2,$$

where $\zeta, \zeta_i \sim N(0,1)^N$, are independent and identically distributed (i.i.d.) Gaussian random vectors, $\mathbb{E}$ is the expected value, $q$ is the number of samples used in the sample average approximation, and $Z \sim N(0,1)^{N \times q}$ is an i.i.d Gaussian random matrix (the matrix with columns $\zeta_i$) [5].

Hence we can form an estimator, $\eta_C \approx e_C$, by forming a random matrix $Z \sim N(0,1)^{N \times q}$, computing

$$Y = A^* Z \quad \text{and} \quad \tilde{Y} = \tilde{A}^* Z,$$

then extracting blocks of the results, and taking norms:

$$\eta_C := \frac{1}{\sqrt{q}} \|\tilde{Y} [C, \cdot] - Y [C, \cdot]\|.$$ 

By performing the randomized sample average approximation with the adjoints $A^*$ and $\tilde{A}^*$, we apply these operators once per sample, then post process the results to get estimators for all cells. Using the original operators $A$ and $\tilde{A}$ instead would require us to apply these operators to new random vectors for each cell.
2.8. Anisotropic refinement: choosing the subdivision direction. We refine anisotropically by estimating the direction that $\varphi_p$ changes the most as a function of $p$, then subdividing in that direction. This allows us to capture changes to $\varphi_p$ in directions where translation-invariance fails, without refining the grid in directions where translation-invariance holds.

Let $C$ be a cell that we have chosen to subdivide based on the randomized a-posteriori error estimator described in Section 2.7. For each coordinate direction $i$ in which $C$ is big enough to be refined ($c_i^{\text{max}} - c_i^{\text{min}} > 2$), we partition the functions $\varphi_k^E$ associated with the corners of $C$ into two groups. One group is the set of $\varphi_k^E$ associated with corners in the “front” of the cell $(+)$ in the $i$th coordinate direction, and the other group is the set of $\varphi_k^E$ associated with the “back” of the cell $(-)$ in the $i$th coordinate direction:

$$
\Psi^{i+} := \{ \varphi_k^E : p_k \in \text{corners}(C), \ p_k^i = c_i^{\text{max}} \},
\Psi^{i-} := \{ \varphi_k^E : p_k \in \text{corners}(C), \ p_k^i = c_i^{\text{min}} \}.
$$

Next, we construct “average” $\varphi_k^E$ functions for the front and back of the cell, respectively:

$$
\varphi^{i+} := \frac{1}{2d-1} \sum_{\varphi_k^E \in \Psi^{i+}} \varphi_k^E \quad \text{and} \quad \varphi^{i-} := \frac{1}{2d-1} \sum_{\varphi_k^E \in \Psi^{i-}} \varphi_k^E.
$$

Then we compute how much these average impulse responses change from the front to the back in direction $i$:

$$
\tau_i := \| \varphi^{i+} - \varphi^{i-} \|_{L^2(\Omega - c_{\text{mid}})}.
$$

Finally, we subdivide $C$ in the coordinate direction $i$ for which $\tau_i$ is maximized.

2.9. Construction cost. Updating $\tilde{A}$ after refining a cell requires us to apply $A$ to delta functions centered at the new sample points created during the refinement. Hence the entire refinement process requires us to apply $A$ to $r$ vectors, where $r$ is the total number of sample points in the final product-convolution approximation.

The dominant cost in the error estimation process is the cost of computing $A^*Z$ and $\tilde{A}^*Z$ for a random $N \times q$ matrix $Z$. Since $A^*Z$ is constant throughout the refinement process, we compute it once at the beginning by computing $A^*\zeta_i$ for all $q$ columns, $\zeta_i$, of $Z$.

Although $\tilde{A}$ changes every time we refine a cell, after performing a refinement we do not have to recompute $\tilde{A}^*Z$ from scratch. To see this, recall from (6) that the adjoint of our product-convolution operator is a convolution-product operator with the convolution functions reflected about the origin and complex conjugated. In order to recompute $\tilde{A}^*Z$ after refining cells, we only need to compute the convolutions

$$
\text{flip} \left( \overline{\varphi_k^E} \right) * \zeta_i
$$

for each column, $\zeta_i$, in $Z$, and each sample point, $p_k$, that is new or has a new neighbor. The convolutions for old sample points without new neighbors have been computed previously and can be re-used within (6). Thus the error estimation process

---

6The function $\varphi_k^E$ depends on neighboring impulse responses due to the extension procedure.
requires computing $O(rq)$ convolutions. As we will discuss in Section 3.2, each of these convolutions can be done with the FFT in $O(N \log N)$ work.

Updating the functions $w_k$ can be done locally. This requires negligible work compared to the other costs already discussed.

Putting all these pieces together, constructing $\tilde{A}$ requires

$$O(rC + qC^* + rqN \log N),$$

work, where $C$ is the work required to apply $A$ to a vector, and $C^*$ is the work required to apply $A^*$ to a vector.

3. Using the product-convolution approximation. The product-convolution format, (13), allows us to perform the following useful operations with $\tilde{A}$:

1. We can compute individual matrix entries of $\tilde{A}$ in $O(1)$ work.
2. We can apply $\tilde{A}$ and $\tilde{A}^*$ to vectors in nearly linear work by using the FFT to compute convolutions.
3. We can apply blocks of $\tilde{A}$ and $\tilde{A}^*$ to vectors in work that is nearly linear in the dimensions of the block.
4. We can efficiently convert $\tilde{A}$ into $H$-matrix format.

Once in $H$-matrix format, $\tilde{A}$ can be inverted, factorized, multiplied with other $H$-matrices, and so on, in nearly linear time using $H$-matrix arithmetic. In particular, we can create preconditioners for $A$ by constructing $\tilde{A}$ using a moderate or large tolerance for error, then converting $\tilde{A}$ into $H$-matrix format, then inverting or factorizing the $H$-matrix.

3.1. Matrix entries of $\tilde{A}$. Our approximate operator $\tilde{A}$ is a product-convolution scheme and therefore (as seen in (5)) has the following matrix entries:

$$\tilde{A}[y,x] = \sum_{k=1}^{r} w_k[x] \varphi^E_k[y-x].$$

Using (20) we can compute individual matrix entries of $\tilde{A}$ in $O(1)$ time even though $\tilde{A}$ is not stored in memory in the conventional sense.

3.2. Applying $\tilde{A}$ or $\tilde{A}^*$ to vectors. Applying $\tilde{A}$ or $\tilde{A}^*$ to a vector requires computing $r$ convolutions, $r$ pointwise vector multiplications, and some vector additions (see equations (13) or (6), respectively). Out of these operations, the $r$ convolutions are the most computationally expensive. Since the convolution theorem allows us to compute each of these convolutions using the FFT (after appropriate zero padding) at $O(N \log N)$ cost, the cost of applying $\tilde{A}$ or $\tilde{A}^*$ to a vector is $O(rN \log N)$. See [47] for details.

3.3. Applying blocks of $\tilde{A}$ or $\tilde{A}^*$ to vectors. Proposition 2 (below) illustrates how to efficiently apply a block of a convolution operator to a vector. That is, to implicitly apply a convolution to a function that is supported in a source box $S$, then restrict the results to another target box $T$, without ever performing operations on the full space (see Figure 6). If the row and column degrees of freedom of the block do not correspond to boxes, extend functions by zero to enclosing boxes as needed.
us to apply a block of a convolution operator to a vector in work that scales nearly linearly with the linear dimensions of the block: $O(\sigma \log \sigma)$ where $\sigma = |S| + |T|$. To apply $\tilde{A}[T,S]$ or $\tilde{A}^+[T,S]$ to a vector, we use this method for each convolution in the sums ((13) and (6)) defining $\tilde{A}$ or $\tilde{A}^+$, respectively, that could be non-zero. Since the functions $w_k$ are supported on the sets $U_k$, the terms in these sums that could be non-zero correspond to sets $U_k$ that intersect $S$ when multiplying $\tilde{A}[T,S]$ with a vector, and $T$ when multiplying with $\tilde{A}^+[T,S]$ with a vector. As a result, it costs

$$O(r_S \sigma \log \sigma) \quad \text{and} \quad O(r_T \sigma \log \sigma)$$

work to apply $\tilde{A}[T,S]$ and $\tilde{A}^+[T,S]$ to vectors, respectively. Here $r_S$ and $r_T$ are the number of sets $U_k$ that intersect $S$ and $T$, respectively.

**Proposition 2 (Applying convolution operator block).** Let $S,T \subset \mathbb{Z}^d$ be boxes, and let $f, \varphi$ be scalar valued functions on $\mathbb{Z}^d$, with $\text{supp}(f) \subset S$. Further, let $G := T - S$, and define the localized boxes

$$S_0 := S - s_{\min}, \quad T' := T - t_{\min} + s_{\max} - s_{\min}, \quad G_0 := G - g_{\min}.$$ 

and the locally supported functions

$$f_0[x] := \begin{cases} f[x_0 + s_{\min}], & x_0 \in S_0, \\ 0, & \text{else}, \end{cases} \quad \varphi_0[z] := \begin{cases} \varphi[z_0 + g_{\min}], & z_0 \in G_0, \\ 0, & \text{else}, \end{cases}$$

(see Figure 6). Then

$$(\varphi \ast f)[T] = (\varphi_0 \ast f_0)[T'].$$

**Proof.** The result follows from expanding $\varphi \ast f$ using the definition of a convolution, then making the change of variables $x_0 := x - s_{\min}$ and $y' := y - s_{\min} - g_{\min}$. \(\square\)

### 3.4. Conversion to hierarchical matrix format

Construction of a hierarchical matrix proceeds in the following steps:

1. The degrees of freedom are partitioned hierarchically into a cluster tree.
2. The matrix entries are partitioned hierarchically into a block cluster tree.
3. A low-rank approximation is constructed for each block of the matrix that is marked as low-rank (i.e., admissible) within the block cluster tree. Figure 7 illustrates this for a simple HODLR\(^8\) $H$-matrix.

The $H$-matrix construction process is scalable if we can construct low-rank approximations (see Section 1.3.1) of the low-rank blocks (Step 3) in work that scales nearly linearly with the dimensions of the block. The method for efficiently applying blocks of $\tilde{A}$ and $\tilde{A}^*$ to vectors, outlined in Section 3.3, allows us to do this using Krylov methods or randomized SVD. Whenever the Krylov method or randomized SVD requires the application of a block or its adjoint to a vector, we perform this computation using the method in Section 3.3. Alternatively, formula (20) for the matrix entries of $\tilde{A}$ allows us to construct a low-rank approximation of a block by forming a CUR approximation. Whenever the CUR approximation algorithm requires a row, column, or individual entry of the block, we access it using (20).

Since applying the block $\tilde{A}[T,S]$ to a vector costs $O(\sigma \log \sigma)$ work, where $\sigma = |S| + |T|$, whereas accessing a row or column costs $O(\sigma)$ work, the CUR approach is asymptotically more scalable than the Krylov or randomized SVD approaches by a log factor. However, the CUR approach is less robust, and typically has poorer dependence on the rank of the blocks. In either case the overall cost of constructing the $H$-matrix scales nearly linearly with $N$. Moreover, the construction process only uses the approximation, $\tilde{A}$. It does not require expensive application of the original operator, $A$.

4. Theory. Here we provide a theoretical foundation for the product-convolution operator approximation scheme we defined in Section 2.

In Section 4.1 we show that the error in $\tilde{A}$ is controlled by the failure of $A$ to be locally translation-invariant with respect to a locally expanded version of the cover $\{U_k\}_{k=1}^L$ (Theorem 6). This provides an a-priori error estimate for the approximation, and shows that the approximation will not contain boundary artifacts.

In Section 4.2 we present an upper bound (Theorem 7) on the number of samples required for our randomized adjoint a-posteriori error estimator in Section 2.7. The bound is adapted from [5]. Given a desired accuracy and success rate, the bound on the number of samples is independent of any properties of $A$. Theorem 7 is primarily of theoretical interest; in practice we find that our scheme achieves good performance

---

\(^8\)We show a HODLR $H$-matrix for illustrative purposes; the content of this section applies to any $H$-matrix structure.
with far fewer samples than Theorem 7 would suggest.

**4.1. A-priori error analysis.** Let $F_k$ be the following functions that measure the failure of $A$ to be translation-invariant near the points $p_k$ (see Figure 1)

\[
F_k[y,x] := A[y - x + p_k, p_k] - A[y,x].
\]

Let $U_k^E$ be the extension of $U_k$ created by unioning $U_k$ with its neighbors:

\[
U_k^E := \bigcup_{j \in \text{nbrs}(k)} U_j.
\]

We define $\mu_k^E \subset \Omega \times \Omega$ as the set of all $(y,x)$ such that:

- $x$ is “near” $p_k$ (in the sense that $x \in U_k^E$).
- $F_k[y,x]$ is well-defined without resorting to extension by zero.

Specifically,

\[
\mu_k^E := \{(y,x) : x \in U_k^E, y \in \Omega, y - x + p_k \in \Omega\}.
\]

In Theorem 6 we will show that if $F_k$ is small on $\mu_k^E$ for all $k$, then the error in our product-convolution approximation is correspondingly small.

A multi-step path leads to Theorem 6. In Proposition 5 we show that our operator approximation can be reinterpreted as a weighted sum involving the original (not extended) impulse response functions $\varphi_k$, but with weighting functions that form a partition of unity on the product space $\Omega \times \Omega$, and are supported in the sets $\mu_k^E$.

Proposition 5 relies on a lemma about the functions $v^{(j)}_k$ used in our impulse response extension procedure (Lemma 4), which in turn relies on a lemma about Minkowski sums of boxes (Lemma 3). After establishing these prerequisites, we use Proposition 5 and a convexity argument to prove Theorem 6.

**Lemma 3.** *If $S$ and $T$ are boxes, and $S$ is at least as large as $T$ in the sense that*

\[
s^i_{\max} - s^i_{\min} \geq t^i_{\max} - t^i_{\min}
\]

*for $i = 1, \ldots, d$, then $S + T = S + \text{corners}(T)$.*

*Proof. Since boxes are Cartesian products of intervals, and since the Cartesian product distributes over Minkowski addition, the result in $d$ dimensions follows from the result in one dimension.*

**Lemma 4.** *We have*

\[
\sum_{j \in \text{nbrs}(k)} v^{(j)}_k[z] = \begin{cases} 
1, & z \in \Omega - U_k, \\
0, & \text{otherwise}.
\end{cases}
\]

*Proof. By construction,*

\[
\sum_{j \in \text{nbrs}(k)} v^{(j)}_k[z] = \begin{cases} 
1, & z \in \text{supp}(c_k), \\
0, & \text{otherwise},
\end{cases}
\]

*and*

\[
\text{supp}(c_k) = \bigcup_{j \in \text{nbrs}(k)} (\Omega - p_j).
\]
All that remains to be shown is that $\Omega - U_k = \bigcup_{j \in \text{nbrs}(k)} (\Omega - p_j)$. To that end, recall that $U_k$ is the union of leaf boxes $C_i$ that contain $p_k$. Thus

$$\Omega - U_k = \Omega - \bigcup_{C_i \in \text{cells}(p_k)} C_i = \bigcup_{C_i \in \text{cells}(p_k)} (\Omega - C_i).$$

Since $C_i \subset \Omega$, we see that $\Omega$ is at least as large as $-C_i$ (in the sense of Lemma 3). Applying Lemma 3 to $\Omega - C_i$ and performing algebraic manipulations yields:

$$\bigcup_{C_i \in \text{cells}(p_k)} (\Omega - C_i) = \bigcup_{C_i \in \text{cells}(p_k)} (\Omega - \text{corners}(C_i)) = \Omega - \bigcup_{C_i \in \text{cells}(p_k)} \text{corners}(C_i).$$

Furthermore, by definition the union of all corners of leaf cells containing a point is the union of all neighboring points, so we have

$$\Omega - \bigcup_{C_i \in \text{cells}(p_k)} \text{corners}(C_i) = \Omega - \bigcup_{j \in \text{nbrs}(k)} p_j = \bigcup_{j \in \text{nbrs}(k)} (\Omega - p_j).$$

Overall this chain of set equalities implies

$$\Omega - U_k = \bigcup_{j \in \text{nbrs}(k)} (\Omega - p_j),$$

as required. \qed

**Proposition 5.** Let

$$(26) \quad W_k[y, x] := \sum_{j \in \text{nbrs}(k)} w_j[x] v_j^{(k)}[y - x].$$

1. The entries of $\tilde{A}$ can be written as:

$$\tilde{A}[y, x] = \sum_{k=1}^{r} W_k[y, x] \varphi_k[y - x].$$

2. The weighting functions $\{W_k\}_{k=1}^{r}$ form a partition of unity:

$$\sum_{k=1}^{r} W_k[y, x] = 1 \quad \text{for all } (y, x) \in \Omega \times \Omega.$$

3. The partition of unity is subordinate to the cover $\{\mu^E_k\}_{k=1}^{r}$:

$$\text{supp}(W_k) \subset \mu^E_k.$$

**Proof.**
1. Substituting the definition of $\varphi^E_k$ from (16) into the definition of $\tilde{A}$ from (13) then performing algebraic manipulations, we have:

\[
\tilde{A}[y,x] = \sum_{k=1}^{r} w_k[x] \sum_{j \in \text{nbrs}(k)} v_j^{(j)}[y-x] \varphi_j[y-x]
\]

\[
= \sum_{k=1}^{r} \sum_{j \in \text{nbrs}(k)} w_k[x] v_j^{(j)}[y-x] \varphi_j[y-x]
\]

\[
= \sum_{j=1}^{r} \sum_{k \in \text{nbrs}(j)} w_k[x] v_j^{(j)}[y-x] \varphi_j[y-x]
\]

\[
= \sum_{j=1}^{r} W_j[y,x] \varphi_j[y-x].
\]

In the third line we used the fact that

\[
\sum_{a \in X} \sum_{\{b : b \sim a\}} f(a,b) = \sum_{b \in X} \sum_{\{a : a \sim b\}} f(a,b)
\]

for any symmetric relation $\sim$. Note the switch of $k$ and $j$.

2. Using the definition of $W_k$ in (26), we have

\[
\sum_{k=1}^{r} W_k[y,x] = \sum_{k=1}^{r} \sum_{j \in \text{nbrs}(k)} w_j[x] v_j^{(k)}[y-x]
\]

\[
= \sum_{j=1}^{r} \sum_{k \in \text{nbrs}(j)} w_j[x] v_j^{(k)}[y-x]
\]

\[
= \sum_{j=1}^{r} w_j[x] \left( \sum_{k \in \text{nbrs}(j)} v_j^{(k)}[y-x] \right). 
\]

If $x \in U_j$ and $y \in \Omega$, then Minkowski set arithmetic implies $y - x \in \Omega - U_j$, so (25) in Lemma 4 implies

\[
\sum_{k \in \text{nbrs}(j)} v_j^{(k)}[y-x] = 1.
\]

Since $\text{supp}(w_j) \subset U_j$, this implies

\[
\sum_{j=1}^{r} w_j[x] \left( \sum_{k \in \text{nbrs}(j)} v_j^{(k)}[y-x] \right) = \sum_{j=1}^{r} w_j[x] = 1.
\]

Thus $\sum_{k=1}^{r} W_k[y,x] = 1$ as required.

3. From the definition of $v_j^{(j)}$ in (15), either $\text{supp}(v_j^{(j)}) = (\Omega - p_j) \setminus (\Omega - p_k)$ when $k \neq j$, or $\text{supp}(v_j^{(j)}) = \Omega - p_j$ when $k = j$. In either case $\text{supp}(v_k^{(j)}) \subset \Omega - p_j$.

Thus

\[
(y-x \notin \Omega - p_j) \implies (v_k^{(j)}[y-x] = 0),
\]
which is equivalent to the statement

\[(y - x + p_j \notin \Omega) \implies \left(v_k^{(j)}[y - x] = 0\right).\]

Since \(W_k\) consists of a sum of terms, each term containing \(v_j^{(k)}[y - x]\), statement (27) implies (note the swap of \(k, j\)):

\[(y - x + p_k \notin \Omega) \implies (W_k[y, x] = 0).\]

Additionally, since each \(w_j\) in the sum defining \(W_k\) is supported in the blocky neighborhood \(U_j\), and since the union of these blocky neighborhoods \(U_j\) is \(U^{E}_k\), we have

\[(x \notin U^{E}_k) \implies (W_k[y, x] = 0).\]

Altogether, (28), (29), and the definition of \(\mu^{E}_k\) in (23) imply \(\text{supp}(W_k) \subset \mu^{E}_k\).

Theorem 6. Let \(F_k\) and \(\mu^{E}_k\) be defined as in (22) and (23), respectively. If

\[\|F_k\|_{l^2(\mu^{E}_k)} \leq \gamma \|A\|_{l^2(\mu^{E}_k)}\]

then

\[\|\tilde{A} - A\| \leq \gamma \sqrt{n_{\text{overlap}}}\|A\|,\]

where \(n_{\text{overlap}}\) is the maximum number of sets in the cover \(\{\mu^{E}_k\}\) that are non-zero for any particular \((y, x)\).

Proof. From Proposition 5 and the fact that \(\varphi_k[z] = A[z + p_k, p_k]\), we know that

\[\tilde{A}[y, x] = \sum_{k=1}^{r} W_k[y, x] A[y - x + p_k, p_k],\]

Hence the pointwise error in the approximation takes the following form:

\[\tilde{A}[y, x] - A[y, x] = \sum_{k=1}^{r} W_k[y, x] A[y - x + p_k, p_k] - A[y, x]\]

\[= \sum_{k=1}^{r} W_k[y, x] (A[y - x + p_k, p_k] - A[y, x])\]

\[= \sum_{k=1}^{r} W_k[y, x] F_k[y, x].\]

In the second line we used the partition of unity property of \(W_k\) from Proposition 5.
Thus we can bound the error in the approximation as follows:

\[
\| \tilde{A} - A \|^2 = \sum_{(y,x) \in \Omega \times \Omega} \left( \sum_{k=1}^{r} W_k[y,x] F_k[y,x] \right)^2 \\
\leq \sum_{(y,x) \in \Omega \times \Omega} \sum_{k=1}^{r} W_k[y,x] F_k[y,x]^2 \\
= \sum_{k=1}^{r} \sum_{(y,x) \in \mu E_k} W_k[y,x] F_k[y,x]^2 \\
\leq \sum_{k=1}^{r} \| F_k \|^2_{\mathcal{L}(\mu E_k)} \\
\leq \gamma \sum_{k=1}^{r} \| A \|^2_{\mathcal{L}(\mu E_k)} \\
\leq \gamma n_{\text{overlap}} \| A \|^2.
\]

In the first inequality we used Jensen’s inequality, in the second inequality we used the fact that \( W[y,x] \leq 1 \), and in the third inequality we used the assumption of the theorem.

### 4.2. Analysis of the a-posteriori error estimator

Theorem 7 adapts a randomized trace estimator Chernoff bound from [5] to bound on the number of samples required for our randomized a-posteriori error estimator.

**Theorem 7.** Let \( e_C \) and \( \eta_C \) be the error in cell \( C \) and error estimator for cell \( C \), respectively, as defined in Section 2.7. Also recall that \( \eta_C \) has been constructed using \( q \) random samples. If

\[
q \geq 20 \varepsilon^{-2} \log \left( \frac{2}{\delta} \right)
\]

and \( \varepsilon \leq 0.1 \), then

\[
P( |\eta_C - e_C| \geq \varepsilon e_C ) \leq \delta.
\]

**Proof.** Theorem 10 from [5] implies \( P \left( |\eta_C^2 - e_C^2| \geq \varepsilon e_C^2 \right) \leq \delta \). The desired result (without the squares) follows from the fact that \( |\eta_C - e_C| \geq \varepsilon e_C |\eta_C - e_C| \).

### 5. Numerical examples

We numerically test our scheme on the non-local component of the Schur complement associated with restricting the Poisson operator to a boundary (Section 5.1), and the data misfit Hessian for an advection-diffusion inverse problem (Section 5.2). For the Schur complement, our scheme is mesh scalable: it requires roughly the same convolution rank (number of terms in (13)) to achieve a desired error tolerance regardless of how fine the mesh is. For the Hessian, our scheme is data scalable: it requires roughly the same convolution rank to achieve a desired error tolerance regardless of how informative the data are about the unknown parameter in the inverse problem. In both examples, our product-convolution scheme (“Conv”) outperforms optimal low-rank approximation via the truncated singular value decomposition (“TSVD”). Additionally, the randomized a-posteriori error estimator achieves good performance with only a handful of random samples: our scheme performs almost as well with \( q = 5 \) as it does with \( q = 100 \).
**5.1. Poisson Schur complement.**

**Problem setup.** Let $K \approx -\Delta$ be the discretized (negative) Laplace operator on the unit cube, $[0,1]^3$, with Neumann boundary conditions. For discretization we use piecewise linear finite elements defined on a regular rectilinear mesh of tetrahedra. Let ‘t’ denote the degrees of freedom on the top face of the unit cube, let ‘r’ denote remaining degrees of freedom that are not on the top face, and let $K_{tt}$, $K_{tr}$, $K_{rt}$, and $K_{rr}$ denote the associated blocks of $K$. We wish to approximate the following operator:

$$A := K_{tt}K_{rr}^{-1}K_{rt}.$$

Note that $-A$ is the non-local component of the Schur complement, $K_{tt} - K_{tr}K_{rr}^{-1}K_{rt}$, for the top degrees of freedom. Matrix entries of $A$ are not directly available; we apply $A$ to vectors by performing matrix-vector products with $K_{rt}$ and $K_{tr}$, and solving a linear system with $K_{rr}$ as the coefficient matrix.
Fig. 10: Poisson Schur complement: The (convolution) rank, $r$, required to achieve a relative approximation error of 20%, for a variety of mesh sizes, $h$.

Results. Figure 8 compares the convergence of our scheme to TSVD on a $40 \times 40 \times 40$ mesh (so that $N = 1600$). Since the Poisson Schur complement is high rank, TSVD performs poorly. In contrast, our scheme performs well: at $r = 200$ our scheme has less than 1% error, whereas TSVD has approximately 69% error. Figure 8 also shows that our scheme performs well even when we use a small number of random samples for the a-posteriori error estimator: the convergence curve for $q = 5$ is almost identical to the convergence curve for $q = 100$. Figure 9 displays the adaptive meshes from four different stages of the adaptive refinement process from Figure 8. Our scheme adaptively refines towards the boundary, then the corners. This is expected since boundary effects are the only source of translation-invariance failure.

Figure 10 compares our scheme to TSVD on a sequence of progressively finer meshes, from $10 \times 10 \times 10$ to $100 \times 100 \times 100$. That is, $h = 0.1$ to $h = 0.01$, where $h$ is the distance between adjacent gridpoints in the mesh. The curves show the (convolution) rank, $r$, required to achieve a relative error tolerance of 20%. The rank for TSVD grows with the number of degrees of freedom on the top surface ($r \sim O(1/h^2)$), offering little improvement over directly building a dense matrix representation of $A$ column-by-column. In contrast, the convolution rank for our scheme remains roughly constant as $h \to 0$.

5.2. Advection-diffusion inverse problem Hessian.

Problem setup. In this section we approximate the data misfit portion of the Hessian for an advection-diffusion inverse problem in which an unknown initial concentration, $m$, of a contaminant, $u$, is inferred from time series data, $y$, of the contaminant flowing through a boundary, $\Gamma$. Specifically, consider the following PDE:

\begin{equation}
\begin{aligned}
\frac{\partial u}{\partial t} &= \frac{1}{\text{Pe}} \Delta u - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \cdot \nabla u, \quad t \in [0, 1], \\
u &= m, \quad t = 0,
\end{aligned}
\end{equation}

where $\text{Pe}$ is the Peclet number. The region of interest and support of $m$ is the unit square, $\Omega = [0, 1]^2$, and the desired unbounded domain for the PDE is $\mathbb{R}^2$. To
simulate the effect of having an unbounded domain, we extend the computational domain beyond $[0,1]^2$ on all sides and use Neumann boundary conditions on the outer, larger, domain. We use $y$ to denote the known time series observations of $u$ on the top boundary: $y(x,t) = u(x,t)$, $x \in \Gamma$, $t \in (0,1]$, where $\Gamma := [0,1] \times \{1\}$.

The inverse problem is: given $y$, determine $m$. This is commonly formulated as a least squares optimization problem of the following form:

\begin{equation}
\min_m J(m) + R(m),
\end{equation}

where $J$ is the data misfit

$$J(m) := \frac{1}{2} \int_0^1 \int_\Gamma (u(m) - y)^2 \, dx \, dt,$$

and $R(m)$ is a quadratic regularization term. Here $u(m)$ denotes the solution of (31) as a function of $m$. Our goal is to approximate the discretized version of the operator

$$A := \frac{d^2 J}{dm^2},$$

which is the Hessian of the data misfit. For discretization we use piecewise linear finite elements defined on a regular rectilinear $100 \times 100$ mesh of triangles, with 100 time steps. We use backward Euler time stepping and SUPG stabilization [17].

The Hessian. Since $y$ depends linearly on $m$, the solution to (32) is the solution to a linear system with the Hessian of the objective function, $\frac{d^2}{dm^2}(J + R) = A + \frac{d^2 R}{dm^2}$, as the coefficient matrix. Newton methods for solving nonlinear advection-diffusion inverse problems require solving linear systems with similar Hessians as coefficient operators. The Hessian of the regularization, $\frac{d^2 R}{dm^2}$, is typically either a differential operator with known entries (smoothing regularization), or a mass matrix ($L^2$ regularization), and thus it is easy to manipulate. In contrast, $A$ is dense and its matrix entries are not directly available. We can only apply $A$ to vectors using an adjoint-based framework (see [2]). This requires solving a pair of advection-diffusion equations: a state equation of the form (31) forward in time, and the adjoint of (31) backward in time. Explicitly forming $A$ is thus prohibitively expensive: a pair of PDEs would need to be solved for every column of $A$. While Krylov methods can be used to solve these linear systems in a matrix-free manner, good general purpose preconditioners have not been available (see [3] for a discussion of these issues). Our scheme offers an avenue for building good preconditioners: first form a product-convolution approximation of $A$, then convert it to $H$-matrix format, then combine it with $\frac{d^2 R}{dm^2}$, and finally invert the combined $H$-matrix with fast $H$-matrix arithmetic.

Data scalability. The Peclet number, $Pe$, controls the ratio of advection to diffusion. As $Pe$ increases, the rank of $A$ increases [37], making the inverse problem more difficult to solve with existing methods. This increase in the rank corresponds to an increase in the informativeness of the data about the parameter in the inverse problem. Roughly speaking, eigenvectors of $A$ corresponding to large eigenvalues represent modes of the parameter that are well-informed by the data, whereas eigenvectors of $A$ corresponding to small eigenvalues represent modes of the parameter that are poorly-informed by the data (see [4] for a discussion of these issues). As a result, for an approximation of $A$ to be data-scalable (perform well regardless of how informative the data is about the parameter), the cost of constructing the approximation must not grow as $Pe$ increases.
Fig. 11: **Advection-diffusion inverse problem Hessian**: Relative error in the TSVD low-rank approximation and our product-convolution approximation as the (convolution) rank, $r$, changes. We show convergence curves for our scheme using both $q = 5$ and $q = 100$ random samples for the a-posteriori error estimator. Black dots correspond to the adaptive grids visualized in Figure 12.

Fig. 12: **Advection-diffusion inverse problem Hessian**: Intermediate stages of adaptive grid refinement corresponding to black dots in Figure 11.

**Results.** Figure 11 compares the convergence of our product-convolution scheme to TSVD when $Pe = 10^4$. Our scheme performs better than TSVD: at $r = 100$ our scheme has less than 1% error whereas TSVD has approximately 71% error. Like the Poisson problem, the convergence curve for $q = 5$ is almost identical to the convergence curve for $q = 100$. Figure 12 shows the adaptive meshes from four different stages of the adaptive refinement process from Figure 11. Our scheme chooses to adaptively refine in the direction of the vertical flow, prioritizing refinement near the top surface. We expect similar results would hold for inverse problems involving non-vertical, non-uniform flow if the convolution grid were aligned with the streamlines of the flow (Lagrangian parameterization).

Figure 13 compares our scheme to TSVD for a sequence of increasing Peclet numbers, from $Pe = 10^1$ to $Pe = 10^5$. The curves show the (convolution) rank, $r$, required to achieve a relative error tolerance of 10%. Whereas the required rank for TSVD grows dramatically as $Pe$ increases, the required convolution rank for our scheme remains roughly constant.
6. Conclusions. In this paper we presented a matrix-free adaptive grid product-convolution operator approximation scheme. The efficiency of our scheme depends on the degree to which the operator being approximated is locally translation-invariant. As a result, our scheme is well-suited for approximating or preconditioning operators that arise in Schur complement methods for solving partial differential equations (PDEs), reduced Hessians in PDE-constrained optimization and inverse problems, integral operators, covariance operators with spatially varying kernels, and Dirichlet-to-Neumann maps or other Poincaré–Steklov operators in multiphysics problems. These operators are often dense, implicitly defined, and high-rank, making them difficult to approximate with standard techniques. Our scheme is best suited to moderate accuracy requirements (say, 80% to 99% accuracy).

Our scheme improves on existing product-convolution schemes by providing an automated, theoretically justified method for performing adaptivity, and by addressing issues related to boundaries. Once constructed, the approximation can be manipulated efficiently and accessed in ways that the original operator cannot: matrix entries of the approximation can be computed at $O(1)$ cost, the approximation (or blocks of the approximation) can be efficiently applied to vectors with the FFT, and the approximation can be efficiently converted to $H$-matrix format. Once in $H$-matrix format, it can be factorized, inverted, or otherwise manipulated with fast hierarchical matrix arithmetic. Since our scheme is best suited to moderate accuracy requirements, the resulting $H$-matrix representation can be exploited to construct a good preconditioner.

We tested our scheme numerically on the non-local component of a boundary Schur complement for the Poisson operator, and on the data misfit Hessian for an advection-diffusion inverse problem. We saw that our scheme is mesh scalable for the Schur complement, and data scalable for the Hessian. In both cases our scheme far outperformed low-rank approximation. Additionally, we found that the scheme performs well even when only a handful of random samples are used to construct the a-posteriori error estimator used in the adaptive refinement procedure.
Acknowledgements. We thank J.J. Alger, Benjamin Babcock, Joe Bishop, Andrew Potter, Georg Stadler, Umberto Villa, and Hongyu Zhu for helpful discussions.

REFERENCES

[1] H.-M. Adorf, *Towards HST restoration with a space-variant PSF, cosmic rays and other missing data*, in The Restoration of HST Images and Spectra-II, 1994, p. 72.

[2] V. Akcèlik, G. Biros, A. Drăganescu, O. Ghattas, J. Hill, and B. van Bloemen Waanders, *Dynamic data-driven inversion for terascale simulations: Real-time identification of airborne contaminants*, in Proceedings of SC2005, Seattle, 2005.

[3] V. Akcèlik, G. Biros, O. Ghattas, J. Hill, D. Keyes, and B. van Bloemen Waanders, *Parallel PDE-constrained optimization*, in Parallel Processing for Scientific Computing, M. Heroux, P. Raghaven, and H. Simon, eds., SIAM, 2006.

[4] N. Alger, U. Villa, T. Bui-Thanh, and O. Ghattas, *A data scalable augmented Lagrangian KKT preconditioner for large-scale inverse problems*, SIAM Journal on Scientific Computing, 39 (2017), pp. A2365–A2393.

[5] H. Avron and S. Toledo, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix*, Journal of the ACM (JACM), 58 (2011), p. 8.

[6] I. Babuska and J. Meленk, *The partition of unity method*, in International Journal of Numerical Methods in Engineering, Citeseer, 1996.

[7] R. E. Bank and T. Dupont, *An optimal order process for solving finite element equations*, Mathematics of Computation, 36 (1981), pp. 35–51.

[8] G. Bao and W. W. Symes, *Computation of pseudo-differential operators*, SIAM Journal on Scientific Computing, 17 (1996), pp. 416–429.

[9] J. Bardsley, S. Jefferies, J. Nagy, and R. Plemmons, *A computational method for the restoration of images with an unknown, spatially-varying blur*, Optics Express, 14 (2006), pp. 1767–1782.

[10] M. Bebendorf, *Approximation of boundary element matrices*, Numerische Mathematik, 86 (2000), pp. 565–589.

[11] R. BéLanger-Rioux and L. Demanet, *Compressed absorbing boundary conditions via matrix probing*, SIAM Journal on Numerical Analysis, 53 (2015), pp. 2441–2471.

[12] J. Bigot, P. Escande, and P. Weiss, *Estimation of linear operators from scattered impulse responses*, arXiv preprint arXiv:1610.04056, (2016).

[13] J. E. Bishop, *A displacement-based finite element formulation for general polyhedra using harmonic shape functions*, International Journal for Numerical Methods in Engineering, 97 (2014), pp. 1–31.

[14] S. Börm, *Efficient numerical methods for non-local operators: H2-matrix compression, algorithms and analysis*, vol. 14, European Mathematical Society, 2010.

[15] S. Börm and L. Grasedyck, *Hybrid cross approximation of integral operators*, Numerische Mathematik, 101 (2005), pp. 221–249.

[16] D. Braess and W. Hackbusch, *A new convergence proof for the multigrid method including the V-cycle*, SIAM Journal on Numerical Analysis, 20 (1983), pp. 967–975.

[17] A. N. Brooks and T. J. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Computer methods in applied mechanics and engineering, 32 (1982), pp. 199–259.

[18] T. Bui-Thanh, C. Burstedde, O. Ghattas, J. Martin, G. Stadler, and L. C. Wilcox, *Extreme-scale UQ for Bayesian inverse problems governed by PDEs*, in SC12: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2012.

[19] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler, *A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion*, SIAM Journal on Scientific Computing, 35 (2013), pp. A2493–A2523.

[20] D. Calvetti, P. J. Hadwin, J. Huttunen, J. P. Kaipio, and E. Somersalo, *Artificial boundary conditions and domain truncation in electrical impedance tomography. Part II: Stochastic extension of the boundary map*, Inverse Problems Imaging, 9 (2015), pp. 767–89.

[21] D. Calvetti, P. J. Hadwin, J. M. Huttunen, D. Isaacson, J. P. Kaipio, D. McGivney, E. Somersalo, and J. Volzer, *Artificial boundary conditions and domain truncation in electrical impedance tomography. Part I: theory and preliminary results*, Inverse Problems Imaging, 9 (2015), pp. 749–66.

[22] D. Calvetti, B. Lewis, and L. Reichel, *Restoration of images with spatially variant blur by
the GMRES method, in Advanced Signal Processing Algorithms, Architectures, and Implementations X, ed. FT Luk, Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE), vol. 4116, 2000, pp. 364–374.

[23] R. H. Chan and M. K. Ng, Conjugate gradient methods for Toeplitz systems, SIAM review, 38 (1996), pp. 427–452.

[24] S. Chandrasekaran, M. Gu, and W. Lyons, A fast adaptive solver for hierarchically semiseparable representations, Calcolo, 42 (2005), pp. 171–185.

[25] P. Chen, U. Villa, and O. Ghattas, Taylor approximation and variance reduction for pde-constrained optimal control problems under uncertainty, preprint, 2017.

[26] H. Cheng, Z. Gimbutas, P.-G. Martinsson, and V. Rokhlin, On the compression of low rank matrices, SIAM Journal on Scientific Computing, 26 (2005), pp. 1389–1404.

[27] J. Chiu and L. Demanet, Matrix probing and its conditioning, SIAM Journal on Numerical Analysis, 50 (2012), pp. 171–193.

[28] P. G. Ciarlet, Discrete maximum principle for finite-difference operators, Aequationes mathematicae, 4 (1970), pp. 338–352.

[29] T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini, Likelihood-informed dimension reduction for nonlinear inverse problems, Inverse Problems, 30 (2014), p. 114015.

[30] L. Demanet, P.-D. Létourneau, N. Boumal, H. Calandra, J. Chiu, and S. Snelson, Matrix probing: a randomized preconditioner for the wave-equation Hessian, Applied and Computational Harmonic Analysis, 32 (2012), pp. 155–168.

[31] L. Demanet and L. Ying, Discrete symbol calculus, SIAM review, 53 (2011), pp. 71–104.

[32] L. Denis, E. Thiébaut, F. Soulez, J.-M. Becker, and R. Mohya, Fast approximations of shift-variant blur, International Journal of Computer Vision, 115 (2015), pp. 253–278.

[33] P. Escande and P. Weiss, Sparse wavelet representations of spatially varying blurring operators, SIAM Journal on Imaging Sciences, 8 (2015), pp. 2976–3014.

[34] P. Escande and P. Weiss, Approximation of integral operators using product-convolution expansions, Journal of Mathematical Imaging and Vision, 58 (2017), pp. 333–348.

[35] P. Escande, P. Weiss, and F. Malgouyres, Spatially varying blur recovery. Diagonal approximations in the wavelet domain, (2012).

[36] D. Fish, J. Grochmalicki, and E. Pike, Scanning singular-value-decomposition method for restoration of images with space-variant blur, JOSA A, 13 (1996), pp. 464–469.

[37] H. P. Flath, L. C. Wilcox, V. Akçelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas, Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations, SIAM Journal on Scientific Computing, 33 (2011), pp. 407–432.

[38] R. C. Flicker and F. J. Rigaut, Anisoplanatic deconvolution of adaptive optics images, JOSA A, 22 (2005), pp. 504–513.

[39] M. Gentile, F. Courbin, and G. Meylan, Interpolating point spread function anisotropy, Astronomy & Astrophysics, 549 (2013).

[40] E. Gilad and J. von Hardenberg, A fast algorithm for convolution integrals with space and time variant kernels, Journal of Computational Physics, 216 (2006), pp. 326–336.

[41] W. Hackbusch, A sparse matrix arithmetic based on H-matrices. Part I: Introduction to H-matrices, Computing, 62 (1999), pp. 89–108.

[42] W. Hackbusch, Hierarchical matrices: algorithms and analysis, vol. 49, Springer, 2015.

[43] W. Hackbusch, B. Khoromskaia, and S. A. Sauter, On H2-matrices, in Lectures on Applied Mathematics: Proceedings of the Symposium Organized by the Sonderforschungsbereich 438 on the Occasion of Karl-Heinz Hoffmann’s 60th Birthday, Munich, June 30–July 1, 1999, Springer Science & Business Media, 2000, p. 9.

[44] S. B. Hadj, L. Blanc-Féraud, G. Aubert, and G. Emile, Blind restoration of confocal microscopy images in presence of a depth-variant blur and poisson noise, in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, IEEE, 2013, pp. 915–919.

[45] N. Halko, P.-G. Martinsson, and J. A. Tropp, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, SIAM review, 53 (2011), pp. 217–288.

[46] F. J. Herrmann and G. Hennenfent, Non-parametric seismic data recovery with curvelet frames, Geophysical Journal International, 173 (2008), pp. 233–248.

[47] M. Hirsch, S. Sra, B. Schölkopf, and S. Harmeling, Efficient filter flow for space-variant multiframe blind deconvolution, in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 607–614.

[48] L. Hörmander, Pseudo-differential operators, in The Analysis of Linear Partial Differential Operators III, Springer, 2007, pp. 63–179.
[49] T. Hrycak, S. Das, G. Matz, and H. G. Feichtinger, 
Low complexity equalization for doubly 
selective channels modeled by a basis expansion, IEEE 
Transactions on Signal Processing, 
58 (2010), pp. 5706–5719.

[50] T. Isaac, N. Petra, G. Stadler, and O. Ghattas, 
Scaleable and efficient algorithms for 
the propagation of uncertainty from data through inference to prediction for large-scale 
problems, with application to flow of the Antarctic ice sheet, 
Journal of Computational Physics, 
296 (2015), pp. 348–368, doi:10.1016/j.jcp.2015.04.047.

[51] J. E. Jones and P. S. Vassilevski, 
AMGe based on element agglomeration, SIAM Journal on 
Scientific Computing, 23 (2001), pp. 109–133.

[52] T. J. Kohn and L. Nirenberg, 
An algebra of pseudo-differential operators, 
Communications on Pure and Applied Mathematics, 
18 (1965), pp. 269–305.

[53] J. Lin, J. Lu, and L. Ying, 
Fast construction of hierarchical matrix representation from 
matrix–vector multiplication, Journal of Computational Physics, 
230 (2011), pp. 4071–4087.

[54] M. W. Mahoney, 
Randomized algorithms for matrices and data, Foundations and Trends® 
in Machine Learning, 3 (2011), pp. 123–224.

[55] P.-G. Martinsson, 
A fast randomized algorithm for computing a hierarchically semiseparable 
representation of a matrix, SIAM Journal on Matrix Analysis and Applications, 
32 (2011), pp. 1251–1274.

[56] P.-G. Martinsson, 
Compressing rank-structured matrices via randomized sampling, SIAM 
Journal on Scientific Computing, 38 (2016), pp. A1959–A1986.

[57] P.-G. Martinsson and V. Rokhlin, 
A fast direct solver for boundary integral equations in 
two dimensions, Journal of Computational Physics, 
205 (2005), pp. 1–23.

[58] J. G. Nagy and D. P. O’Leary, 
Restoring images degraded by spatially variant blur, SIAM 
Journal on Scientific Computing, 19 (1998), pp. 1063–1082.

[59] J. G. Nagy, K. Palmer, and L. Perrone, 
Iterative methods for image deblurring: a Matlab 
object-oriented approach, Numerical Algorithms, 
36 (2004), pp. 73–93.

[60] J. Ng, R. Prager, N. Kingsbury, G. Treece, and A. Gee, 
Wavelet restoration of medical 
pulse-echo ultrasound images in an EM framework, IEEE Transactions on Ultrasonics, 
Ferroelectrics, and Frequency Control, 54 (2007).

[61] N. Petra, J. Martin, G. Stadler, and O. Ghattas, 
A computational framework for infinite- 
dimensional Bayesian inverse problems, Part II: Stochastic Newton MCMC with applica- 
tion to ice sheet flow inverse problems, SIAM Journal on Scientific Computing, 
36 (2014), pp. A1525–A1555.

[62] C. Preza and J.-A. Conchello, 
Depth-variant maximum-likelihood restoration for three- 
dimensional fluorescence microscopy, JOSA A, 21 (2004), pp. 1593–1601.

[63] A. Rogers and J. D. Fiege, 
Strong gravitational lens modeling with spatially variant point- 
spread functions, The Astrophysical Journal, 743 (2011), p. 68.

[64] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk, 
Optimal low- 
rank approximations of Bayesian linear inverse problems, SIAM Journal on Scientific Computing, 
37 (2015), pp. A2451–A2487.

[65] H. Trussell and B. Hunt, 
Sectioned methods for image restoration, IEEE Transactions on 
Acoustics, Speech, and Signal Processing, 26 (1978), pp. 157–164.

[66] H. J. Trussell and S. Fogel, 
Identification and restoration of spatially variant motion blurs 
in sequential images, IEEE Transactions on Image Processing, 
1 (1992), pp. 123–126.

[67] E. Tyrtyshnikov, 
Incomplete cross approximation in the mosaic-skeleton method, Computing, 
64 (2000), pp. 367–380.

[68] J. Xia, S. Chandrasekaran, M. Gu, and X. S. Li, 
Superfast multifrontal method for large 
structured linear systems of equations, SIAM Journal on Matrix Analysis and Applications, 
31 (2009), pp. 1382–1411.

[69] H. Zhu, S. Li, S. Fomel, G. Stadler, and O. Ghattas, 
A Bayesian approach to estimate 
uncertainty for full-waveform inversion using a priori information from depth migration, 
Geophysics, (2016).