MULTISCALE MODEL. SIMUL.
Vol. 19, No. 2, pp. 688–730

COMPRESSION, INVERSION, AND APPROXIMATE PCA OF DENSE KERNEL MATRICES AT NEAR-LINEAR COMPUTATIONAL COMPLEXITY*

FLORIAN SCHÄFER†, T. J. SULLIVAN‡, AND HOUMAN OWHADI§

Abstract. Dense kernel matrices $\Theta \in \mathbb{R}^{N\times N}$ obtained from point evaluations of a covariance function G at locations $\{x_i\}_{1\leq i\leq N} \subset \mathbb{R}^d$ arise in statistics, machine learning, and numerical analysis. For covariance functions that are Green's functions of elliptic boundary value problems and homogeneously distributed sampling points, we show how to identify a subset $S \subset \{1,\dots,N\}^2$, with $\#S = \mathcal{O}(N\log(N)\log^d(N/\epsilon))$, such that the zero fill-in incomplete Cholesky factorization of the sparse matrix $\Theta_{ij}\mathbf{1}_{(i,j)\in S}$ is an ϵ-approximation of Θ. This factorization can provably be obtained in complexity $\mathcal{O}(N\log(N)\log^d(N/\epsilon))$ in space and $\mathcal{O}(N\log^2(N)\log^{2d}(N/\epsilon))$ in time, improving upon the state of the art for general elliptic operators; we further present numerical evidence that d can be taken to be the intrinsic dimension of the data set rather than that of the ambient space. The algorithm only needs to know the spatial configuration of the x_i and does not require an analytic representation of G. Furthermore, this factorization straightforwardly provides an approximate sparse PCA with optimal rate of convergence in the operator norm. Hence, by using only subsampling and the incomplete Cholesky factorization, we obtain, at nearly linear complexity, the compression, inversion, and approximate PCA of a large class of covariance matrices. By inverting the order of the Cholesky factorization we also obtain a solver for elliptic PDE with complexity $\mathcal{O}(N\log^d(N/\epsilon))$ in space and $\mathcal{O}(N\log^{2d}(N/\epsilon))$ in time, improving upon the state of the art for general elliptic operators.

Key words. Cholesky factorization, covariance function, gamblet transform, kernel matrix, sparsity, principal component analysis

AMS subject classifications. 65F30, 42C40, 65F50, 65N55, 65N75, 60G42, 68Q25, 68W40

DOI. 10.1137/19M129526X

1. Introduction.

1.1. Dense kernel matrices and the N^3-bottleneck.

Kernel matrices, i.e., square matrices Θ of the form

$$\Theta_{ij} \coloneqq G(x_i, x_j), \tag{1.1}$$

obtained from pointwise evaluation of a symmetric positive-definite kernel G at a collection of points $\{x_i\}_{i\in I}$ in a domain $\Omega \subset \mathbb{R}^d$, play an important role in statistics, machine learning, and scientific computing. In statistics, they are used as covariance matrices of Gaussian process priors. In machine learning, they equip the feature space with a meaningful inner product via the *kernel trick* [42]. In scientific computing,

*Received by the editors October 24, 2019; accepted for publication (in revised form) October 9, 2020; published electronically April 15, 2021.
https://doi.org/10.1137/19M129526X
Funding: The first and third authors gratefully acknowledge support by the Air Force Office of Scientific Research and the DARPA EQUiPS Program (award FA9550-16-1-0054, Computational Information Games) and the Air Force Office of Scientific Research (award FA9550-18-1-0271, Games for Computation and Learning). The second author has been supported by the Freie Universitat Berlin within the Excellence Initiative of the German Research Foundation. This collaboration has been facilitated by the Statistical and Applied Mathematical Sciences Institute through the National Science Foundation award DMS-1127914.

†Corresponding author. Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125 USA (florian.schaefer@caltech.edu).
‡Mathematics Institute and School of Engineering, University of Warwick, Coventry, CV4 7AL, UK, and Zuse Institute Berlin, 14195 Berlin, Germany (t.j.sullivan@warwick.ac.uk).
§Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125 USA (owhadi@caltech.edu).

they appear as Green's functions (i.e., fundamental solutions) of linear elliptic partial differential equations (PDEs).

For all these applications, it is usually necessary to perform some or all of the following tasks:

(1) compute $v \mapsto \Theta v$, given $v \in \mathbb{R}^I$;
(2) compute $v \mapsto \Theta^{-1} v$, given $v \in \mathbb{R}^I$;
(3) compute $\log\det\Theta$;
(4) sample from the normal/Gaussian distribution $\mathcal{N}(0, \Theta)$;
(5) approximate eigenspaces corresponding to the leading eigenvalues of Θ.

The first four of these tasks can be performed by computing the Cholesky factorization of Θ (i.e., the decomposition $\Theta = LL^T$ where L is lower triangular). For many popular covariance functions, most notably those of smooth random processes, the matrices Θ will be *dense*. For large $N := \#I$ this results in a computational complexity of $\mathcal{O}(N^3)$ for the Cholesky factorization and a complexity of $\mathcal{O}(N^2)$ to even store the matrix. When Θ is *sparse*, i.e., has relatively few nonzero entries, better complexity can be achieved—the obvious limiting case being $\mathcal{O}(N)$ (i.e., linear) complexity if Θ is diagonal. However, for practical problems, the cubic scaling restricts dense Cholesky factorization to problems with $N \lessapprox 10^5$. The breadth of kernel matrices' uses means that there is correspondingly high interest in achieving approximate Cholesky factorization of Θ at linear or near-linear cost.

1.2. Existing approaches. Many fast methods are available for approximating dense kernel matrices and their applicability depends on specific assumptions made on Θ. If the precision matrix Θ^{-1} is sparse and can be approximated directly (e.g., by discretizing a PDE), then sparse linear solvers can be used. These include multigrid solvers [22, 16, 36, 38] and sparse Cholesky factorization methods with nested dissection ordering [29, 28, 55, 30]. This approach has been proposed for problems arising in spatial statistics [54, 69, 70, 68]. In other situations, available methods directly approximate the covariance matrix based on low-rank approximations, sparsity, and hierarchy. Low-rank techniques such as the Nyström approximation [82, 77, 26] or rank-revealing Cholesky factorization [4, 24] seek to approximate Θ by low-rank matrices whereas sparsity-based methods like *covariance tapering* [27] seek to approximate Θ with a sparse matrix by setting entries corresponding to long-range interactions to zero. These two approximations can also be combined to obtain *sparse low-rank* approximations [73, 66, 75, 5, 78], which can be interpreted as imposing a particular graphical structure on the Gaussian process. When Θ is neither sufficiently sparse nor of sufficiently low rank, these approaches can be implemented in a hierarchical manner. For low-rank methods, this leads to *hierarchical* ($\mathcal{H}$- and $\mathcal{H}^2$-) matrices [40, 37, 39], *hierarchical off-diagonal low rank* (HODLR) matrices [2, 3], and hierarchically semiseparable (HSS) matrices [18, 83, 52] that rely on computing low-rank approximations of subblocks of Θ corresponding to far-field interactions on different scales. The interpolative factorization developed by [41] combines hierarchical low-rank structure with the sparsity obtained from an elimination ordering of nested-dissection type. Hierarchical low-rank structure was originally developed as an algebraic abstraction of the fast multipole method of [33]. In order to construct hierarchical low-rank approximations from entries of the kernel matrix efficiently, both deterministic and randomized algorithms have been proposed [8, 58]. For many popular covariance functions, including Green's functions of elliptic PDEs [7], hierarchical matrices allow for (near-)linear-in-N complexity algorithms for the inversion and approximation of Θ, at exponential accuracy. Wavelet-based methods [12, 31], using the

separation and truncation of interactions on different scales, can be seen as a hierarchical application of sparse approximation approaches. The resulting algorithms have near-linear computational complexity and rigorous error bounds for asymptotically smooth covariance functions. [23] uses operator-adapted wavelets to compress the expected solution operators of random elliptic PDEs. In [48], although no rigorous accuracy estimates are provided, the authors establish the near-linear computational complexity of algorithms resulting from the multiscale generalization of probabilistically motivated sparse and low-rank approximations [73, 66, 75, 5, 78].

1.3. Our main result and and overview of the paper. Our main result is to show that a small modification of the Cholesky factorization algorithm is both accurate and scalable, when applied to kernel matrices obtained from kernels G identified as Green's functions of elliptic PDEs and a (roughly) homogeneously distributed cloud of points. Such kernels are oftentimes used as covariance functions of smooth Gaussian processes (to enforce a smoothness prior on the function to be recovered/interpolated) and therefore a large class of popular kernels fall into this category. The cheap, accurate, approximate Cholesky factors provided by our method thereby serve tasks (1)–(4) from subsection 1.1. We furthermore show that by reversing the elimination order we obtain a fast direct solver for elliptic PDEs.

Contrary to the present belief that fast solvers for elliptic integral operators require the use of hierarchical low-rank structure or wavelets with a high order of vanishing moments, we show that state-of-the-art performance can be obtained just by zero fill-in Cholesky factorization (which just amounts to skipping some steps in the Cholesky factorization algorithm—wavelets are only used in the detailed rigorous analysis of the algorithm). While there is a huge literature on the sparse Cholesky factorization of *sparse* matrices, we are not aware of any prior literature on the sparse Cholesky factorization of *dense* matrices.

For elliptic PDEs with arbitrary L^∞-coefficients, $\mathcal{H}$-matrices can be used to compute ϵ-approximate Cholesky factors of both differential and integral operators in computational complexity $\mathcal{O}(N\log^2(N)\log^{2d+2}(\epsilon^{-1}))$ [7, 37, 40, 6]. $\mathcal{H}^2$-matrices can improve these complexities to $\mathcal{O}(N\log(N)\log^{2d+2}(\epsilon^{-1}))$ [39, 13, 14]. The "fast gamblet transform" of [62, 63] can invert stiffness matrices of arbitrary elliptic operators in computational complexity $\mathcal{O}(\log^{2d+1}(\epsilon^{-1}))$. Our computational complexities of $\mathcal{O}(N\log^{2d}(N/\epsilon))$ for the Cholesky factorization of differential operators and $\mathcal{O}(N\log^2(N)\log^{2d}(N/\epsilon))$ for the Cholesky factorization of integral operators improve upon the state of the art while using a much simpler algorithm.

Our method relies upon a cleverly constructed elimination ordering and sparsity pattern, which we use in the incomplete Cholesky factorization of the matrix Θ. Simplified versions of these constructions are given in section 2; section 3 gives an overview, without detailed proof, of why the method yields the desired results. In particular, subsection 2.4 shows how the method provides a sparse approximate principal component analysis (PCA), thereby serving task (5).

Section 4 presents detailed numerical experiments that illustrate the power of our method, and section 5 gives the mathematical proofs of correctness and accuracy vs. complexity. Section 8 contains concluding remarks, and some technical results are deferred to an appendix.

2. Overview of the algorithm and its setting. In this introductory section we give a brief overview of the setting in which our theoretical results apply (the class of kernels associated to elliptic operators) and highlight its main features. All detailed numerical experiments and analysis will be deferred to sections 4 and 5, respectively.

2.1. The class of elliptic operators. In order to establish rigorous, a priori, complexity-vs.-accuracy estimates in section 5 we will assume that G is the Green's function of an elliptic operator $\mathcal{L}$ of order $2s > d$ ($s, d \in \mathbb{N}$), defined on a bounded domain $\Omega \subset \mathbb{R}^d$ with Lipschitz boundary, and acting on $H_0^s(\Omega)$, the Sobolev space of (zero boundary value) functions having derivatives of order s in $L^2(\Omega)$. More precisely, writing $H^{-s}(\Omega)$ for the dual space of $H_0^s(\Omega)$ with respect to the $L^2(\Omega)$ scalar product, our rigorous estimates will be stated for an arbitrary linear bijection

$$\mathcal{L}\colon H_0^s(\Omega) \to H^{-s}(\Omega) \tag{2.1}$$

that is *symmetric* (i.e., $\int_\Omega u\mathcal{L}v\,\mathrm{d}x = \int_\Omega v\mathcal{L}u\,\mathrm{d}x$), *positive* (i.e., $\int_\Omega u\mathcal{L}u\,\mathrm{d}x \geq 0$), and *local* in the sense that

$$\int_\Omega u\mathcal{L}v\,\mathrm{d}x = 0 \ \forall u, v \in H_0^s(\Omega) \text{ such that } \operatorname{supp} u \cap \operatorname{supp} v = \emptyset. \tag{2.2}$$

Let $\|\mathcal{L}\| \coloneqq \sup_{u\in H_0^s} \|\mathcal{L}u\|_{H^{-s}}/\|u\|_{H_0^s}$ and $\|\mathcal{L}^{-1}\| \coloneqq \sup_{f\in H^{-s}} \|\mathcal{L}^{-1}f\|_{H_0^s}/\|f\|_{H^{-s}}$ denote the operator norms of $\mathcal{L}$ and $\mathcal{L}^{-1}$. The complexity and accuracy estimates for our algorithm will depend on (and only on) $d, s, \Omega, \|\mathcal{L}\|$, $\|\mathcal{L}^{-1}\|$ and the parameter

$$\delta \coloneqq \frac{\min_{i\neq j\in I} \operatorname{dist}(x_i, \{x_j\} \cup \partial\Omega)}{\max_{x\in\Omega} \operatorname{dist}(x, \{x_i\}_{i\in I} \cup \partial\Omega)}, \tag{2.3}$$

which is a measure of the homogeneity of the distribution of the cloud of points x_i.

Since our algorithm only requires the locations of the points x_i and is oblivious to the exact knowledge of G, for our numerical experiments in section 4 we will consider (2.1), general elliptic operators with or without boundary conditions (these include Matérn kernels with fractional values of s) and exponential kernels.

2.2. Zero fill-in incomplete Cholesky factorization (`ICHOL(0)`). A simple approach to decreasing the computational complexity of Cholesky factorization is the *zero fill-in incomplete Cholesky factorization* [60] (`ICHOL(0)`). When performing Gaussian elimination using `ICHOL(0)`, we treat all entries of both the input matrix and the output factors outside a prescribed *sparsity pattern* $S \subset I \times I$ as zero and correspondingly ignore all operations in which they are involved. Figure 2.1 shows a comparison of ordinary Cholesky factorization and `ICHOL(0)`. Our approach to kernel matrices consists of applying Algorithm 2.2 with an elimination ordering $\prec$ and a sparsity pattern S that are chosen based on the locations of the x_i; Construction 5.25 gives the details of this elimination ordering and sparsity pattern.

Write $\|\cdot\|_{\mathrm{Fro}}$ for the Frobenius matrix norm and C for a constant depending only on d, Ω, s, $\|\mathcal{L}\|$, $\|\mathcal{L}^{-1}\|$, and δ. To simplify notation, the asymptotic bounds in this paper are stated in the case where the logarithmic factors are at least one. Our main result is the following.

THEOREM 2.1. *Let $\mathcal{L}$ and δ be defined as in* (2.1) *and* (2.3). *For $\rho \geq C\log(N/\epsilon)$, the sparse Cholesky factor L^ρ, obtained from Algorithm* 2.2 *with the elimination ordering $\prec_\rho$ and sparsity pattern $\tilde{S}_\rho \subset I \times I$ described in Construction* 5.25, *satisfies*

$$\left\|\Theta - L^\rho L^{\rho,\top}\right\|_{\mathrm{Fro}} \leq \epsilon. \tag{2.4}$$

The selection of the ordering and sparsity pattern, as well as Algorithm 2.2, *can be performed in computational complexity $C\rho^{2d}N\log^2 N$ in time and $C\rho^d N\log N$ in space. In particular, we can obtain an ϵ-accurate approximation in Frobenius norm in complexity $CN\log^2(N)\log(N/\epsilon)^{2d}$ in time and $CN\log(N)\log(N/\epsilon)^d$ in space.*

Algorithm 2.1 Standard dense Cholesky factorization.

Input: $A \in \mathbb{R}^{N\times N}$ symmetric
Output: $L \in \mathbb{R}^{N\times N}$ lower triang.

for $i \in \{1,\dots,N\}$ **do**
  $L_{:i} \leftarrow A_{:i}/\sqrt{A_{ii}}$
  for $j \in \{i+1,\dots,N\}$ **do**
    for $k \in \{j,\dots,N\}$ **do**
      $A_{kj} \leftarrow A_{kj} - \frac{A_{ki}A_{ji}}{A_{ii}}$
    end for
  end for
end for
return L

Algorithm 2.2 Incomplete Cholesky factorization with sparsity pattern S.

Input: $A \in \mathbb{R}^{N\times N}$ symmetric, $\mathrm{nz}(A) \subset S$
Output: $L \in \mathbb{R}^{N\times N}$ lower triang. $\mathrm{nz}(L) \subset S$

for $(i,j) \notin S$ **do**
  $A_{ij} \leftarrow 0$
end for
for $i \in \{1,\dots,N\}$ **do**
  $L_{:i} \leftarrow A_{:i}/\sqrt{A_{ii}}$
  for $j \in \{i+1,\dots,N\} : (i,j) \in S$ **do**
    for $k \in \{j,\dots,N\} : (k,i),(k,j) \in S$ **do**
      $A_{kj} \leftarrow A_{kj} - \frac{A_{ki}A_{ji}}{A_{ii}}$
    end for
  end for
end for
return L

FIG. 2.1. *Comparison of ordinary and incomplete Cholesky factorization. Here, for a matrix* A, $\mathrm{nz}(A) \coloneqq \{(i,j) \mid A_{ij} \neq 0\}$ *denotes the index set of the nonzero entries of* A.

Remark 2.2. For problems arising in Gaussian process regression, there will typically be no domain Ω on the boundary of which the process is conditioned to be zero; equivalently, Ω will be all of $\mathbb{R}^d$. This introduces an additional error, but we still observe good approximation of the covariances even of points close to the boundary (see subsection 4.2 for a detailed discussion).

We will now present a simplified version of the elimination ordering and sparsity pattern (compared to the one mentioned in Theorem 2.1). Although the proof of Theorem 2.1 does not cover the stability of `ICHOL(0)` under this simplified version (rather, it covers the one described in Construction 5.25), extensive numerical experiments suggest that `ICHOL(0)` remains stable under this simplified version, and since it is also *user-friendly* we recommend this as the "go-to" version for a simple, practical implementation.[1]

2.3. The elimination ordering and sparsity pattern. We use a *maximum-minimum distance ordering* (maximin ordering) [34] as the elimination ordering. This ordering is obtained by successively picking the point x_i that is furthest away from $\partial\Omega$ and the points that were already picked. If $\partial\Omega = \emptyset$, then we select an arbitrary $i \in I$ as first index to eliminate; otherwise, we choose the first index as

$$i_1 \coloneqq \operatorname*{arg\,max}_{i\in I} \operatorname{dist}(x_i, \partial\Omega). \tag{2.5}$$

Then, for the first k indices of the ordering already chosen, we choose

$$i_{k+1} \coloneqq \operatorname*{arg\,max}_{i\in I\setminus\{i_1,\dots,i_k\}} \operatorname{dist}(x_i, \{x_{i_1},\dots,x_{i_k}\} \cup \partial\Omega) \tag{2.6}$$

until we have ordered all the N points (see Figure 2.2).

[1] Although more complex, the ordering used in Theorem 2.1 has more potential for optimization by exploiting parallelism and dense linear algebra operations.

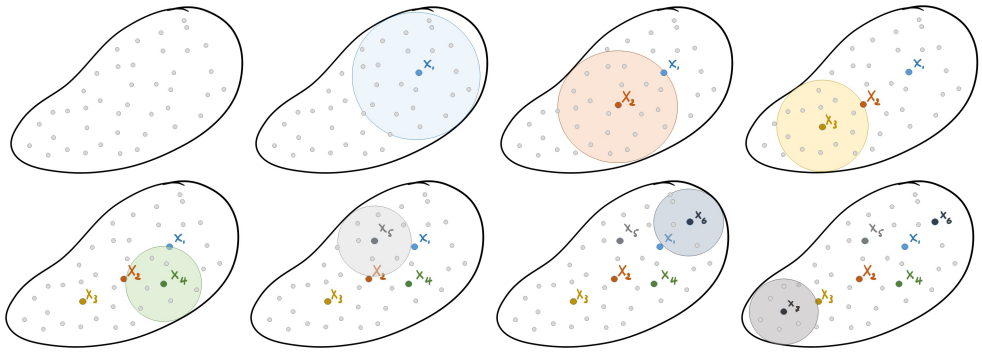

FIG. 2.2. *The maximin order successively adds the point that is furthest away from both $\partial\Omega$ and the set of points already added. The radius of the shaded circle is $l[i]$.*

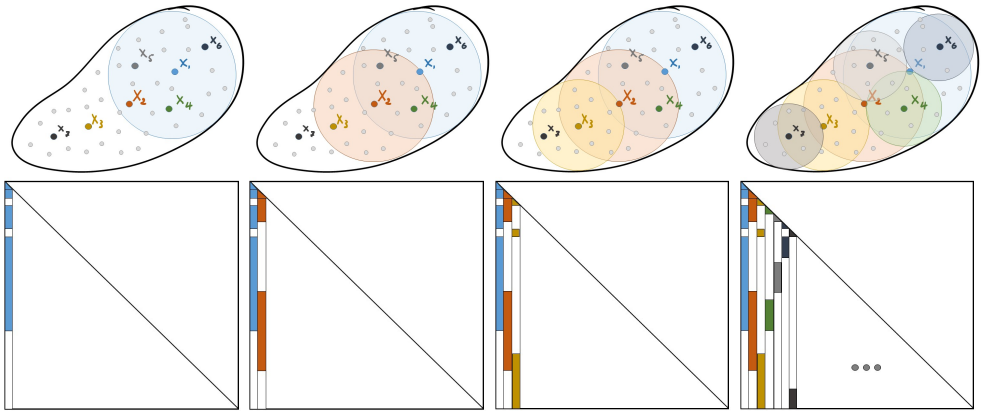

FIG. 2.3. *Upper row: radii of interaction of the different degrees of freedom, for $\rho = 1$. Lower row: corresponding columns of the sparsity pattern. While the first columns are relatively dense, subsequent columns become more and more sparse.*

Let

$$l[i_k] \coloneqq \operatorname{dist}(x_{i_k}, \{x_{i_1}, \dots, x_{i_{k-1}}\} \cup \partial\Omega) \tag{2.7}$$

be the distance between x_{i_k} and $\partial\Omega$ and the earlier points in the ordering. For $\rho > 0$, let $S_\rho \subset I \times I$ be the sparsity pattern defined by

$$S_\rho \coloneqq \{(i,j) \in I \times I \mid \operatorname{dist}(x_i, x_j) \leq \rho \max(l[i], l[j])\}. \tag{2.8}$$

Here, ρ parameterizes a trade-off between computational efficiency and accuracy. For a given ρ, the sparsity pattern will have $C\rho^d N \log N$ entries and the Cholesky factorization will require $C\rho^{2d} N \log^2 N$ floating-point operations. Figure 2.3 shows the sparsity pattern for $\rho = 1$. While a naïve implementation requires $\mathcal{O}(N^2)$ distance evaluations, Theorem 4.1 shows that Algorithm 4.1 delivers this sparsity pattern at computational complexity $C\rho^d N \log^2 N$.

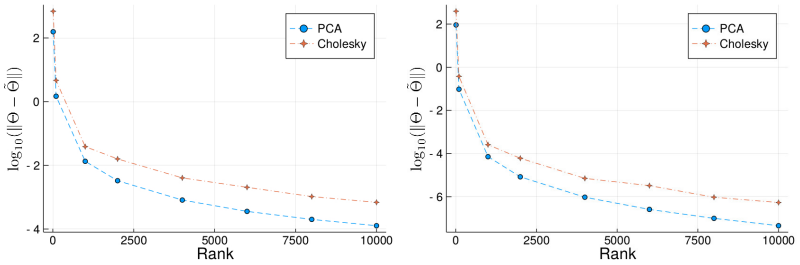

FIG. 2.4. *Near-optimal sparse PCA: Approximation errors comparisons between low-rank Cholesky ($\rho = \infty$) and PCA for a Matérn kernel with smoothness parameters $\nu = 1$ (left) and $\nu = 2$ (right).*

2.4. Sparse approximate PCA. The sparse Cholesky factorization described in section 2 is also *rank revealing* in the sense that the low-rank approximation obtained by using only the first k columns of the Cholesky factorization achieves an accuracy within a constant factor of optimal rank-k approximation (measured in operator norm). This is illustrated by Figure 2.4 and the following theorem.

THEOREM 2.3. *In the setting of Theorem* 2.1*, let $L^{(k)}$ be the rank-k matrix defined by the first k columns of the (dense) Cholesky factor L of Θ. Then*

$$\left\|\Theta - L^{(k)}L^{(k),\top}\right\| \leq C\|\Theta\|k^{-\frac{2s}{d}}, \tag{2.9}$$

where $\|\Theta\|$ is the operator norm of Θ and $C > 0$ depends only on d, Ω, s, $\|\mathcal{L}\|$, $\|\mathcal{L}^{-1}\|$, and δ.

The rank-k approximation estimate (2.9) is a numerical homogenization accuracy estimate similar those obtained in [57, 65, 62, 63, 44]. Numerical homogenization basis functions can be identified by the *last* k rows of the lower triangular Cholesky factor of $A \coloneqq \Theta^{-1}$, obtained with the reverse elimination ordering described in subsection 6.2.

3. Why it works—justification of the method. The method described in section 2 combines two crude approximations. First, it discards all but $\mathcal{O}(\rho^d N \log N)$ entries of the *dense* $N \times N$ matrix Θ. Second, it skips all but $\mathcal{O}(\rho^{2d} N \log^2 N)$ operations of the Cholesky factorization of Θ (which has complexity $\mathcal{O}(N^3)$). The obvious question is, why is the resulting approximation of Θ accurate for $\rho \gtrsim \log N$?

3.1. Sparse Cholesky factors of dense matrices. The first part of the answer is that the Cholesky factors of Θ decay exponentially quickly away from the sparsity pattern S_ρ when the maximin ordering is used as the elimination ordering. This decay is illustrated in Figure 3.1 and by Theorem 3.1. Write C for a constant depending only on d, Ω, s, $\|\mathcal{L}\|$, $\|\mathcal{L}^{-1}\|$, and δ.

THEOREM 3.1. *In the setting of Theorem* 2.1*, let L be the full Cholesky factor of Θ in the maximin ordering of section* 2*. Then, for $\rho \geq C\log(N/\epsilon)$, S_ρ as defined in section* 2*, and*

$$L_{ij}^{S_\rho} \coloneqq L_{ij}\mathbf{1}_{(i,j)\in S_\rho} = \begin{cases} L_{ij} & \text{for } (i,j) \in S_\rho, \\ 0 & \text{else,} \end{cases} \tag{3.1}$$

the inequality $\|\Theta - L^{S_\rho}L^{S_\rho,\top}\|_{\mathrm{Fro}} \leq \epsilon$ holds.

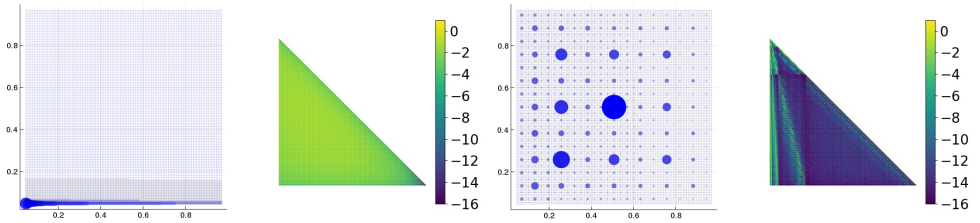

FIG. 3.1. *The lexicographic (left) and maximin (right) ordering of points in* $\Omega := (0,1)^2$ *with larger and darker nodes corresponding to earlier elements of the ordering, together with the corresponding Cholesky factors of* Θ *with entries plotted on a* $\log_{10}$*-scale.*

Algorithm 2.2 computes the exact Cholesky factorization under the assumption that the entries of L lying outside S_ρ are zero. Theorem 3.1 shows that this assumption holds true up to an approximation error that decays exponentially in ρ, which supports the claim of accuracy of Algorithm 2.2 for $\rho \gtrsim \log N$. We will now explain the exponential decay of L based on a probabilistic interpretation of Gaussian elimination.

3.2. Gaussian elimination, conditioning of Gaussian random variables, and the screening effect. The dense (block-)Cholesky factorization of a matrix Θ can be seen as the recursive application of the matrix identity

$$\begin{pmatrix} \Theta_{1,1} & \Theta_{1,2} \\ \Theta_{2,1} & \Theta_{2,2} \end{pmatrix} = \begin{pmatrix} \mathrm{Id} & 0 \\ \Theta_{2,1}(\Theta_{1,1})^{-1} & \mathrm{Id} \end{pmatrix} \begin{pmatrix} \Theta_{1,1} & 0 \\ 0 & \Theta_{2,2} - \Theta_{2,1}(\Theta_{1,1})^{-1}\Theta_{1,2} \end{pmatrix} \begin{pmatrix} \mathrm{Id} & (\Theta_{1,1})^{-1}\Theta_{1,2} \\ 0 & \mathrm{Id} \end{pmatrix},$$

where, at each step of the outermost loop, the above identity is applied to the Schur complement $\Theta_{2,2} - \Theta_{2,1}(\Theta_{1,1})^{-1}\Theta_{1,2}$ obtained at the previous step. If the Schur complements appearing during the factorization are sparse, then the final Cholesky factorization will also be sparse.

For $X = (X_1, X_2) \sim \mathcal{N}(0, \Theta)$, the well-known identities

$$\mathbb{E}[X_2 \mid X_1 = a] = \Theta_{2,1}(\Theta_{1,1})^{-1}a, \tag{3.2}$$

$$\mathrm{Cov}[X_2 \mid X_1] = \Theta_{2,2} - \Theta_{2,1}(\Theta_{1,1})^{-1}\Theta_{1,2} \tag{3.3}$$

imply that the sparsity of Cholesky factors of Θ is equivalent to conditional independence of Gaussian vectors with covariance matrix Θ. In the spatial statistics literature, it is well known that many smooth Gaussian processes are subject to the *screening effect* [79]. This effect, illustrated in Figure 3.2, means that the value of the process at a given site, conditioned on the values at nearby sites, is only weakly dependent on the values at distant sites.

Consider now the kth step of Cholesky factorization in the ordering described in section 2. Any pair x_i, x_j with $\mathrm{dist}(x_i, x_j) \gtrapprox l[k]$ will have points between them that have already been eliminated, as illustrated in Figure 3.3. Thus, the screening effect suggests that their correlation will be weak, which supports choosing $\rho l[k]$ as a truncation radius.

3.3. Cholesky factorization and operator-adapted wavelets. Cholesky factorization in the maximin ordering is intimately related to computing operator-adapted wavelets. In section 5 we will use this connection to prove the accuracy of our approximation.

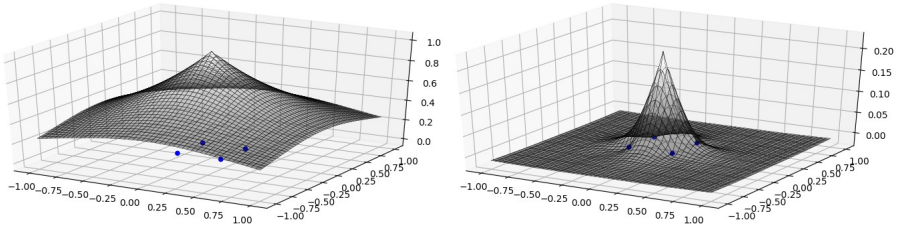

FIG. 3.2. *Left: covariance between a single site of a Matérn field with the values at the remaining sites. Right: conditional covariance given the values at the sites marked in blue. The conditional covariance decays significantly faster.*

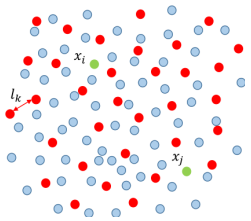

FIG. 3.3. *Step k of the Cholesky factorization in the ordering described in section 2. The red points have already been eliminated and form a covering of radius $l[k]$. The separation of the green points x_i and x_j by points that have already been eliminated implies the weak correlation between X_{x_i} and X_{x_j} conditional on $\{X_{x_i}\}_{i \preceq k}$.*

Operator-adapted wavelets. [62] and [63] introduced a novel class of operator-adapted wavelets called *gamblets* (see also [64]). For an operator $\mathcal{L}$ defined as in (2.1), gamblets can be identified as conditional expectations of the Gaussian process $\xi \sim \mathcal{N}\left(0, \mathcal{L}^{-1}\right)$. To construct the gamblets up to level $q \in \mathbb{N}$ we start with a hierarchy of *measurement functions* $\{\phi_i^{(k)}\}_{1 \leq k \leq q, i \in I^{(k)}} \subset H^{-s}(\Omega)$; heuristically, k labels a scale, and i a location at that scale. These measurement functions are linearly nested in the sense that, for $k < l$,

$$\phi_i^{(k)} = \sum_{j \in I^{(l)}} \pi_{i,j}^{(k,l)} \phi_j^{(l)} \tag{3.4}$$

for some rank-$|I^{(k)}|$ matrices $\pi^{(k,l)} \in \mathbb{R}^{I^{(k)} \times I^{(l)}}$. Writing $[\cdot, \cdot]$ for the duality product between $H^{-s}(\Omega)$ and $H_0^s(\Omega)$, the conditional expectations

$$\psi_i^{(k)} \coloneqq \mathbb{E}\left[\xi \,\middle|\, [\phi_j^{(k)}, \xi] = \delta_{ij} \ \forall j \in I^{(k)}\right] \quad \text{for } i \in I^{(k)} \tag{3.5}$$

act as $\mathcal{L}$-adapted prewavelets. These prewavelets can be identified as optimal recovery splines in the sense of [61] through the representation formula

$$\psi_i^{(k)} = \sum_{j \in I^{(k)}} \Theta_{i,j}^{(k),-1} \mathcal{L}^{-1} \phi_j^{(k)} \quad \text{for } i \in I^{(k)}, \tag{3.6}$$

where $\Theta_{i,j}^{(k),-1}$ is the (i,j)th entry of the inverse $\Theta^{(k),-1}$ of the matrix $\Theta^{(k)} \in \mathbb{R}^{I^{(k)} \times I^{(k)}}$ with entries $\Theta_{i,j}^{(k)} \coloneqq \int_\Omega \phi_i^{(k)} \mathcal{L}^{-1} \phi_j^{(k)} \, \mathrm{d}x$. The linear nesting of the $\phi_i^{(k)}$ across scales implies that the linear spaces $\mathfrak{V}^{(k)} \coloneqq \operatorname{span}\{\psi_i^{(k)} \mid i \in I^{(k)}\}$ are nested (i.e., $\mathfrak{V}^{(k-1)} \subset$

FIG. 3.4. *From left to right: an exemplary $\phi_i^{(k)}$, the corresponding $\psi_i^{(k)}$, a $\phi_j^{(k),W}$, and the corresponding $\chi_j^{(k)}$, all in the setting of $d = 1$.*

$\mathfrak{V}^{(k)}$). The multiresolution decomposition $\mathfrak{V}^{(q)} \coloneqq \mathfrak{V}^{(1)} \oplus \mathfrak{W}^{(2)} \oplus \cdots \oplus \mathfrak{W}^{(q)}$ is then obtained by defining $\mathfrak{W}^{(k)}$ as the orthogonal complement $\mathfrak{W}^{(k)}$ of $\mathfrak{V}^{(k-1)}$ in $\mathfrak{V}^{(k)}$ with respect to the energy scalar product $\langle u, v\rangle \coloneqq \int_\Omega u\mathcal{L}v \,\mathrm{d}x$. Basis functions for $\mathfrak{W}^{(k)}$ are identified (for $2 \leq k \leq q$) by

$$\chi_i^{(k)} \coloneqq \sum_j W_{ij}^{(k)} \psi_j^{(k)} \quad \text{for } i \in J^{(k)}, \tag{3.7}$$

or, equivalently, by

$$\chi_i^{(k)} \coloneqq \mathbb{E}\left[\xi \,\middle|\, \left[\phi_j^{(k),W}, \xi\right] = \delta_{ij}\delta_{kl} \;\forall 1 \leq l \leq k, j \in J^{(l)}\right] \quad \text{for } i \in J^{(k)}, \tag{3.8}$$

with $\phi_i^{(k),W} \coloneqq \sum_{j \in I^{(k)}} W_{i,j}^{(k)} \phi_j^{(k)}$, where $J^{(k)} \cong \left(I^{(k)} \setminus I^{(k-1)}\right)$ and $W^{(k)}$ is a $J^{(k)} \times I^{(k)}$ matrix such that $\operatorname{Im} W^{(k),\top} = \operatorname{Ker} \pi^{(k-1,k)}$ (writing $W^{(k),\top}$ for the transpose of $W^{(k)}$). See Figure 3.4 for an illustration.

For simplicity we write $J^{(1)} \coloneqq I^{(1)}$ and $\chi_i^{(1)} \coloneqq \psi_i^{(1)}$. Write $B^{(k)}$ for the $J^{(k)} \times J^{(k)}$ stiffness matrices $B^{(k)} \coloneqq \left\langle \chi_i^{(k)}, \chi_j^{(k)} \right\rangle$. The gamblets $\chi_i^{(k)}$ are $\mathcal{L}$-adapted wavelets in the sense that, under sufficient conditions on the $\phi_i^{(k)}$, they satisfy the following three properties:

- *Scale orthogonality in the energy scalar product*, i.e.,

$$\left\langle \chi_i^{(k)}, \chi_j^{(l)} \right\rangle = 0 \text{ for } l \neq k \text{ and } (i,j) \in J^{(k)} \times J^{(l)}\,. \tag{3.9}$$

  This leads to the block-diagonalization of the operator (with the $B^{(k)}$ as diagonal blocks).
- *Uniform Riesz stability* in the energy norm: the condition numbers of the blocks $B^{(k)}$ are uniformly bounded in k.
- *Exponential decay*, which leads to sparse blocks $B^{(k)}$: the gamblets $\chi_i^{(k)}$ exhibit exponential decay on the scale associated with k.

Although the scale-orthogonality property (3.9) is always satisfied, the two others (exponential decay and uniform Riesz stability) depend on the properties of $\mathcal{L}$ and the $\phi_i^{(k)}$. In the setting of the localization of numerical homogenization basis functions (where $\mathcal{L}$ is an elliptic PDE and the measurements $\phi_i^{(k)}$ are local and possibly not explicitly introduced), rigorous exponential decay estimates were pioneered in [57] and generalized in [50, 62, 44, 63]; see subsection 5.3.2 for detailed comparisons. For $\phi_i^{(k)}$ spanning the space of local polynomials of order up to $s-1$, bounded condition numbers are shown by [62, 63]. The homogenization results obtained in the special case $q = 2$ [57, 65, 44] are closely related to the lower bound on the spectrum of $B^{(2)}$ (see subsection 5.3.3).

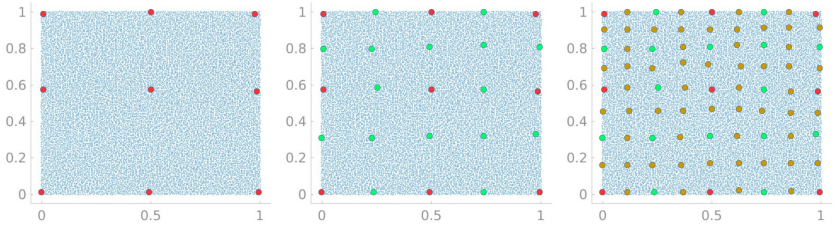

FIG. 3.5. *The implicit hierarchy of the maximin ordering. The maximin ordering has a hidden hierarchical structure, which can be discovered by picking a scale factor* $h \in (0,1)$ *and defining* $J^{(k)} \coloneqq \left\{ j \in I \,\middle|\, h^k \leq l[i]/l[1] < h^{k-1} \right\}$ *for* $1 \leq k \leq q \coloneqq \lceil \log_h (l[k]/l[1]) \rceil$. *In the figure, we see* $J^{(1)}$ *in red,* $J^{(2)}$ *in green, and* $J^{(3)}$ *in brown, for* $h = 1/2$.

Relation to Cholesky factorization. To explain the connection between gamblets and Cholesky factorization, let $J \coloneqq J^{(1)} \cup \cdots \cup J^{(q)}$, let $W^{(1)}$ be the $I^{(1)} \times I^{(1)}$ identity matrix, let $\pi^{(k,k)}$ be the $I^{(k)} \times I^{(k)}$ identity matrix, and let $\bar{\Theta}$ be the $J \times J$ symmetric matrix with the $J^{(k)} \times J^{(l)}$ block defined for $k \leq l$ by

$$\bar{\Theta}_{k,l} \coloneqq W^{(k)} \Theta^{(k)} \pi^{(k,l)} W^{(l),T}, \tag{3.10}$$

or equivalently by

$$\left(\bar{\Theta}_{k,l}\right)_{ij} \coloneqq \left[\phi_i^{(k),W}, \mathcal{L}^{-1} \phi_j^{(l),W}\right]. \tag{3.11}$$

Then, the block-Cholesky factorization of $\bar{\Theta}$ satisfies the identity

$$\bar{\Theta} = \bar{L} D \bar{L}^{\top}, \tag{3.12}$$

where D is a block-diagonal matrix with the $J^{(k)} \times J^{(k)}$ diagonal block equal to $B^{(k),-1}$ and

$$\bar{L}_{i,j} \coloneqq \begin{cases} \delta_{i,j} & \text{if } i,j \in J^{(k)}, \\ 0 & \text{if } i \in J^{(k)}, j \in J^{(k')}, \text{ and } k' > k, \\ [\phi_i^{(k)}, \chi_j^{(k')}] & \text{if } i \in J^{(k)}, j \in J^{(k')}, \text{ and } k' < k. \end{cases} \tag{3.13}$$

Therefore, computing gamblets associated to the operator $\mathcal{L}$ and measurement functions ϕ_i is equivalent to computing a block-Cholesky factorization of Θ in the multiresolution basis given by the $\phi_i^{(k),W}$.

The Cholesky decomposition of Θ (1.1) belongs to this setting. Indeed, although the maximin ordering of section 2 has no explicit multiscale structure, this structure can be introduced, as described in Figure 3.5, by decomposing $x_1, \ldots, x_N$ into a nested hierarchy $\{x_i\}_{i \in I^{(1)}} \subset \{x_i\}_{i \in I^{(2)}} \subset \cdots \subset \{x_i\}_{i \in I^{(q)}}$, and choosing $\phi_i^{(k)} = \boldsymbol{\delta}(\cdot - x_i)$ for $i \in I^{(k)}$ and $k \in \{1, \ldots, q\}$, where $\boldsymbol{\delta}$ denotes the unit (unscaled) Dirac delta function. Under this choice, $\pi_{i,j}^{(k,k+1)} = 1$ for $j \in I^{(k)}$ and $\pi_{i,j}^{(k,k+1)} = 0$ for $j \notin I^{(k)}$. Letting $J^{(k)}$ label the indices in $I^{(k)}/I^{(k-1)}$ and choosing $W_{i,j}^{(k)} = 1$ for $j \in I^{(k)}/I^{(k-1)}$ and $W_{i,j}^{(k)} = 0$ for $j \in I^{(k-1)}$ implies $\Theta = \bar{\Theta}$. The exponential decay of $\bar{L}$ and D^{-1} follows from known results [63] on exponential decay of the $\chi_j^{(k)}$. The uniform bound on the condition number of the $B^{(k)}$ is proved in subsection 5.3.3. The exponential decay and uniform bound on the condition numbers of the blocks $B^{(k)}$ imply the exponential decay of the Cholesky factors $\hat{L}$ of D and hence of $L = \bar{L}\hat{L}$. The approximation error estimate (2.4) is then obtained by matching the sparsity set S with the near-sparse structure of L.

Algorithm 4.1 Ordering and sparsity pattern algorithm.

Input: Real $\rho \geq 2$ and Oracles $\texttt{dist}(\cdot,\cdot), \texttt{dist}_{\partial\Omega}(\cdot)$ such that $\texttt{dist}(i,j) = \operatorname{dist}(x_i, x_j)$ and $\texttt{dist}_{\partial\Omega}(i) = \operatorname{dist}(x_i, \partial\Omega)$

Output: An array $l[:]$ of distances, an array P encoding the multiresolution ordering, and an array of index pairs S containing the sparsity pattern.

$P = \emptyset$
for $i \in \{1, \ldots, N\}$ **do**
  $l[i] \leftarrow \texttt{dist}_{\partial\Omega}(i)$
  $p[i] \leftarrow \emptyset$
  $c[i] \leftarrow \emptyset$
end for
{Creates a mutable binary heap, containing pairs of indices and distances as elements:}
$H \leftarrow \texttt{MutableMaximalBinaryHeap}\left(\{(i, l[i])\}_{i\in\{1,\ldots,N\}}\right)$
{Instates the Heap property, with a pair with maximal distance occupying the root of the heap:}

$\texttt{heapSort!}(H)$
{Processing the first index:}
{Get the root of the heap, remove it, and restore the heap property:}
$(i, l) = \texttt{pop}(H)$
{Add the index as the next element of the ordering} $\texttt{push}\,(P, i)$
for $j \in \{1, \ldots, N\}$ **do**
  $\texttt{push}(c[i], j)$
  $\texttt{push}(p[j], i)$
  $\texttt{sort!}\,(c[i], \texttt{dist}(\cdot, i))$
  $\texttt{decrease!}\,(H, j, \texttt{dist}(i, j))$
end for
{Processing remaining indices:}
while $H \neq \emptyset$ **do**
  {Get the root of the heap, remove it, and restore the heap property:} $(i, l) = \texttt{pop}(H)$ $l[i] \leftarrow l$
  {Select the parent node that has all possible children of i amongst its children, and is closest to i:}
  $k = \arg\min_{j\in p[i]:\texttt{dist}(i,j)+\rho l[i]\leq\rho l[j]} \texttt{dist}\,(i, j)$
  {Loop through those children of k that are close enough to k to possibly be children of i:}
  for $j \in c[k] : \texttt{dist}(j, k) \leq \texttt{dist}(i, k) + \rho l[i]$ **do**
    $\texttt{decrease!}\,(H, j, \texttt{dist}(i, j))$
    if $\texttt{dist}(i, j) \leq \rho l[i]$ **then**
      $\texttt{push}(c[i], j)$
      $\texttt{push}(p[j], i)$
    end if
  end for
  {Add the index as the next element of the ordering}
  $\texttt{push}\,(P, i)$
  {Sort the children according to distance to the parent node, so that the closest children can be found more easily} $\texttt{sort!}\,(c[i], \texttt{dist}(\cdot, i))$
end while
{Aggregating the lists of children into the sparsity pattern:}
for $i \in \{1, \ldots, N\}$ **do**
  for $j \in c[i]$ **do**
    $\texttt{push!}\,(S, (i, j))$
    $\texttt{push!}\,(S, (j, j))$
  end for
end for

4. Implementation and numerical results.

4.1. Selection of the sparsity pattern and ordering. This section introduces an $\mathcal{O}(\rho^d N \log^2 N)$-complexity algorithm (Algorithm 4.1) for selecting the sparsity pattern and ordering used as inputs in Algorithm 2.2. This algorithm does not

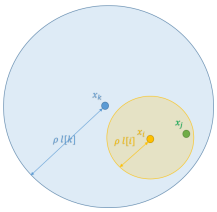

FIG. 4.1. *Localization of computation in Algorithm* 4.1 *based on hierarchy. When adding* i *to the ordering, only consider indices* j *such that* $\operatorname{dist}(x_i, x_j) \leq \rho l[i]$. *Those indices are a subset of the* children *of the coarse-level index* k *if* $\rho l[k] \geq \operatorname{dist}(x_i, x_j) + \rho l[i]]$. *Thus, the search for candidates* j *can be restricted to those children of* k.

explicitly query the position of the $\{x_i\}_{i\in I}$ and only uses pairwise distances by processing points one by one by updating a mutable binary heap, keeping track of the point to be processed at each step. With this approach, our proposed algorithm is oblivious to the dimension d of the ambient space and, in particular, can automatically exploit low-dimensional structure in the point cloud $\{x_i\}_{i\in I}$. In order to avoid computing all $\mathcal{O}(N^2)$ pairwise distances, as illustrated in Figure 4.1, Algorithm 4.1 uses the sparsity pattern obtained on the coarser scales to restrict computation at the finer scales to local neighborhoods.

THEOREM 4.1. *The output of Algorithm* 4.1 *is the ordering and sparsity pattern described in section* 2. *Furthermore, in the setting of Theorem* 3.1, *if the oracles* $\texttt{dist}(\cdot,\cdot)$ *and* $\texttt{dist}_{\partial\Omega}(\cdot)$ *can be queried in complexity* $\mathcal{O}(1)$, *then the complexity of Algorithm* 4.1 *is bounded by* $C\rho^d N \log^2 N$, *where* C *is a constant depending only on* d, Ω, *and* δ.

Theorem 4.1 is proved in section SM1. As discussed therein, in the case $\Omega = \mathbb{R}^d$, Algorithm 4.1 has the advantage that its computational complexity depends only on the intrinsic dimension of the dataset, which can be much smaller than d.

4.2. The case of the whole space ($\Omega = \mathbb{R}^d$). Many applications in Gaussian process statistics and machine learning are in the $\Omega = \mathbb{R}^d$ setting. In that setting, the Matérn family of kernels (4.5) is a popular choice that is equivalent to using the whole-space Green's function of an elliptic PDE as covariance function [80, 81]. Let $\bar{\Omega}$ be a bounded domain containing the $\{x_i\}_{i\in I}$. The case $\Omega = \mathbb{R}^d$ is not covered in Theorem 3.1 because in this case the screening effect is weakened near the boundary of $\bar{\Omega}$ by the absence of measurement points outside of $\bar{\Omega}$. Therefore, distant points close to the boundary of $\bar{\Omega}$ will have stronger conditional correlations than similarly distant points in the interior of $\bar{\Omega}$ (see Figure 4.2). As observed by [68] and [19], Markov random field (MRF) approaches that use a discretization of the underlying PDE face similar challenges at the boundary. While the weakening of the exponential decay at the boundary worsens the accuracy of our method, the numerical results of subsection 4.4 (which are all obtained without imposing boundary conditions) suggest that its overall impact is limited. In particular, as shown in Figure 4.2, it does not cause significant artifacts in the quality of the approximation near the boundary. This differs from the significant boundary artifacts of MRF methods, which have to be mitigated against by a careful calibration of boundary conditions [68, 19]. Although the numerical results presented in this section are mostly obtained with $x_i \sim \mathrm{UNIF}([0,1]^d)$, in many practical applications, the density of measurement points will slowly (rather than abruptly) decrease toward zero near the boundary of the sampled domain, which drastically de-

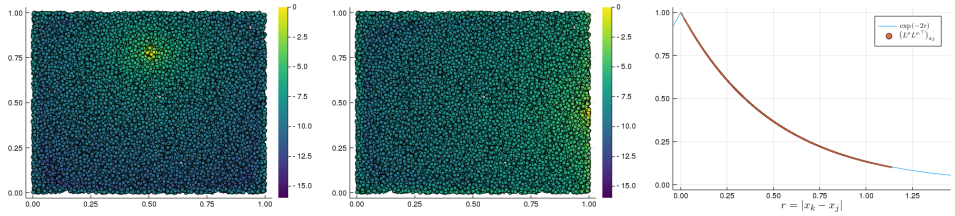

FIG. 4.2. *Weaker screening between boundary points. Left and center: ith (left) and jth (center) column of the Cholesky factor L (normalized to unit diagonal) of* Θ *in maximin ordering, where* x_i *is an interior point and* x_j *is near the boundary. Although* $l[i]$ *is of the order of* $l[j]$*, the exponential decay of* $L_{:,j}$ *near the boundary is significantly weakened by the absence of Dirichlet boundary conditions. Right: approximate correlations* $\{(L^\rho L^{\rho,\top})_{kj}\}_{k\in I}$ *(with* $\rho = 3.0$*) and true covariance function* $\exp(-2r)$ *with* $r = |x_k - x_j|$*. Correlations between* x_j *and remaining points are captured accurately, despite the weakened exponential decay near the boundary.*

creases the boundary errors shown above. Accuracy can also be enhanced by adding *artificial* points $\{x_i\}_{i\in\tilde{I}}$ at the boundary. By applying the Cholesky factorization to $\{x_i\}_{i\in I\cup\tilde{I}}$, and then restricting the resulting matrix to $I \times I$, we can obtain a very accurate approximate matrix-vector multiplication. Although not in the form of a Cholesky factorization, this approximation can be efficiently inverted using iterative methods such as conjugate gradient [76] preconditioned with the Cholesky factorization obtained from the original set of points.

4.3. Nuggets and measurement errors. In the Gaussian process regression setting it is common to model measurement error by adding a *nugget* $\sigma^2\mathrm{Id}$ to the covariance matrix:

$$\tilde{\Theta} = \Theta + \sigma^2\mathrm{Id}. \tag{4.1}$$

The addition of a diagonal matrix diminishes the screening effect and thus the accuracy of Algorithm 2.2. This problem can be avoided by rewriting the modified covariance matrix $\tilde{\Theta}$ as

$$\tilde{\Theta} = \Theta(\sigma^2 A + \mathrm{Id}), \tag{4.2}$$

where $A \coloneqq \Theta^{-1}$. As noted in subsection 6.2, A can be interpreted as a discretized partial differential operator and has near-sparse Cholesky factors in the reverse elimination ordering. Adding a multiple of the identity to A amounts to adding a zeroth-order term to the underlying PDE and thus preserves the sparsity of the Cholesky factors. This leads to the sparse decomposition

$$\tilde{\Theta} = LL^\top P^\updownarrow \tilde{L}\tilde{L}^\top P^\updownarrow, \tag{4.3}$$

where $P^\updownarrow$ is the order-reversing permutation and $\tilde{L}$ is the Cholesky factor of $P^\updownarrow(\sigma^2 A + \mathrm{Id})P^\updownarrow$. Figure 4.3 shows that the exponential decay of these Cholesky factors is robust with respect σ.

This idea can be turned into an algorithm by first approximately computing L using Algorithm 2.2; then using L to approximate A, which can be done in near-linear complexity by exploiting sparsity; and then approximating $\tilde{L}$, again using Algorithm 2.2. While this algorithm is asymptotically efficient, our preliminary results suggest that the additional inversion step significantly increases the constants featured in the approximation accuracy. Therefore, when low accuracy is sufficient, we instead

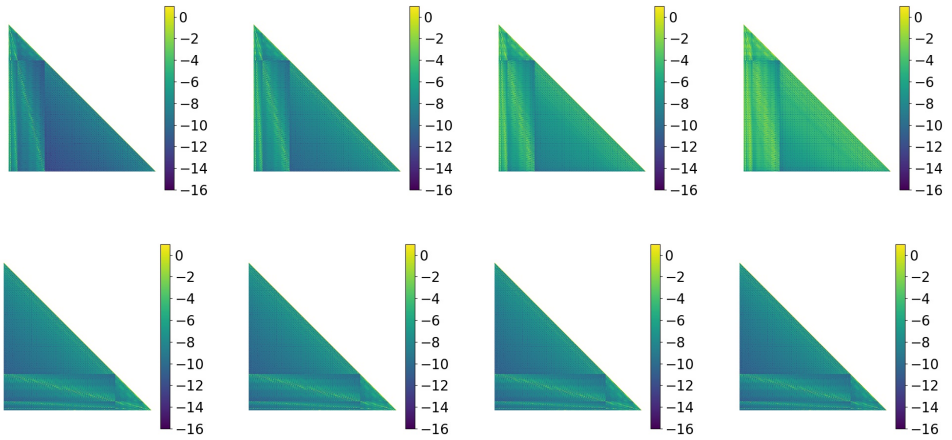

FIG. 4.3. *(Lack of) robustness to varying size of the nugget: We plot the* $\log_{10}$ *of the magnitude of the Cholesky factors of* $\Theta+\sigma^2\mathrm{Id}$ *in maximin ordering (first row) and of* $A+\sigma^2$ *in reverse maximin ordering (second row). As we increase* $\sigma^2 \in [0.0, 0.1, 1.0, 10.0]$ *from left to right the decay of the Cholesky factors of* $\Theta + \sigma^2\mathrm{Id}$ *deteriorates and that of the factors of* $A + \sigma^2\mathrm{Id}$ *is preserved.*

recommend simply applying Algorithm 2.2 to the matrix Θ. This preserves the original approximation accuracy and the matrix inversion can then efficiently be performed using iterative methods such as conjugate gradient (CG) [76] by taking advantage of the fast matrix-vector multiplication obtained from the sparse factorization. For small values of σ (which would lead to slow convergence of CG) we can directly apply Algorithm 2.2 to $\tilde{\Theta}$. For large values of σ, $\tilde{\Theta}$ will be well conditioned and the convergence of CG is fast. For intermediate values of σ, we can apply Algorithm 2.2 to $\tilde{\Theta}$ and use the resulting factors as a preconditioner for CG. Sampling from $\mathcal{N}(0, \tilde{\Theta})$ can be done by adding independent samples from $\mathcal{N}(0, \Theta)$ and $\mathcal{N}(0, \sigma^2\mathrm{Id})$. Approximations of the log-determinant could be obtained either by applying Algorithm 2.2 directly to $\tilde{\Theta}$ (with some loss of accuracy) or by combining iterative methods [72, 25] with the fast matrix-vector multiplication obtained from the sparse factorization of Θ. Just like CG, these methods benefit from the fact that we can work with well-conditioned matrices for small and large σ. A detailed investigation of the efficiency of the above mentioned strategies for computing with nuggets is beyond the scope of this work.

4.4. Numerical results. We will now present numerical evidence in support of our results. All experiments reported below were run on a workstation using an Intel Core i7-6400 CPU with 4.00GHz and 64 GB of RAM. The time-critical parts of the code are run on a single thread, leaving the exploration of parallelism to future work. The Julia scripts implementing the experiments can be found online under https://github.com/f-t-s/nearLinKernel. In the following, $\mathrm{nnz}(L)$ denotes the number of nonzero entries of the lower-triangular factor L; $t_{\texttt{SortSparse}}$ denotes the time taken by Algorithm 4.1 to compute the maximin ordering $\prec$ and sparsity pattern S_ρ; $t_{\texttt{Entries}}$ denotes the time taken to compute the entries of Θ on S_ρ; and $t_{\texttt{ICHOL(0)}}$ denotes the time taken to perform Algorithm 2.2 (`ICHOL(0)`), all measured in seconds. The relative error in Frobenius norm is approximated by

$$E \coloneqq \frac{\|LL^\top - \Theta\|_{\mathrm{Fro}}}{\|\Theta\|_{\mathrm{Fro}}} \approx \frac{\sqrt{\sum_{k=1}^{m} \left\| \left(LL^\top - \Theta\right)_{i_k j_k} \right\|^2}}{\sqrt{\sum_{k=1}^{m} \|\Theta_{i_k j_k}\|^2}}, \tag{4.4}$$

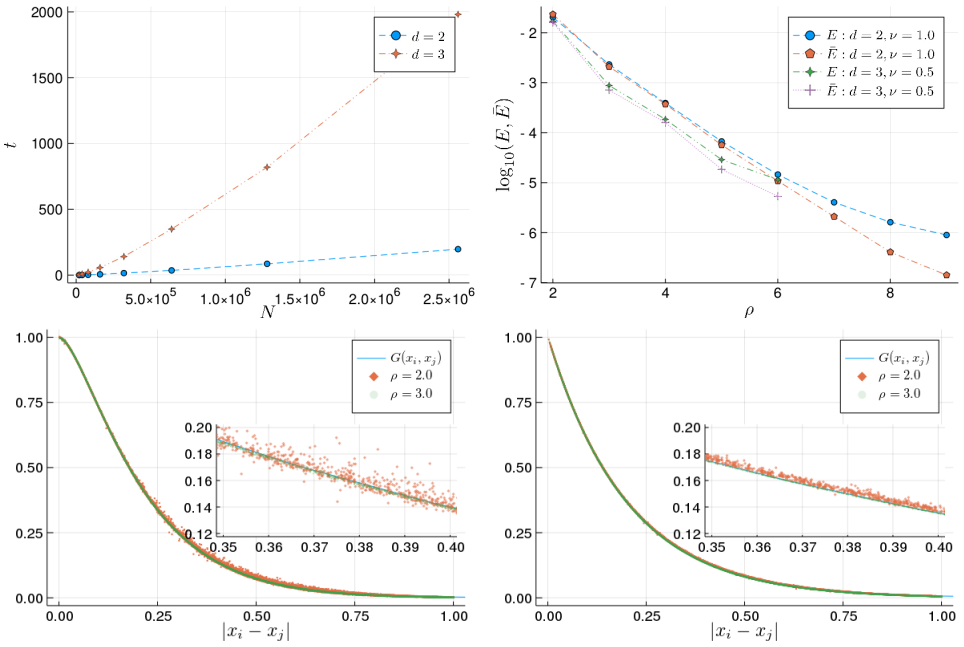

FIG. 4.4. *First panel: the increase in computational time taken by the Cholesky factorization, as N increases (for $\rho = 3.0$). Second panel: the exponential decay of the relative error in Frobenius norm, as ρ is increased. In the third ($d = 2$) and fourth panel ($d = 3$), we see the comparison of the approximate and true covariance for $\rho = 2.0$ and $\rho = 3.0$.*

where the $m = 500000$ pairs of indices $i_k, j_k \sim \mathrm{UNIF}(I)$ are independently and uniformly distributed in I. This experiment is repeated 50 times and the resulting mean and standard deviation (in brackets) are reported. For measurements in $[0,1]^d$, in order to isolate the boundary effects, we also consider the quantity $\bar{E}$ which is defined as E, but with only those samples i_k, j_k for which $x_{i_k}, x_{j_k} \in [0.05, 0.95]^d$. Most of our experiments will use the Matérn class of covariance functions [59], defined by

$$G_{l,\nu}^{\text{Matérn}}(x,y) \coloneqq \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|x-y|}{l} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}|x-y|}{l} \right), \tag{4.5}$$

where K_ν is the modified Bessel function of second kind [1, section 9.6] and ν, l are parameters describing the degree of smoothness, and the length-scale of interactions, respectively [67]. In Figure 4.5, the Matérn kernel is plotted for different degrees of smoothness. The Matérn covariance function is used in many branches of statistics and machine learning to model random fields with finite order of smoothness [35, 67].

As observed by [80, 81], the Matérn kernel is the Green's function of an elliptic PDE of possibly fractional order $2(\nu + d/2)$ in the whole space. Therefore, for $2(\nu + d/2) \in \mathbb{N}$, the Matérn kernel falls into the framework of our theoretical results, up to the behavior at the boundary discussed in subsection 4.2. Since the locations of our points will be chosen at random, some of the points will be very close to each other, resulting in an almost singular matrix Θ that can become nonpositive under the approximation introduced by `ICHOL(0)`. If Algorithm 2.2 encounters a nonpositive pivot A_{ii}, then we set the corresponding column of L to zero, resulting in a low-rank approximation of the original covariance matrix. We report the rank of L in our

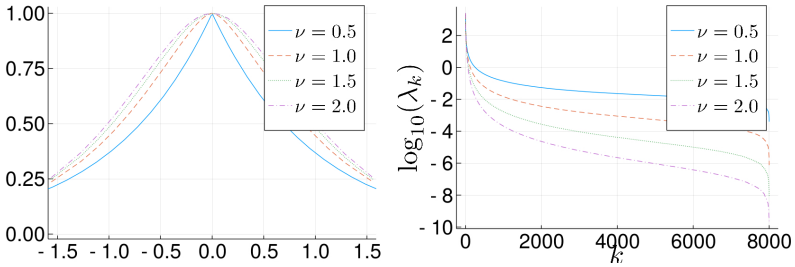

FIG. 4.5. *Matérn kernels for different values of ν (left), and the spectrum of Θ, for 2000 points $x_i \in [0,1]^2$ (right). Smaller values of ν correspond to stronger singularities at zero and hence lower degrees of smoothness of the associated Gaussian process.*

TABLE 4.1
$G_{\nu,l}^{Matérn}$ *with* $\nu = 0.5$, $l = 0.2$, $\rho = 3.0$, *and* $d = 2$.

N	$\mathrm{nnz}(L)/N^2$	$\mathrm{rank}(L)$	$t_{\texttt{SortSparse}}$	$t_{\texttt{Entries}}$	$t_{\texttt{ICHOL(0)}}$	E	$\bar{E}$
20000	5.26e-03	20000	0.71	0.81	0.42	1.25e-03 (3.68e-06)	1.11e-03 (3.01e-06)
20000	5.26e-03	20000	0.71	0.81	0.42	1.25e-03 (3.68e-06)	1.11e-03 (3.01e-06)
40000	2.94e-03	40000	1.21	1.19	1.00	1.27e-03 (3.32e-06)	1.12e-03 (3.56e-06)
80000	1.62e-03	80000	2.72	2.82	2.55	1.30e-03 (3.20e-06)	1.21e-03 (3.29e-06)
160000	8.91e-04	160000	6.86	6.03	6.11	1.28e-03 (3.57e-06)	1.16e-03 (3.32e-06)
320000	4.84e-04	320000	17.22	13.79	15.66	1.23e-03 (3.19e-06)	1.11e-03 (2.40e-06)
640000	2.63e-04	640000	41.40	31.02	36.02	1.24e-03 (2.58e-06)	1.09e-03 (3.02e-06)
1280000	1.41e-04	1280000	98.34	65.96	85.99	1.23e-03 (3.72e-06)	1.10e-03 (3.74e-06)
2560000	7.55e-05	2560000	233.92	148.43	197.52	1.16e-03 (2.82e-06)	1.04e-03 (3.36e-06)

TABLE 4.2
$G_{\nu,l}^{Matérn}$ *with* $\nu = 0.5$, $l = 0.2$, $\rho = 3.0$, *and* $d = 3$.

N	$\mathrm{nnz}(L)/N^2$	$\mathrm{rank}(L)$	$t_{\texttt{SortSparse}}$	$t_{\texttt{Entries}}$	$t_{\texttt{ICHOL(0)}}$	E	$\bar{E}$
20000	1.30e-02	20000	1.61	1.44	2.94	1.49e-03 (5.00e-06)	1.20e-03 (5.09e-06)
40000	7.60e-03	40000	3.26	3.32	8.33	1.21e-03 (4.29e-06)	9.91e-04 (3.72e-06)
80000	4.35e-03	80000	7.46	7.64	22.46	1.06e-03 (3.74e-06)	8.51e-04 (2.93e-06)
160000	2.45e-03	160000	20.95	18.42	57.64	9.81e-04 (2.33e-06)	7.88e-04 (3.23e-06)
320000	1.37e-03	320000	53.58	40.72	141.46	9.27e-04 (2.26e-06)	7.53e-04 (2.72e-06)
640000	7.61e-04	640000	133.55	96.67	350.10	8.98e-04 (3.25e-06)	7.25e-04 (3.02e-06)
1280000	4.19e-04	1280000	312.43	212.57	820.07	8.59e-04 (2.79e-06)	7.00e-04 (2.87e-06)
2560000	2.29e-04	2560000	795.68	480.17	1981.92	8.96e-04 (2.76e-06)	7.73e-04 (4.28e-06)

experiments and note that we obtain a full-rank approximation for moderate values of ρ.

We begin by investigating the scaling of our algorithm as N increases. To this end, we consider $\nu = 0.5$ (the exponential kernel), $l = 0.2$, and choose N randomly distributed points in $[0,1]^d$ for $d \in \{2,3\}$. The results are summarized in Tables 4.1 and 4.3, and in Figure 4.4 and they confirm the near-linear computational complexity of our algorithm.

Next, we investigate the trade-off between computational efficiency and accuracy of the approximation. To this end, we choose $d = 2$, $\nu = 1.0$ and $d = 3$, $\nu = 0.5$, corresponding to fourth-order equations in two and three dimensions. We choose $N = 10^6$ data points $x_i \sim \mathrm{UNIF}([0,1]^d)$ and apply our method with different values of ρ. The results of these experiments are tabulated in Tables 4.3 and 4.4 and the impact of ρ on the approximation error is visualized in Figure 4.4.

TABLE 4.3
$G_{\nu,l}^{Matérn}$ *with* $\nu = 1.0$, $l = 0.2$, $N = 10^6$, *and* $d = 2$.

	$\mathrm{nnz}(L)/N^2$	$\mathrm{rank}(L)$	$t_{\texttt{SortSparse}}$	$t_{\texttt{Entries}}$	$t_{\texttt{ICHOL(0)}}$	E	$\bar{E}$
$\rho = 2.0$	8.78e-05	254666	38.06	33.72	17.54	2.04e-02 (1.73e-02)	2.34e-02 (2.75e-02)
$\rho = 3.0$	1.76e-04	964858	71.07	67.85	61.35	2.32e-03 (6.02e-06)	2.09e-03 (7.50e-06)
$\rho = 4.0$	2.90e-04	999810	115.07	112.56	152.93	3.92e-04 (1.44e-06)	3.72e-04 (2.32e-06)
$\rho = 5.0$	4.26e-04	999999	165.91	166.60	312.19	6.70e-05 (2.98e-07)	5.68e-05 (2.55e-07)
$\rho = 6.0$	5.83e-04	1000000	227.62	229.76	566.94	1.45e-05 (6.69e-08)	1.08e-05 (5.01e-08)
$\rho = 7.0$	7.59e-04	1000000	292.52	300.65	944.33	4.05e-06 (4.96e-08)	2.10e-06 (1.69e-08)
$\rho = 8.0$	9.53e-04	1000000	363.90	380.07	1476.71	1.62e-06 (2.30e-08)	4.08e-07 (9.47e-09)
$\rho = 9.0$	1.16e-03	1000000	447.47	467.07	2200.32	8.98e-07 (1.44e-08)	1.42e-07 (5.14e-09)

TABLE 4.4
$G_{\nu,l}^{Matérn}$ *with* $\nu = 0.5$, $l = 0.2$, $N = 10^6$, *and* $d = 3$.

	$\mathrm{nnz}(L)/N^2$	$\mathrm{rank}(L)$	$t_{\texttt{SortSparse}}$	$t_{\texttt{Entries}}$	$t_{\texttt{ICHOL(0)}}$	E	$\bar{E}$
$\rho = 2.0$	1.87e-04	998046	87.83	56.44	85.20	1.69e-02 (6.89e-04)	1.60e-02 (3.36e-04)
$\rho = 3.0$	5.17e-04	1000000	226.84	158.42	599.86	8.81e-04 (3.21e-06)	7.15e-04 (2.99e-06)
$\rho = 4.0$	1.05e-03	1000000	446.52	326.27	2434.52	1.85e-04 (5.37e-07)	1.59e-04 (5.30e-07)
$\rho = 5.0$	1.82e-03	1000000	747.65	567.06	7227.45	2.89e-05 (1.94e-07)	1.84e-05 (1.15e-07)
$\rho = 6.0$	2.82e-03	1000000	1344.59	928.27	17640.58	1.15e-05 (1.06e-07)	5.34e-06 (5.34e-08)

TABLE 4.5
We tabulate the approximation rank and error for $\rho = 5.0$ *and* $N = 10^6$ *points uniformly distributed in* $[0,1]^3$. *The covariance function is* $G_{\nu,0.2}^{Matérn}$ *for* ν *ranging around* $\nu = 0.5$ *and* $\nu = 1.5$. *Even though the intermediate values of* ν *correspond to a fractional-order elliptic PDE, the behavior of the approximation stays the same.*

	$\nu = 0.3$	$\nu = 0.5$	$\nu = 0.7$	$\nu = 0.9$	$\nu = 1.1$	$\nu = 1.3$	$\nu = 1.5$	$\nu = 1.7$
$\mathrm{rank}(L)$	1000000	1000000	1000000	1000000	1000000	1000000	1000000	999893
E	7.04e-05	2.89e-05	2.49e-05	3.58e-05	6.03e-05	8.77e-05	1.18e-04	1.46e-04
	(3.98e-07)	(1.79e-07)	(1.11e-07)	(1.19e-07)	(2.37e-07)	(3.06e-07)	(4.52e-07)	(5.39e-07)
$\bar{E}$	5.19e-05	1.85e-05	1.77e-05	2.82e-05	4.88e-05	6.87e-05	9.06e-05	1.13e-04
	(2.26e-07)	(1.18e-07)	(8.11e-08)	(1.30e-07)	(2.37e-07)	(3.50e-07)	(5.14e-07)	(5.45e-07)

TABLE 4.6
$G_{l,\alpha,\beta}^{Cauchy}$ *for* $(l,\alpha,\beta) = (0.4, 0.5, 0.025)$ *(first table) and* $(l,\alpha,\beta) = (0.2, 1.0, 0.20)$ *(second table) for* $N = 10^6$ *and* $d = 2$.

	$\rho = 2.0$	$\rho = 3.0$	$\rho = 4.0$	$\rho = 5.0$	$\rho = 6.0$	$\rho = 7.0$	$\rho = 8.0$	$\rho = 9.0$
$\mathrm{rank}(L)$	999923	1000000	1000000	1000000	1000000	1000000	1000000	1000000
E	4.65e-04	5.98e-05	2.36e-05	1.19e-05	4.84e-06	4.17e-06	2.25e-06	1.42e-06
	(4.23e-07)	(1.56e-07)	(9.53e-08)	(6.32e-08)	(4.14e-08)	(4.99e-08)	(1.86e-08)	(1.64e-08)
$\bar{E}$	3.81e-04	3.49e-05	9.83e-06	4.65e-06	1.47e-06	8.49e-07	4.25e-07	2.12e-07
	(4.98e-07)	(1.59e-07)	(5.56e-08)	(2.63e-08)	(7.73e-09)	(1.04e-08)	(4.81e-09)	(3.24e-09)

	$\rho = 2.0$	$\rho = 3.0$	$\rho = 4.0$	$\rho = 5.0$	$\rho = 6.0$	$\rho = 7.0$	$\rho = 8.0$	$\rho = 9.0$
$\mathrm{rank}(L)$	999547	1000000	1000000	1000000	1000000	1000000	1000000	1000000
E	1.08e-03	1.36e-04	2.89e-05	2.35e-05	5.33e-06	3.25e-06	2.53e-06	1.68e-06
	(5.02e-06)	(6.27e-07)	(2.63e-07)	(3.01e-07)	(6.15e-08)	(5.74e-08)	(4.84e-08)	(4.25e-08)
$\bar{E}$	7.23e-04	8.96e-05	1.17e-05	5.65e-06	1.09e-06	5.84e-07	4.03e-07	2.40e-07
	(4.07e-06)	(2.63e-07)	(7.10e-08)	(1.47e-07)	(7.71e-09)	(5.48e-09)	(3.44e-09)	(2.23e-09)

While our theoretical results only cover integer-order elliptic PDEs, we observe no practical difference between the numerical results for Matérn kernels corresponding to integer- and fractional-order smoothness. As an illustration, for the case $d = 3$,

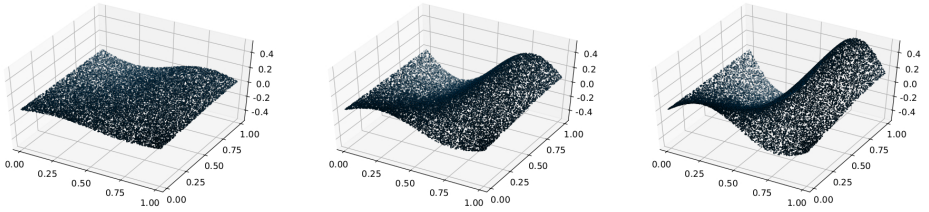

FIG. 4.6. *A two-dimensional point cloud deformed into a two-dimensional submanifold of* $\mathbb{R}^3$ *with* $\delta_z \in \{0.1, 0.3, 0.5\}$.

TABLE 4.7
$G_{\nu,l}^{Matérn}$ *for* $\nu = 0.5$, $l = 0.2$, *and* $\rho = 3.0$ *with* $N = 10^6$ *points chosen as in Figure* 4.6.

	$\delta_z = 0.0$	$\delta_z = 0.1$	$\delta_z = 0.2$	$\delta_z = 0.3$	$\delta_z = 0.4$	$\delta_z = 0.5$	$\delta_z = 0.6$
$\frac{\mathrm{nnz}(L)}{N^2}$	1.76e-04	1.77e-04	1.78e-04	1.80e-04	1.82e-04	1.84e-04	1.85e-04
$t_{\texttt{ICHOL(0)}}$	61.92	62.15	62.81	64.27	64.87	65.50	66.12
$\mathrm{rank}(L)$	1000000	1000000	1000000	1000000	1000000	1000000	1000000
E	1.17e-03	1.11e-03	1.28e-03	1.60e-03	1.72e-03	1.89e-03	2.11e-03
	(2.74e-06)	(3.00e-06)	(2.73e-06)	(4.28e-06)	(3.95e-06)	(5.11e-06)	(5.07e-06)

we provide approximation results for ν ranging around $\nu = 0.5$ (corresponding to a fourth-order elliptic PDE) and $\nu = 1.5$ (corresponding to a sixth-order elliptic PDE). As seen in Table 4.5, the results vary continuously as ν changes, with no qualitative differences between the behavior for integer- and fractional-order PDEs. To further illustrate the robustness of our method, we consider the Cauchy class of covariance functions introduced in [32]

$$G_{l,\alpha,\beta}^{\mathrm{Cauchy}}(x,y) \coloneqq \left(1 + \left(\frac{|x-y|}{l}\right)^{\alpha}\right)^{-\frac{\beta}{\alpha}}. \tag{4.6}$$

As far as we are aware, the Cauchy class has not been associated to an elliptic PDE. Furthermore, it does not have exponential decay in the limit $|x - y| \to \infty$, which allows us to emphasize the point that the exponential decay of the error is *not* due to the exponential decay of the covariance function itself. Table 4.6 gives the results for $(l, \alpha, \beta) = (0.4, 0.5, 0.025)$ and $(l, \alpha, \beta) = (0.2, 1.0, 0.2)$.

In Gaussian process regression, the ambient dimension d is typically too large to ensure computational efficiency of our algorithm. However, since our algorithm only requires access to pairwise distances between points, it can take advantage of the low intrinsic dimension of the dataset. We might be concerned that in this case, interaction through the higher-dimensional ambient space will disable the screening effect. As a first demonstration that this is not the case, we will draw $N = 10^6$ points in $[0,1]^2$ and equip them with a third component according to $x_i^{(3)} \coloneqq -\delta_z \sin(6x_i^{(1)})\cos(2(1 - x_i^{(2)})) + \xi_i 10^{-3}$ for ξ_i a standard Gaussian vector. Figure 4.6 shows the resulting point sets for different values of δ_z, and Table 4.7 shows that the approximation is robust to increasing values of δ_z.

An appealing feature of our method is that it can be formulated in terms of the pairwise distances alone. This means that the algorithm will automatically exploit any low-dimensional structure in the dataset. In order to illustrate this feature, we artificially construct a dataset with low-dimensional structure by randomly rotating four low-dimensional structures into a 20-dimensional ambient space (see Figure 4.7).

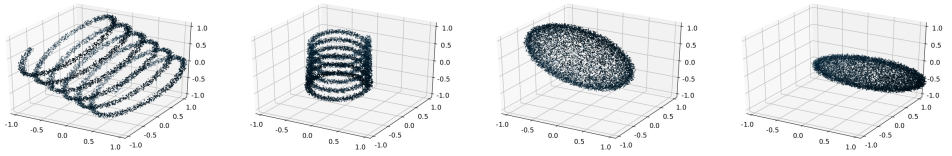

FIG. 4.7. *We construct a high-dimensional dataset with low-dimensional structure by rotating the above structures at random into a* 20*-dimensional ambient space.*

TABLE 4.8
$G_{\nu,l}^{Matérn}$ *for* $\nu = 0.5$*,* $l = 0.5$*, and* $N = 10^6$ *points as in Figure* 4.7*.*

	$\mathrm{nnz}(L)/N^2$	$\mathrm{rank}(L)$	$t_{\texttt{SortSparse}}$	$t_{\texttt{Entries}}$	$t_{\texttt{ICHOL(0)}}$	E
$\rho = 2.0$	1.62e-04	997635	80.60	57.11	52.49	1.57e-02 (1.13e-03)
$\rho = 3.0$	3.76e-04	1000000	173.86	135.61	248.78	2.88e-03 (1.14e-05)
$\rho = 4.0$	6.76e-04	1000000	302.98	247.74	748.62	8.80e-04 (4.97e-06)
$\rho = 5.0$	1.05e-03	1000000	462.98	397.42	1802.44	3.44e-04 (2.54e-06)
$\rho = 6.0$	1.49e-03	1000000	645.56	556.72	3696.31	1.44e-04 (8.76e-07)
$\rho = 7.0$	2.02e-03	1000000	891.08	758.88	6855.23	7.61e-05 (5.66e-07)
$\rho = 8.0$	2.62e-03	1000000	1248.90	990.86	11598.66	4.57e-05 (4.36e-07)

Table 4.8 shows that the resulting approximation is even better than the one obtained in dimension 3, illustrating that our algorithm did indeed exploit the low intrinsic dimension of the dataset.

5. Analysis of the algorithm.

5.1. General setting. We will start the analysis in a more general setting than that of section subsection 2.1. Let $\mathcal{B}$ be a separable Banach space with dual space $\mathcal{B}^*$, and write $[\,\cdot\,,\,\cdot\,]$ for the duality product between $\mathcal{B}^*$ and $\mathcal{B}$. Let $\mathcal{L}\colon \mathcal{B} \to \mathcal{B}^*$ be a linear bijection and let $G \coloneqq \mathcal{L}^{-1}$. Assume $\mathcal{L}$ to be symmetric and positive (i.e., $[\mathcal{L}u, v] = [\mathcal{L}v, u]$ and $[\mathcal{L}u, u] \geq 0$ for $u, v \in \mathcal{B}$). Let $\|\cdot\|$ be the quadratic (energy) norm defined by $\|u\|^2 \coloneqq [\mathcal{L}u, u]$ for $u \in \mathcal{B}$ and let $\|\cdot\|_*$ be its dual norm defined by

$$\|\phi\|_* \coloneqq \sup_{0 \neq u \in \mathcal{B}} \frac{[\phi, u]}{\|u\|} = [\phi, G\phi] \text{ for } \phi \in \mathcal{B}^*. \tag{5.1}$$

Let $\{\phi_i\}_{i \in I}$ be linearly independent elements of $\mathcal{B}^*$ (known as *measurement functions*) and let $\Theta \in \mathbb{R}^{I \times I}$ be the symmetric positive-definite matrix defined by

$$\Theta_{ij} \coloneqq [\phi_i, G\phi_j] \quad \text{for } i, j \in I. \tag{5.2}$$

We assume that we are given $q \in \mathbb{N}$ and a partition $I = \bigcup_{1 \leq k \leq q} J^{(k)}$ of I. We represent $I \times I$ matrices as $q \times q$ block matrices according to this partition. Given an $I \times I$ matrix M we write $M_{k,l}$ for the (k,l)th block of M and $M_{k_1:k_2,l_1:l_2}$ for the submatrix of M defined by blocks ranging from k_1 to k_2 and l_1 to l_2. Unless specified otherwise we write L for the lower-triangular Cholesky factor of Θ and define

$$\Theta^{(k)} \coloneqq \Theta_{1:k,1:k}, \qquad A^{(k)} \coloneqq \Theta^{(k),-1}, \qquad B^{(k)} \coloneqq A^{(k)}_{k,k} \qquad \text{for } 1 \leq k \leq q. \tag{5.3}$$

We interpret the $\{J^{(k)}\}_{1 \leq k \leq q}$ as labeling a hierarchy of scales with $J^{(1)}$ representing the coarsest and $J^{(q)}$ the finest. We write $I^{(k)}$ for $\bigcup_{1 \leq k' \leq k} J^{(k')}$.

Throughout this section we assume that the ordering of the set I of indices is compatible with the partition $I = \bigcup_{k=1^q} J^{(k)}$, i.e., $k < l$, $i \in J^{(k)}$ and $j \in J^{(l)}$

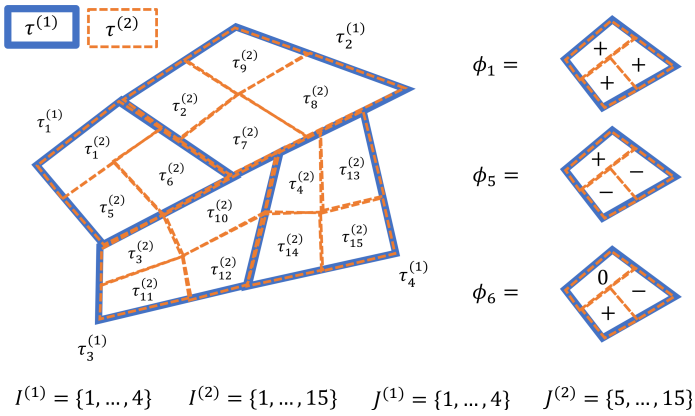

FIG. 5.1. *We illustrate the construction described in Example* 5.2 *in the case* $q = 2$. *On the left we see the nested partition of the domain, and on the right we see (the signs of) a possible choice for* ϕ_1, ϕ_5, *and* ϕ_6.

together imply $i \prec j$. We will write L or $\mathrm{chol}(\Theta)$ for the Cholesky factor of Θ in that ordering.

5.2. Main examples. We will prove the main results of this section in the setting where $\mathcal{L}$ is defined as in subsection 2.1 and the ϕ_i are chosen as in Examples 5.1 and 5.2. We will assume (without loss of generality after rescaling) that $\mathrm{diam}(\Omega) \leq 1$. As described in Figure 3.5, successive points of the maximin ordering can be gathered into levels, so that, after appropriate rescaling of the measurements, the Cholesky factorization in the maximin ordering falls in the setting of Example 5.1.

Example 5.1. Let $s > d/2$. For $h, \delta \in (0,1)$ let $\{x_i\}_{i \in I^{(1)}} \subset \{x_i\}_{i \in I^{(2)}} \subset \cdots \subset \{x_i\}_{i \in I^{(q)}}$ be a nested hierarchy of points in Ω that are homogeneously distributed at each scale in the sense of the following three inequalities:

(1) $\sup_{x \in \Omega} \min_{i \in I^{(k)}} |x - x_i| \leq h^k$,
(2) $\min_{i \in I^{(k)}} \inf_{x \in \partial\Omega} |x - x_i| \geq \delta h^k$, and
(3) $\min_{i,j \in I^{(k)}: i \neq j} |x_i - x_j| \geq \delta h^k$.

Let $J^{(1)} := I^{(1)}$ and $J^{(k)} := I^{(k)} \setminus I^{(k-1)}$ for $k \in \{2, \dots, q\}$. Let $\boldsymbol{\delta}$ denote the unit Dirac delta function and choose

$$\phi_i := h^{\frac{kd}{2}} \boldsymbol{\delta}(x - x_i) \text{ for } i \in J^{(k)} \text{ and } k \in \{1, \dots, q\}. \tag{5.4}$$

Given subsets $\tilde{I}, \tilde{J} \subset I$ we extend a matrix $M \in \mathbb{R}^{\tilde{I} \times \tilde{J}}$ to an element of $\mathbb{R}^{I \times J}$ by padding it with zeros.

Example 5.2 (see Figure 5.1). For $h, \delta \in (0,1)$, let $(\tau_i^{(k)})_{i \in I^{(k)}}$ be uniformly Lipschitz convex sets forming a regular nested partition of Ω in the following sense. For $k \in \{1, \dots, q\}$, $\Omega = \bigcup_{i \in I^{(k)}} \tau_i^{(k)}$ is a disjoint union except for the boundaries. $I^{(k)}$ is a nested set of indices, i.e., $I^{(k)} \subset I^{(k+1)}$ for $k \in \{1, \dots, q-1\}$. For $k \in \{2, \dots, q\}$ and $i \in I^{(k-1)}$, there exists a subset $c_i \subset I^{(k)}$ such that $i \in c_i$ and $\tau_i^{(k-1)} = \bigcup_{j \in c_i} \tau_j^{(k)}$. Assume that each $\tau_i^{(k)}$ contains a ball $B_{\delta h^k}(x_i^{(k)})$ of center $x_i^{(k)}$ and radius δh^k and is contained in the ball $B_{h^k}(x_i^{(k)})$. For $k \in \{2, \dots, q\}$ and $i \in I^{(k-1)}$, let the submatrices $\mathfrak{w}^{(k),i} \in \mathbb{R}^{(c_i \setminus \{i\}) \times c_i}$ satisfy $\sum_{j \in c_i} \mathfrak{w}_{m,j}^{(k),i} \mathfrak{w}_{n,j}^{(k),i} |\tau_j^{(k)}| = \delta_{mn}$ and $\sum_{j \in c_i} \mathfrak{w}_{l,j}^{(k),i} |\tau_j^{(k)}| = 0$ for each $l \in c_i \setminus \{i\}$, where $|\tau_i^{(k)}|$ denotes the volume of $\tau_i^{(k)}$. Let $J^{(1)} := I^{(1)}$ and

$J^{(k)} \coloneqq I^{(k)} \setminus I^{(k-1)}$ for $k \in \{2, \dots, q\}$. Let $W^{(1)}$ be the $J^{(1)} \times I^{(1)}$ matrix defined by $W^{(1)}_{ij} \coloneqq \delta_{ij}$. Letting $W^{(k)}$ be the $J^{(k)} \times I^{(k)}$ matrix defined by $W^{(k)} \coloneqq \sum_{i \in I^{(k-1)}} \mathfrak{w}^{(k),i}$ for $k > 2$, we set

$$\phi_i \coloneqq h^{-kd/2} \sum_{j \in I^{(k)}} W^{(k)}_{i,j} \mathbf{1}_{\tau_j^{(k)}} \quad \text{for each } i \in J^{(k)} \tag{5.5}$$

and define $[\phi_i, u] \coloneqq \int_\Omega \phi_i u \, \mathrm{d}x$. In order to keep track of the distance between the different ϕ_i of Example 5.2, we choose an arbitrary set of points $\{x_i\}_{i \in I} \subset \Omega$ with the property that $x_i \in \operatorname{supp}(\phi_i)$ for each $i \in I$.

5.3. Exponential decay of Cholesky factors. Our bound on the `ICHOL(0)` approximation error will be based on the following exponential decay estimate on the entries of the Cholesky factor L of Θ:

$$|L_{ij}| \leq \operatorname{poly}(N) \exp(-\gamma d(i,j)) \tag{5.6}$$

for a constant $\gamma > 0$ and a suitable distance measure $d(\cdot, \cdot) \colon I \times I \to \mathbb{R}$.

5.3.1. Algebraic identities and roadmap. The following block-Cholesky decomposition of Θ will be used to obtain the estimate (5.6).

LEMMA 5.3. *We have $\Theta = \bar{L} D \bar{L}^T$, with $\bar{L}$ and D defined by*

$$D \coloneqq \begin{pmatrix} B^{(1),-1} & 0 & \dots & 0 \\ 0 & B^{(2),-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & B^{(q),-1} \end{pmatrix}, \bar{L} \coloneqq \begin{pmatrix} \mathrm{Id} & \dots & \dots & 0 \\ B^{(2),-1} A^{(2)}_{2,1} & \ddots & 0 & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ B^{(q),-1} A^{(q)}_{q,1} & \dots & B^{(q),-1} A^{(q)}_{q,q-1} & \mathrm{Id} \end{pmatrix}^{-1}. \tag{5.7}$$

In particular, if $\tilde{L}$ is the lower-triangular Cholesky factor of D, then the lower-triangular Cholesky factor L of Θ is given by $L = \bar{L}\tilde{L}$.

Proof. To obtain Lemma 5.3 we successively apply Lemma 5.4 to Θ (see section SM2 for details). Lemma 5.4 summarizes classical identities satisfied by Schur complements. ∎

LEMMA 5.4 ([84, Chapter 1.1]). *Let $\Theta = \left(\begin{smallmatrix} \Theta_{1,1} & \Theta_{1,2} \\ \Theta_{2,1} & \Theta_{2,2} \end{smallmatrix}\right)$ be symmetric positive definite and $A = \left(\begin{smallmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{smallmatrix}\right)$ its inverse. Then*

$$\Theta = \begin{pmatrix} \mathrm{Id} & 0 \\ L_{2,1} & \mathrm{Id} \end{pmatrix} \begin{pmatrix} D_{1,1} & 0 \\ 0 & D_{2,2} \end{pmatrix} \begin{pmatrix} \mathrm{Id} & L_{2,1}^\top \\ 0 & \mathrm{Id} \end{pmatrix}, \tag{5.8}$$

$$A = \begin{pmatrix} \mathrm{Id} & -L_{2,1}^\top \\ 0 & \mathrm{Id} \end{pmatrix} \begin{pmatrix} D_{1,1}^{-1} & 0 \\ 0 & D_{2,2}^{-1} \end{pmatrix} \begin{pmatrix} \mathrm{Id} & 0 \\ -L_{2,1} & \mathrm{Id} \end{pmatrix}, \tag{5.9}$$

where

$$L_{2,1} = \Theta_{2,1}\Theta_{1,1}^{-1} = -A_{2,2}^{-1} A_{2,1}, \tag{5.10}$$

$$D_{1,1} = \Theta_{1,1} = \left(A_{1,1} - A_{1,2} A_{2,2}^{-1} A_{2,1}\right)^{-1}, \tag{5.11}$$

$$D_{2,2} = \Theta_{2,2} - \Theta_{2,1}\Theta_{1,1}^{-1}\Theta_{1,2} = A_{2,2}^{-1}. \tag{5.12}$$

Based on Lemma 5.3, (5.6) can be established by ensuring that

(1) the matrices $A^{(k)}$ (and hence also $B^{(k)}$) decay exponentially according to $d(\cdot,\cdot)$;
(2) the matrices $B^{(k)}$ have uniformly bounded condition numbers;
(3) the products of exponentially decaying matrices decay exponentially;
(4) the inverses of well-conditioned exponentially decaying matrices decay exponentially;
(5) the Cholesky factors of the inverses of well-conditioned exponentially decaying matrices decay exponentially; and
(6) if a $q \times q$ block lower-triangular matrix $\bar{L}$ with unit block-diagonal decays exponentially, then so does its inverse.

We will carry out this program in the setting of Examples 5.1 and 5.2 and prove that (5.6) holds with

$$d(i,j) \coloneqq h^{-\min(k,l)} \operatorname{dist}(x_i, x_j) \quad \text{for each } i \in J^{(k)},\ j \in J^{(l)}. \tag{5.13}$$

To prove (1), the matrices $\Theta^{(k)}$, $A^{(k)}$ (interpreted as coarse-grained versions of G and $\mathcal{L}$), and $B^{(1)}$ will be identified as stiffness matrices of the $\mathcal{L}$-adapted wavelets described in subsection 3.3. This identification is established on the general identities $\Theta^{(k)}_{i,j} = [\phi_i, G\phi_j]$ for $i,j \in I^{(k)}$, $A^{(k)} = (\Theta^{(k)})^{-1}$, $A^{(k)}_{i,j} = [\mathcal{L}\psi^{(k)}_i, \psi^{(k)}_j]$, and $B^{(k)}_{i,j} = [\mathcal{L}\chi^{(k)}_i, \chi^{(k)}_j]$, where the $\psi^{(k)}_i$ and $\chi^{(k)}_i$ are defined as in (3.6) and (3.7).

5.3.2. Exponential decay of $A^{(k)}$. Our proof of the exponential decay of L will be based on that of $A^{(k)}$ as expressed in the following condition.

CONDITION 5.5. *Let $\gamma, C_\gamma \in \mathbb{R}_+$ be constants such that for $1 \le k \le q$ and $i,j \in I^{(k)}$,*

$$\left|A^{(k)}_{ij}\right| \le C_\gamma \sqrt{A^{(k)}_{ii} A^{(k)}_{jj}} \exp(-\gamma d(i,j)). \tag{5.14}$$

The matrices $A^{(k)}$ are coarse-grained versions of the local operator $\mathcal{L}$ and as such inherit some of its locality in the form of exponential decay. Such exponential localization results were first obtained by [57] for the coarse-grained operators obtained from local orthogonal decomposition (LOD) applied to second-order elliptic PDEs with rough coefficients. [62] gives similar results for measurement functions chosen as in Example 5.2. [44] extends the results on exponential decay to higher-order operators satisfying a strong ellipticity condition. These results were obtained using similar *mass chasing* techniques that are difficult to extend to general higher-order operators. [50] present a simpler proof of the exponential decay of the LOD basis functions of [57] based on the exponential convergence of subspace iteration methods. [63] extends this technique (by presenting necessary and sufficient conditions expressed as frame inequalities in dual spaces) to elliptic PDEs of arbitrary (integer) order and new classes of (possibly nonconforming) measurements, including those of Examples 5.1 and 5.2. More recently, [17] showed localization results for the fractional partial differential operators by using the Caffarelli–Silvestre extension. The results of [63] are sufficient to show that Condition 5.5 holds true in the setting of Example 5.1 and Example 5.2.

THEOREM 5.6 ([63]). *In Example 5.1, the matrices $A^{(k)}$ satisfy*

$$\left|A^{(k)}_{ij}\right| \le C_\gamma \sqrt{A^{(k)}_{ii} A^{(k)}_{jj}} \exp\left(-\frac{\gamma}{h^k} \operatorname{dist}(x_i, x_j)\right) \le C_\gamma \sqrt{A^{(k)}_{ii} A^{(k)}_{jj}} \exp(-\gamma d(i,j)) \tag{5.15}$$

and in Example 5.2 *they satisfy*

$$\left|A_{ij}^{(k)}\right| \leq C_\gamma \exp\left(\frac{\gamma}{h}\right) \sqrt{A_{ii}^{(k)} A_{jj}^{(k)}} \exp(-\gamma d(i,j)) \tag{5.16}$$

with the constants C_γ *and* γ *depending only on* $\|\mathcal{L}\|$, $\|\mathcal{L}^{-1}\|$, s, d, Ω, *and* δ. *In particular, they satisfy Condition* 5.5 *with the constants described above.*

Proof. Our Example 5.1 is equivalent to Example 2.29 of [63]. In [63, Theorems 2.25 and 2.26] it is shown that in the gamblets $\{\psi_i^{(k)}\}_{i\in I^{(k)}}$ computed in this setting decay exponentially on the length-scale h^k, with respect to the energy norm. By [63, Theorem 3.8] we have $A_{ij}^{(k)} = [\psi_i^{(k)}, \mathcal{L}\psi_j^{(k)}]$ and, therefore, the exponential decay of gamblets implies the exponential decay of the $A^{(k)}$.

We further note that Example 5.2 is equivalent to Example 2.27 in [63]. Therefore, by the same theorems, as above, the results of [63] imply exponential decay of the $A^{(k)}$ in this setting.[2]

See also [64, Theorem 15.45] for a detailed proof and [64, Theorem 15.43] for required sufficient lower bounds on $A_{ii}^{(k)}$. ∎

5.3.3. Bounded condition numbers. In this section, we will bound the condition numbers of $B^{(k)}$ based on the following condition, which we will show to be satisfied for Examples 5.1 and 5.2.

CONDITION 5.7. *Let* $H \in (0,1), C_\Phi \geq 1$ *be constants such that for* $1 \leq k < l \leq q$,

$$\lambda_{\min}\big(\Theta^{(k)}\big) \geq \frac{1}{C_\Phi} H^{2k}, \tag{5.17}$$

$$\lambda_{\max}\big(\Theta_{l,l}^{(q)} - \Theta_{l,1:k}^{(q)} \Theta_{1:k,1:k}^{(q),-1} \Theta_{1:k,l}^{(q)}\big) \leq C_\Phi H^{2k}. \tag{5.18}$$

THEOREM 5.8. *Condition* 5.7 *implies that, for all* $1 \leq k \leq q$,

$$C_\Phi^{-1} H^{-2(k-1)} \mathrm{Id} \prec B^{(k)} \prec C_\Phi H^{-2k} \mathrm{Id}, \tag{5.19}$$

and, for $\kappa \coloneqq H^{-2} C_\Phi^2$,

$$\mathrm{cond}\big(B^{(k)}\big) \leq \kappa. \tag{5.20}$$

Proof. The lower bound in (5.19) follows from (5.18) and

$$B^{(k)} = \big(\Theta_{k,k}^{(q)} - \Theta_{k,1:(k-1)}^{(q)} \Theta_{1:k,1:k}^{(q),-1} \Theta_{1:(k-1),k}^{(q)}\big)^{-1}. \tag{5.21}$$

The upper bound in (5.19) follows from (5.17) and $B^{(k)} = \big(\big(\Theta^{(k)}\big)^{-1}\big)_{k,k}$. ∎

The following theorem shows that (5.18) is a Poincaré inequality closely related to the accuracy of numerical homogenization basis functions [57, 65, 44] and (5.17) is an inverse Sobolev inequality related to the regularity of the discretization of $\mathcal{L}$.

THEOREM 5.9. *Condition* 5.7 *holds true if the constants* $C_\Phi \geq 1$ *and* $H \in (0,1)$ *satisfy*

(1) $\frac{1}{C_\Phi} H^{2k} \leq \frac{\|\phi\|_*^2}{|\alpha|^2}$ *for* $\alpha \in \mathbb{R}^{I^{(k)}}$ *and* $\phi = \sum_{i\in I^{(k)}} \alpha_i \phi_i$ *and*

[2] We point out that the block $A_{m,l}^{(k)}$ in our notation is $W^{(m)} \pi^{(m,k)} A^{(k)} \pi^{(k,l)} W^{(l),\top}$ in the notation of [63].

(2) $\min_{\varphi\in\operatorname{span}(\phi_i)_{i\in I^{(k-1)}}} \frac{\|\phi-\varphi\|_*^2}{|\alpha|^2} \leq C_\Phi H^{2(k-1)}$ *for* $\alpha\in\mathbb{R}^{J^{(l)}}$, $k<l\leq q$, *and* $\phi=\sum_{i\in J^{(l)}}\alpha_i\phi_i$.

Proof. Inequality (5.17) is a direct consequence of the first assumption of the theorem, whereas (5.18) follows from the variational property [84, Theorem 5.1] of the Schur complement:

$$\alpha^\top\left(\Theta_{l,l}-\Theta^{(q)}_{l,1:k}\Theta^{(q),-1}_{1:k,1:k}\Theta^{(q)}_{1:k,l}\right)\alpha=\inf_{\beta\in\mathbb{R}^{I^{(k)}}}(\alpha-\beta)^\top\Theta^{(q)}(\alpha-\beta) \tag{5.22}$$

$$=\min_{\varphi\in\operatorname{span}\{\phi_i|i\in I^{(k)}\}}\|\phi-\varphi\|_*^2\leq C_\Phi H^{2k}|\alpha|^2. \tag{5.23}$$ ∎

We will now show that Examples 5.1 and 5.2 satisfy the conditions of Theorem 5.9. For simplicity, for $\tilde\Omega\subset\Omega$ and $\phi\in H^{-s}(\Omega)$ we still write ϕ for the unique element $\tilde\phi\in H^{-s}(\tilde\Omega)$ such that $[\tilde\phi,u]=[\phi,u]$ for $u\in H^s_0(\tilde\Omega)$. The following Fenchel conjugate identity [15, Example 3.27, p. 93] will be useful throughout this section:

$$\|\phi\|^2_{H^{-s}(\Omega)}=\sup_{v\in H^s_0(\Omega)}2[\phi,v]-\|v\|^2_{v\in H^s_0(\Omega)}. \tag{5.24}$$

The first condition can be verified similarly as is done in [63].

LEMMA 5.10. *Let* Θ *be given as in Examples* 5.1 *and* 5.2. *Then there exists a constant* C *depending only on* δ, s, *and* d, *such that*

$$\frac{1}{C_\Phi}h^{2sk}\leq\frac{\|\phi\|_*^2}{|\alpha|^2} \tag{5.25}$$

for $C_\Phi=\|\mathcal{L}\|C$, $\alpha\in\mathbb{R}^{I^{(k)}}$, *and* $\phi=\sum_i\alpha_i\phi_i$.

Proof. The proof can be found in section SM2. ∎

In order to verify the second condition in Theorem 5.9, we will construct a φ such that $\phi-\varphi$ integrates to zero against polynomials of order at most $s-1$ on domains of size h^k. Then an application of the Bramble–Hilbert lemma [20] will yield the desired factor h^{ks}. To avoid scaling issues we define, for $1\leq k\leq q$ and $i\in I^{(k)}$,

$$\phi_i^{(k)}\coloneqq\begin{cases}\boldsymbol{\delta}_{x_i} & \text{in Example 5.1,}\\ \mathbf{1}_{\tau_i^{(k)}}/|\tau_i^{(k)}| & \text{in Example 5.2,}\end{cases} \tag{5.26}$$

noting that $\operatorname{span}\{\phi_i^{(k)}\mid i\in I^{(k)}\}=\operatorname{span}\{\phi_i\mid i\in I^{(k)}\}$. To obtain estimates independent of the regularity of Ω, for the simplicity of the proof and without loss of generality, we will partially work in the extended space $\mathbb{R}^d$ (rather than on Ω). We write v for the zero extension of $v\in H^s_0(\Omega)$ to $H^s(\mathbb{R}^d)$ and $\phi_i^{(k)}$ for the extension of $\phi_i^{(k)}\in H^{-s}(\Omega)$ to an element of the dual space of $H^s_{\mathrm{loc}}(\mathbb{R}^d)$. We introduce new measurement functions in the complement of Ω as follows. For $1\leq k\leq q$ we consider countably infinite index sets $\tilde I^{(k)}\supset I^{(k)}$. We choose points $(x_i)_{i\in\tilde I^{(q)}\setminus I^{(q)}}$ satisfying

$$\sup_{x\in\mathbb{R}^d\setminus\Omega}\min_{i\in\tilde I^{(k)}}\operatorname{dist}(x_i,x)\leq\delta^{-1}h^k,\qquad\min_{i\neq j\in\tilde I^{(k)}\setminus I^{(k)}}\operatorname{dist}(x_i,x_j\cup\partial\Omega)\geq\delta h^k. \tag{5.27}$$

We then define, for $1\leq k\leq q$ and $i\in\tilde I^{(k)}$, $\phi_i^{(k)}\coloneqq\delta_{x_i}$ for Example 5.1, and $\phi_i^{(k)}\coloneqq\frac{\mathbf{1}_{B_{\delta h^k}(x_i)}}{|B_{\delta h^k}(x_i)|}$ for Example 5.2. Let $\mathcal{P}^{s-1}$ denote the linear space of polynomials of degree at most $s-1$ (on $\mathbb{R}^d$).

LEMMA 5.11. *Let Θ be as in Example* 5.1 *or Example* 5.2. *Given $\rho \in (2,\infty)$ and $1 \le k < l \le q$ let $w \in \mathbb{R}^{J^{(l)} \times \tilde{I}^{(k)}}$ be such that*

$$\int_{B_{\rho h^k}(x_i)} \left(\phi_i - \sum_{j \in \tilde{I}^{(k)}} w_{ij} \phi_j^{(k)} \right)(x) p(x) \, \mathrm{d}x = 0 \quad \forall p \in \mathcal{P}^{s-1} \text{ and } i \in J^{(l)} \tag{5.28}$$

and $w_{ij} \neq 0 \Rightarrow \operatorname{supp}(\phi_j^{(k)}) \subset B_{\rho h^k}(x_i)$. Then, for $\alpha \in \mathbb{R}^{J^{(l)}}$, $\phi \coloneqq \sum_{i \in J^{(l)}} \alpha_i \phi_i$ and $\varphi \coloneqq \sum_{i \in J^{(l)}, j \in I^{(k)}} \alpha_i w_{ij} \phi_j^{(k)}$ satisfy

$$\|\phi - \varphi\|_*^2 \le \|\mathcal{L}^{-1}\| C(d,s) \frac{\rho^{d+2s}}{\delta^d} \left(1 + h^{-ld} \omega_{l,k}^2\right) h^{2sk} |\alpha|^2 \tag{5.29}$$

with $\omega_{l,k} \coloneqq \sup_{i \in J^{(l)}} \sum_{j \in \tilde{I}^{(k)}} |w_{ij}|$ and $\|\phi\|_* \coloneqq \sup_{u \in H_0^s(\Omega)} [\phi, u] / [\mathcal{L}u, u]^{\frac{1}{2}}$ as in (5.1).

We proceed by proving Lemma 5.11 in the setting of Example 5.1. The proof in the setting of Example 5.2 can be found in section SM2. For $u \in H^s(\Omega)$ write $\mathrm{D}^0 u \coloneqq u$ and for $1 \le k \le s$ write $\mathrm{D}^k u$ for the vector of partial derivatives of u of order k, i.e., $\mathrm{D}^k u \coloneqq (\frac{\partial^k u}{\partial_{i_1} \cdots \partial_{i_k}})_{i_1,\ldots,i_k=1,\ldots,d}$. The proof of Lemma 5.11 will use the following version of the Bramble–Hilbert lemma.

LEMMA 5.12 ([20]). *Let $\Omega \subset \mathbb{R}^d$ be convex and let ϕ be a sublinear functional on $H^s(\Omega)$ for $s \in \mathbb{N}$ such that*

(1) *there exists a constant $\tilde{C}$ such that, for all $u \in H^s(\Omega)$,*

$$|\phi(u)| \le \tilde{C} \sum_{k=0}^{s} \operatorname{diam}(\Omega)^k \|\mathrm{D}^k u\|_{L^2(\Omega)}; \tag{5.30}$$

(2) *and $\phi(p) = 0$ for all $p \in \mathcal{P}^{s-1}$.*

Then, for all $u \in H^s(\Omega)$,

$$|\phi(u)| \le \tilde{C} C(d,s) \operatorname{diam}(\Omega)^s \|\mathrm{D}^s u\|_{L^2(\Omega)}. \tag{5.31}$$

The following lemma is obtained from Lemma 5.12.

LEMMA 5.13. *For $1 \le k < l \le q$ and $i \in J^{(l)}$, let ϕ_i, w_{ij} be as in Lemma* 5.11 *and Example* 5.2 *and define $\varphi_i \coloneqq \sum_{j \in I^{(k)}} w_{ij} \phi_j^{(k)}$. Then there exists a constant $C(d,s)$ such that, for all $v \in H_0^s(\Omega)$,*

$$\left| \int_{B_{\rho h^k}(x_i)} (\phi_i - \varphi_i)(x) v(x) \, \mathrm{d}x \right| \le C(d,s) \rho^{s-d/2} h^{(s-d/2)k} \left(h^{ld/2} + \sum_{j \in \tilde{I}^{(k)}} |w_{ij}| \right) \|\mathrm{D}^s v\|_{L^2\left(B_{\rho h^k}(x_i)\right)}. \tag{5.32}$$

Proof. We apply Lemma 5.12 to the linear functional $u \mapsto \int_{B_{\rho h^k}} (\phi_i - \varphi_i) u$. Since the second requirement of Lemma 5.12 is fulfilled by definition, it remains to bound $\tilde{C}$. We only execute the proof for Example 5.1; the proof for Example 5.2 is analogous. We first note that while the sum in the definition of φ_i only ranges over $j \in I^{(k)}$, we can increase it to run over all of $j \in \tilde{I}^{(k)}$, since for $j \in \tilde{I}^{(k)} \setminus I^{(k)}$, the support of $\phi_j^{(k)}$

is disjoint from that of $v \in H_0^s(\Omega)$. Let $u \in H^s(\Omega)$. Writing $C(d,s)$ for the continuity constant of the embedding of $H^s(B_1(0))$ into $C_b(B_1(0))$, the inequalities

$$\begin{aligned}&\max_{B_{\rho h^k}(x_i)} |u(\cdot)| \\ &\quad = \max_{x \in B_1(0)} \left|u\left(\rho h^k (x - x_i)\right)\right| \leq C(d,s) \sum_{m=0}^{s} (\rho h^k)^m \left\|[\mathrm{D}^m u]\left(\rho h^k (\cdot - x_i)\right)\right\|_{L^2(B_1(0))}\end{aligned}$$

and

$$\left\|[\mathrm{D}^m u]\left(\rho h^k (\cdot - x_i)\right)\right\|_{L^2(B_1(0))} = (\rho h^k)^{-d/2} \|\mathrm{D}^m u\|_{L^2(B_{\rho h^k}(x_i))}$$

imply that

(5.33)

$$\begin{aligned}&|\phi_i(u) - \varphi_i(u)| \\ &\quad \leq \left(h^{ld/2} + \sum_{j \in \tilde{I}^{(k)}} |w_{ij}|\right) \max_{x \in B_{\rho h^k}(x_i)} |u(x)| \\ &\quad \leq C(d,s) \rho^{-d/2} h^{-kd/2} \left(h^{ld/2} + \sum_{j \in \tilde{I}^{(k)}} |w_{ij}|\right) \sum_{m=0}^{s} (\rho h^k)^m \|\mathrm{D}^m u\|_{L^2(B_{\rho h^k}(x_i))}.\end{aligned}$$

Therefore the first condition of Lemma 5.12 holds with

$$\tilde{C} = C(d,s) \rho^{-d/2} h^{-kd/2} \left(h^{ld/2} + \sum_{j \in \tilde{I}^{(k)}} |w_{ij}|\right), \tag{5.34}$$

and we conclude the proof by writing $C(d,s)$ for any constant depending only on d and s. ∎

We can now conclude the proof of Lemma 5.11.

Proof of Lemma 5.11. Write $\varphi := \sum_{i \in J^{(l)}} \alpha_i \varphi_i$ and $\varphi_i := \sum_{j \in I^{(k)}} w_{ij} \phi_j^{(k)}$. Equation (5.24) implies that

(5.35)

$$\|\phi - \varphi\|_{H^{-s}(\Omega)}^2 = \sup_{v \in H_0^s(\Omega)} \left(\sum_{i \in J^{(l)}} 2\alpha_i \int_{B_{\rho h^k}(x_i)} (\phi_i - \varphi_i)(x) v(x) \, \mathrm{d}x\right) - \|v\|_{H_0^s(\Omega)}^2.$$

The packing inequality $\sum_{i \in J^{(l)}} \|\mathrm{D}^s v\|_{L^2\left(B_{\rho h^k}(x_i)\right)}^2 \leq C(d) \left(h^{k-l}\rho/\delta\right)^d \|v\|_{H_0^s(\Omega)}^2$ together with Lemma 5.13 yields

(5.36)

$$\begin{aligned}&\|\phi - \varphi\|_{H^{-s}(\Omega)}^2 \\ &\quad \leq \sup_{v \in H_0^s(\Omega)} \sum_{i \in J^{(l)}} \left[2|\alpha_i| C(d,s) \rho^{s - \frac{d}{2}} h^{(s - \frac{d}{2})k} \left(h^{\frac{ld}{2}} + \sum_{j \in I^{(k)}} |w_{ij}|\right) \|\mathrm{D}^s v\|_{L^2(B_{\rho h^k}(x_i))}\right. \\ &\qquad \left. - (C(d))^{-1} \left(h^{k-l}\rho/\delta\right)^{-d} \|\mathrm{D}^s v\|_{L^2\left(B_{\rho h^k}(x_i)\right)}^2\right].\end{aligned}$$

Applying the inequality $2ax - bx^2 \leq a^2/b$ to each summand yields

$$\|\phi - \varphi\|^2_{H^{-s}(\Omega)} \leq C(d) \left(h^{k-l}\rho/\delta\right)^d \sum_{i \in J^{(l)}} \left(\alpha_j C(d,s) \rho^{s-\frac{d}{2}} h^{(s-\frac{d}{2})k} \left(h^{\frac{ld}{2}} + \sum_{j \in J^{(k)}} |w_{ij}| \right) \right)^2$$
$$\leq C(d,s) \frac{\rho^{2s}}{\delta^d} \left(1 + h^{-ld}\omega^2_{l,k}\right) h^{2sk} |\alpha|^2 .$$

Since, for all $f \in H^{-s}(\Omega)$,

$$\|f\|^2_* = [f, \mathcal{L}^{-1} f] \leq \|f\|_{H^{-s}(\Omega)} \|\mathcal{L}^{-1} f\|_{H^s_0(\Omega)} \leq \|\mathcal{L}^{-1}\| \|f\|^2_{H^{-s}(\Omega)}, \tag{5.37}$$

we have $\|\phi - \varphi\|_* \leq \sqrt{\|\mathcal{L}^{-1}\|} \|\phi - \varphi\|_{H^{-s}(\Omega)}$, and this completes the proof. ∎

The following geometric lemma shows that the assumption (5.28) of Lemma 5.11 can be satisfied with a uniform bound on the value of ρ and the norm of weights $w_{i,j}$.

LEMMA 5.14. *There exist constants $\rho(d,s)$ and $C(d,s,\delta)$ such that for all $1 \leq k < l \leq q$ there exist weights $w \in \mathbb{R}^{J^{(l)} \times \tilde{I}^{(k)}}$ satisfying* (5.28) *and (with $\omega_{l,k}$ defined as in Lemma* 5.11*)*

$$\omega^2_{l,k} \leq h^{ld} C(d,s,\delta) . \tag{5.38}$$

Proof. For Example 5.1, (5.28) is equivalent to

$$h^{ld/2} p(x_i) = \sum_{j \in \tilde{I}^{(k)}_\rho} w_{ij} p(x_j) \ \forall p \in \mathcal{P}^{s-1}, \tag{5.39}$$

where $\tilde{I}^{(k)}_\rho \coloneqq \{j \in \tilde{I}^{(k)} \mid x_j \in B(x_i, \rho h^k)\}$.

Fix $i \in J^{(l)}$, let $\lambda > 0$, and write $x^\lambda_j \coloneqq \frac{x_j - x_i}{\lambda}$. Write $\mathbf{0} \coloneqq (0, \ldots, 0) \in \mathbb{R}^d$. Since the function $p(\,\cdot\,) \mapsto p(\frac{\cdot - x_i}{\lambda})$ is surjective on $\mathcal{P}^{s-1}$, (5.39) is satisfied if

$$h^{ld/2} p(\mathbf{0}) = \sum_{j \in \tilde{I}^{(k)}_\rho} w_{ij} p(x^\lambda_j) \ \forall p \in \mathcal{P}^{s-1}. \tag{5.40}$$

For a multiindex $n = (n_1, \ldots, n_d) \in \mathbb{N}^d$ and a point $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$, write $z^n \coloneqq \prod_{m=1}^d z_m^{n_m}$. Use the convention $\mathbf{0}^n = 0$ if $n \neq \mathbf{0}$ and $\mathbf{0}^{\mathbf{0}} = 1$. To satisfy (5.40) it is sufficient to identify a subset σ of $\tilde{I}^{(k)}_\rho$ and $w_{i,\,\cdot} \in \mathbb{R}^{\tilde{I}^{(k)}}$ such that $\#\sigma = s^d$, $w_{i,j} = 0$ for $j \notin \sigma$, and

$$h^{ld/2} \mathbf{0}^n = \sum_{j \in \sigma} w_{ij} (x^\lambda_j)^n \ \forall n \in \{0, \ldots, s-1\}^d . \tag{5.41}$$

Let $\mathbb{V}^\lambda \in \mathbb{R}^{\{0,1,\ldots,s-1\}^d \times \sigma}$ be the $s^d \times s^d$ matrix defined by

$$\mathbb{V}^\lambda_{n,j} \coloneqq \left(x^\lambda_j\right)^n \tag{5.42}$$

for a multiindex $n \in \mathbb{N}^d$ and a point $x \in \mathbb{R}^d$ $x^n \coloneqq \prod_{m=1}^d x^{n_m}$. Let $\mathbf{w} \in \mathbb{R}^\sigma$ be defined by $\mathbf{w}_j \coloneqq w_{i,j}$ for $j \in \sigma$. Equation (5.41) is then equivalent to

$$h^{ld/2} \mathbf{e} = \mathbb{V}^\lambda \mathbf{w}, \tag{5.43}$$

where $\mathbf{e} \in \mathbb{R}^{\{0,1,\dots,s-1\}^d}$ is defined by $\mathbf{e}_n \coloneqq \mathbf{0}^n$ for $n \in \{0,1,\dots,s-1\}^d$. We will now identify $\mathbf{w}$ by inverting (5.43). To achieve this while keeping the norm of $\mathbf{w}$ under control we will seek to identify the subset σ and $\lambda > 0$ such that $\sigma_{\min}(\mathbb{V}^\lambda)$ (the minimal singular value of $\mathbb{V}^\lambda$) is bounded from below by a constant depending only on s and d.

For $\alpha \geq 0$ let $(\epsilon_j)_{j\in\{0,1,\dots,s-1\}^d}$ be elements of $\mathbb{R}^d$ satisfying $|\epsilon_j| \leq \alpha$ for all $j \in \{0,1,\dots,s-1\}^d$. Let $\mathbf{1} \coloneqq (1,\dots,1) \in \mathbb{R}^d$ and, for $j \in \{0,1,\dots,s-1\}^d$, let $z_j \coloneqq \mathbf{1} + j + \epsilon_j$. Observe that for $\alpha = 0$ the points z_j are on a regular grid. Let $\bar{\mathbb{V}}^\alpha \in \mathbb{R}^{\{0,1,\dots,s-1\}^d\times\{0,1,\dots,s-1\}^d}$ be the $s^d \times s^d$ matrix defined by $\bar{\mathbb{V}}^\alpha_{n,j} \coloneqq (z_j)^n$. Let V be the $s \times s$ Vandermonde matrix defined by $V_{i,j} = i^j$. Writing $\sigma_{\min}(V)$ for the minimal singular value of V we have, for $\alpha = 0$, by [43, Theorem 4.2.12],

$$\sigma_{\min}\left(\bar{\mathbb{V}}^0\right) = \left(\sigma_{\min}(V)\right)^d. \tag{5.44}$$

Since univariate polynomial interpolation on s points with polynomials of degree $s-1$ is uniquely solvable, we have $\sigma_{\min}(V) > 0$ and $\sigma_{\min}(\bar{\mathbb{V}}^0) > C(d,s) > 0$. Therefore, the continuity of the minimal singular value with respect the entries of $\bar{\mathbb{V}}^\alpha$ implies that there exists $\alpha^*, \sigma^* > 0$ depending only on s, d such that $\alpha \leq \alpha^*$ implies $\sigma_{\min}(\bar{\mathbb{V}}^\alpha) > \sigma^*$. Since (by construction) the $(x_i)_{i\in\tilde{I}^{(k)}}$ form a covering of $\mathbb{R}^d$ of radius h^k, the $(x_i^\lambda)_{i\in\tilde{I}^{(k)}}$ form a covering of $\mathbb{R}^d$ of radius h^k/λ and for each $n \in \{0,1,\dots,s-1\}^d$ there exists an $x^\lambda_{j_n}$ that is at distance at most h^k/λ from n. Let $\sigma \coloneqq \{j_n \mid n \in \{0,1,\dots,s-1\}^d\} \subset \tilde{I}^{(k)}$ be the collection of corresponding labels. It follows from $|x^\lambda_{j_n}| \leq \sqrt{d}s + h^k/\lambda$ that $|x_{j_n} - x_i| \leq \lambda\sqrt{d}s + h^k$, and $\sigma \subset \tilde{I}^{(k)}_\rho$ for $\rho > 1 + \lambda\sqrt{d}s/h^k$. Selecting $\lambda = h^k/\alpha^*$ implies that $\sigma_{\min}(\mathbb{V}^\lambda) > \sigma^*$ and $\sigma \subset \tilde{I}^{(k)}_\rho$ for $\rho > 1 + \sqrt{d}s/\alpha^*$. Defining

$$w_{ij} \coloneqq \begin{cases} \left((\mathbb{V}^\lambda)^{-1} h^{ld/2}\mathbf{e}\right)_n & \text{if } j = j_n \in \sigma, \\ 0 & \text{otherwise,} \end{cases} \tag{5.45}$$

the weights w_{ij} satisfy $\omega_{kl} \leq C(s,d)h^{ld/2}$ and (5.28) with a ρ depending only on s and d. This concludes the proof for Example 5.1. The proof is similar for Example 5.2 with minor changes (the bound on ω also depends on δ). ∎

The following lemma concerns the satisfaction of the second condition of Theorem 5.9.

LEMMA 5.15. *In the setting of Examples* 5.1 *and* 5.2*, there exists some constant* $C(d,s,\delta) > 0$ *such that, for* $2 \leq k < l \leq q$, $\alpha \in \mathbb{R}^{J^{(l)}}$ *and* $\phi = \sum_i \alpha_i \phi_i$,

$$\min_{\varphi\in\operatorname{span}(\phi_i)_{i\in I^{(k-1)}}} \frac{\|\phi - \varphi\|_*^2}{|\alpha|^2} \leq C(d,s,\delta)\|\mathcal{L}^{-1}\| h^{2s(k-1)}. \tag{5.46}$$

Proof. Apply Lemma 5.11 with the bounds on ρ and ω obtained in Lemma 5.14. ∎

The following theorem is a direct consequence of Theorem 5.9 and Lemmas 5.10 and 5.15.

THEOREM 5.16. *In the setting of Examples* 5.1 *and* 5.2 *there exists a constant* $C(d,s,\delta)$ *such that Condition* 5.7 *is fulfilled with* $C_\Phi \coloneqq \max(\|\mathcal{L}\|, \|\mathcal{L}^{-1}\|)C(d,s,\delta)$ *and* $H \coloneqq h^s$.

5.3.4. Propagation of exponential decay. We will now derive the exponential decay of the Cholesky factors L by combining the algebraic identities of Lemma 5.3

with the bounds on the condition numbers of the $B^{(k)}$ (implied by Condition 5.7) and the exponential decay of the $A^{(k)}$ (specified in Condition 5.5). The core of our proof is based on a combination/extension of the results of [21, 47, 10, 9, 51, 11] on decay algebras. The pseudodistance $d(\cdot,\cdot)$ appearing in (5.6) is not a pseudometric because it does not satisfy the triangle inequality. However, to prove (5.6) we we will only need the following weaker version of the triangle inequality.

DEFINITION 5.17. *A function* $d\colon I \times I \longrightarrow \mathbb{R}_+$ *is called a* hierarchical pseudometric *if*

(1) $d(i,i) = 0$ *for all* $i \in I$;
(2) $d(i,j) = d(j,i)$ *for all* $i,j \in I$;
(3) *for all* $1 \le k \le q$, $d(\cdot,\cdot)$ *restricted to* $J^{(k)} \times J^{(k)}$ *is a pseudometric;*
(4) *for all* $1 \le k \le l \le m \le q$ *and* $i \in J^{(k)}, s \in J^{(l)}, j \in J^{(m)}$, *we have* $d(i,j) \le d(i,s) + d(s,j)$.

Note that the $d(\cdot,\cdot)$ specified in (5.13) for Examples 5.1 and 5.2 is a hierarchical pseudometric. For a hierarchical pseudometric $d(\cdot,\cdot)$ and $\gamma \in \mathbb{R}_+$, let

$$c_d(\gamma) \coloneqq \sup_{1\le k\le l\le q} \sup_{j\in J^{(l)}} \sum_{i\in J^{(k)}} \exp(-\gamma d(i,j)). \tag{5.47}$$

The following theorem states the main result of this section.

THEOREM 5.18 (exponential decay of the Cholesky factors). *Assume that* Θ *fulfils Conditions* 5.5 *and* 5.7 *with the constants* $\gamma, C_\gamma, H, C_\Phi$ *and the hierarchical pseudometric* $d(\cdot,\cdot)$. *Then*

$$\left|\left(\operatorname{chol}(\Theta)\right)_{ij}\right| \le \frac{2C_\Phi c_d\left(\tilde\gamma/8\right)^2}{(1-r)^2}\left(4c_d\left(\tilde\gamma/4\right)\frac{C_\Phi C_\gamma\left(c_d\left(\tilde\gamma/2\right)\right)^2}{(1-r)^2}\right)^q \exp\left(-\frac{\tilde\gamma}{8}d(i,j)\right), \tag{5.48}$$

where

$$C_R \coloneqq \max\left\{1, \frac{2C_\gamma C_\Phi}{1+\kappa}\right\}, r \coloneqq \frac{1-\kappa^{-1}}{1+\kappa^{-1}}, \tilde\gamma \coloneqq \frac{-\log(r)}{1+\log(c_d(\gamma/2))+\log(C_R)-\log(r)}\frac{\gamma}{2},$$

and $\kappa = H^{-2}C_\Phi^2$ *is defined as in Theorem* 5.8.

The remaining part of this section will present the proof of Theorem 5.18. We will use the following lemma on the stability of exponential decay under matrix multiplication, the proof of which is a minor modification of that of [47].

LEMMA 5.19. *Let* I *be an index set that is partitioned as* $I = J^{(1)} \cup \cdots J^{(q)}$ *and let* $d\colon I \times I \to \mathbb{R}_{\ge 0}$ *satisfy*

$$d(i_1, i_{n+1}) \le \sum_{k=1}^{n} d(i_k, i_{k+1}) \quad \forall 1 \le n \le q-1 \text{ and } i_k \in J^{(k)}.$$

Let $M^{(k)} \in \mathbb{R}^{J^{(k)}\times J^{(k+1)}}$ *be such that* $|M^{(k)}_{i,j}| \le C\exp(-\gamma d(i,j))$ *for* $1 \le k \le q-1$ *and let*

$$c_d(\gamma/2) \coloneqq \sup_{1\le k\le q-1} \sup_{j\in J^{(k+1)}} \sum_{i\in J^{(k)}} \exp\left(-\frac{\gamma}{2}d(i,j)\right) \text{ for } \gamma \in \mathbb{R}_+. \tag{5.49}$$

Then, for $1 \leq n \leq q-1$,

$$\left| \left(\prod_{k=1}^{n} M^{(k)} \right)_{i,j} \right| \leq \left(c_d \left(\gamma/2 \right) C \right)^n \exp\left(-\frac{\gamma}{2} d(i,j) \right).$$

Proof. Set $i_1 \coloneqq i$, $i_{n+1} \coloneqq j$. Then

$$\begin{aligned} \left| \left(\prod_{k=1}^{n} M^{(k)} \right)_{i,j} \right| &\leq C^n \sum_{i_2,\ldots,i_n \in J^{(2)},\ldots,J^{(n)}} \exp\left(-\gamma \sum_{k=1}^{n} d\left(i_k, i_{k+1}\right) \right) \\ &\leq C^n \exp\left(-\frac{\gamma}{2} d\left(i_1, i_{n+1}\right) \right) \sum_{i_2,\ldots i_n \in I} \exp\left(-\frac{\gamma}{2} \sum_{k=1}^{n} d\left(i_k, i_{k+1}\right) \right) \\ &\leq \left(c_d \left(\gamma/2 \right) C \right)^n \exp\left(-\frac{\gamma}{2} d(i,j) \right). \end{aligned}$$

∎

The proof of the following lemma (on the stability of exponential decay under matrix inversion for well conditioned matrices) is nearly identical to that of [47]. (We only keep track of constants; see also [21] for a related result on the inverse of sparse matrices.)

LEMMA 5.20. *Let* $A \in \mathbb{R}^{I \times I}$ *be symmetric and positive definite with* $|A_{i,j}| \leq C \exp(-\gamma d(i,j))$ *for some* $C, \gamma > 0$ *and a metric* $d(\,\cdot\,,\,\cdot\,)$ *on* I. *It holds true that*

(5.50)

$$\begin{aligned} &\left| (A^{-1})_{i,j} \right| \\ &\leq \frac{4}{\left(\|A\| + \|A^{-1}\|^{-1} \right) (1-r)^2} \exp\left(-\frac{\log(\frac{1}{r})}{\left(1 + \log\left(c_d\left(\gamma/2\right)\right) + \log(C_R)\right) + \log(\frac{1}{r})} \frac{\gamma}{2} d(i,j) \right), \end{aligned}$$

where

$$\begin{aligned} c_d(\gamma/2) &\coloneqq \sup_{j \in I} \sum_{i \in I} \exp\left(-\frac{\gamma}{2} d(i,j) \right), \\ C_R &\coloneqq \max\left\{ 1, \frac{2C}{\|A\| + \|A^{-1}\|^{-1}} \right\} = \max\left\{ 1, \frac{2C\|A^{-1}\|}{1+\kappa} \right\}, \\ r &\coloneqq \frac{1 - \frac{1}{\|A\|\|A^{-1}\|}}{1 + \frac{1}{\|A\|\|A^{-1}\|}} = \frac{1-\kappa^{-1}}{1+\kappa^{-1}}, \end{aligned}$$

and $\kappa \coloneqq \|A\|\|A^{-1}\|$ *is the condition number of* A.

Proof. On a compact set not containing 0, the function $x \mapsto x^{-1}$ can be accurately approximated by low-order polynomials in x. Then, the spread of the exponential decay can be controlled by Lemma 5.19. See section SM2 for details. ∎

By representing Schur complements as matrix inverses, Lemma 5.20 can also be used to show that the Cholesky factors of well-conditioned exponentially decaying matrices are exponentially decaying. The following lemma appears in a similar form in [11] for banded matrices and in [51] without explicit constants.

LEMMA 5.21. *Let* $B \in \mathbb{R}^{I \times I} \simeq \mathbb{R}^{N \times N}$ *be symmetric and positive definite with condition number* κ *and such that* $|B_{i,j}| \leq C \exp(-\gamma d(i,j))$ *for some constant* $C > 0$

and some metric d on I. Let L be the Cholesky factor (in an arbitrary order) of B^{-1} ($B^{-1} = LL^T$). Then

(5.51)
$$|L_{i,j}| \leq \frac{4\sqrt{\|B\|}}{(\|B\| + \|B^{-1}\|^{-1})\,(1-r)^2} \exp\left(\frac{\log(r)}{1 + \log\left(c_d\left(\gamma/2\right)\right) + \log(C_R) - \log(r)} \frac{\gamma}{2} d(i,j)\right),$$

where

$$c_d(\gamma/2) \coloneqq \sup_{j\in I} \sum_{i\in I} \exp(-\frac{\gamma}{2} d(i,j)), C_R \coloneqq \max\{1, \frac{2C\|B^{-1}\|}{1+\kappa}\},$$

and $r \coloneqq \frac{1-\kappa^{-1}}{1+\kappa^{-1}}$.

Proof. Lemma 5.4 implies that the Schur complements of B^{-1} can be expressed as inverses of submatrices of B. The result then follows from Lemma 5.20 (see Proof 9 for details). ∎

The last ingredient needed to prove the exponential decay of the Cholesky factors of Θ is the following lemma showing the stability of exponential decay under inversion for block-lower-triangular matrices. (This operation appears in the definition of $\bar{L}$ in (5.7)).

LEMMA 5.22. *Let I be an index set that is partitioned as $I = J^{(1)} \cup \cdots J^{(q)}$ and assume that the matrix $L \in \mathbb{R}^{I\times I}$ is block-lower triangular with respect to this partition, with identity matrices as diagonal blocks. If $d(\cdot,\cdot)$ is a hierarchical pseudometric such that $|L_{ij}| \leq C\exp\left(-\gamma d(i,j)\right)$ (for some $C \geq 1$ and $\gamma > 0$), then it holds true that*

$$\left|(L^{-1})_{ij}\right| \leq 2^q \left(c_d\left(\gamma/2\right) C\right)^q \exp\left(-\frac{\gamma}{2} d(i,j)\right) \tag{5.52}$$

with $c_d(\gamma) \coloneqq \sup_{1\leq k\leq l\leq q} \sup_{j\in J^{(l)}} \sum_{i\in J^{(k)}} \exp\left(-\gamma d(i,j)\right)$.

Proof. The Neumann series of a $q \times q$ block-lower-triangular matrix with identity matrices on the (block) diagonal can be written as

$$L^{-1} = \sum_{k=0}^{q} \left(\mathrm{Id} - L\right)^k . \tag{5.53}$$

Since the sum terminates in q steps, the thickening of the exponential decay can be bounded using Lemma 5.19. See section SM2 in the supplementary material for details. ∎

By applying the above results to the decomposition obtained in Lemma 5.3, we conclude the proof of Theorem 5.18. See section SM2 in the supplementary material for details.

5.4. Complexity and error estimates. The results of the previous sections allow us to prove the following theorem on the exponential decay of the Cholesky factors and the accuracy of their truncation.

THEOREM 5.23. *In the setting of Examples* 5.1 *and* 5.2 *there exist constants $C, \gamma, \alpha > 0$ depending only on d, Ω, s, $\|\mathcal{L}\|$, $\|\mathcal{L}^{-1}\|$, h, and δ, such that the entries of the Cholesky factor L of Θ satisfy*

$$|L_{ij}| \leq CN^{\alpha} \exp(-\gamma d(i,j)), \tag{5.54}$$

where $d\colon I \times I \to \mathbb{R}$ *is the hierarchical pseudometric defined by*

$$d(i,j) \coloneqq h^{-\min(k,l)} \operatorname{dist}\left(\operatorname{supp}\left(\phi_i\right), \operatorname{supp}\left(\phi_j\right)\right) \quad \forall i \in J^{(k)}, j \in J^{(l)}. \tag{5.55}$$

As a consequence, writing

$$L^S_{ij} \coloneqq \begin{cases} L_{ij} & \text{for } (i,j) \in S, \\ 0 & \text{else} \end{cases} \tag{5.56}$$

with $S \supset S_{d,\rho} \coloneqq \{(i,j) \mid d(i,j) \leq \rho\}$*, we have* $\left\|\Theta - L^S L^{S,\top}\right\|_{\mathrm{Fro}} \leq \epsilon$ *for* $\rho \geq \tilde{C}(C,\gamma)\log(N/\epsilon)$*. Furthermore, writing* $E \coloneqq \Theta - L^S L^{S,\top}$*, using the* ϵ*-perturbation* $\Theta - E$ *of* Θ *as the input to Algorithm* 2.2 *returns* L^S *as the output.*

Proof. Theorems 5.6 and 5.16 imply that Conditions 5.5 and 5.7 are fulfilled with constants depending only on d, s, $\|\mathcal{L}\|$, $\|\mathcal{L}^{-1}\|$, h, and δ. Theorem 5.18 concludes the exponential decay of L. The accuracy of the truncated factors follows directly from the exponential decay. ∎

Theorem 3.1 is a direct consequence of Theorem 5.23.

Proof of Theorem 3.1. As described in subsection 3.3, the maximin ordering can be represented as a hierarchical ordering satisfying the conditions of Example 5.1. The result follows from Theorem 5.23 by observing that the sparsity pattern S_ρ specified in section 2 satisfies

$$S_{d,(\delta h)^{-1}\rho} \supset S_\rho \supset S_{d,\delta h \rho}\,. \tag{5.57}$$

Scaling the weights of the measurement functions ϕ_i to 1 increases the error by a factor that is at most polynomial in N, which can be subsumed into the $\log(N)$-dependence of ρ by increasing the constants in the decay estimates. ∎

While accurate (per Theorem 5.23), it is computationally inefficient to compute the full Cholesky factor first (with Algorithm 2.1) and then truncate it according to S_ρ. Instead, we want to directly compute an approximation of L from the incomplete factorization Algorithm 2.2, whose complexity is bounded by the following theorem.

THEOREM 5.24. *In the setting of Examples* 5.1 *and* 5.2*, there exists a constant* $C(d,\delta)$*, such that, for* $S \subset \{(i,j) \mid d(i,j) \leq \rho\}$*, the application of Algorithm* 2.2 *has computational complexity* $C(d,\delta)Nq\rho^d$ *in space and* $C(d,\delta)Nq^2\rho^{2d}$ *in time. In particular,* $q \propto \log N / \ln \frac{1}{h^d}$ *implies the upper bounds of* $C(d,\delta,h)\rho^d N \log N$ *on the space complexity, and of* $C(d,\delta,h)\rho^{2d} N \log^2 N$ *on the time complexity.*

Proof. Defining $m \coloneqq \max_{j \in I, 1 \leq k \leq q} \#\{i \in J^{(k)} \mid i \prec j \text{ and } d(i,j) \leq \rho\}$, $|x_i - x_j| \geq \delta^{-1} h^l$ for $i,j \in I^{(l)}$ implies that $m \leq C(d,\delta)\rho^d$. Therefore $\#\{i \in I \mid i \prec j \text{ and } d(i,j) \leq \rho\} \leq qmN$ implies the bound on space complexity.

Consider the structure of the nested for-loops of Algorithm 2.2 and observe that, for every k in the innermost loop, the number of distinct (i,j) satisfying $i \prec j \prec k$, $(j,k) \in S$, and $(i,j) \in S$ is at most $(qm)^2$. This implies the upper bound $N(qm)^2$ on the time complexity. ∎

Theorems 5.23 and 5.24 imply that the application of Algorithm 2.2 to $\Theta - E$ (the ϵ-perturbation of Θ described in Theorem 5.23) returns an ϵ-accurate Cholesky factorization of Θ in computational complexity $\mathcal{O}(N \log^2(N) \log^{2d}(N/\epsilon))$. In practice we do not have access to E, so we need to rely on the stability of Algorithm 2.2 to deduce that Θ and $\Theta - E$ (used as inputs) would yield similar outputs, for sufficiently

small E. Even though such a stability property of `ICHOL(0)` would also be required by prior works on incomplete LU-factorization such as [31], we did not find this type of result in the literature. We also found it surprisingly difficult to prove (and were unable to do so) for the maximin ordering and sparsity pattern, although we always observed the stability of Algorithm 2.2 in practice, for reasonable values of ρ. We can, however, prove the stability of Algorithm 2.2 when using a slight modification of the ordering and sparsity pattern that compromises neither the computational complexity nor the accuracy of the factorization. The modified ordering and sparsity pattern, being inspired by the concepts of red-black orderings [46] and supernodal factorizations [71, 56], also allows one to take advantage of parallelism and dense linear algebra operations and could therefore be used to improve the practical performance of the algorithm. For $r > 0$, $1 \leq k \leq q$, and $i \in J^{(k)}$, write

$$B_r^{(k)}(i) \coloneqq \{j \in J^{(k)} \mid d(i,j) \leq r\}. \tag{5.58}$$

CONSTRUCTION 5.25 (supernodal multicolor ordering and sparsity pattern). *Let $\Theta \in \mathbb{R}^{I \times I}$ with $I \coloneqq \bigcup_{1 \leq k \leq q} J^{(k)}$ and let $d(\cdot,\cdot)$ be a hierarchical pseudometric. For $\rho \geq 1$, define the* supernodal multicolor ordering $\prec_\rho$ *and* sparsity pattern S_ρ *as follows. For each $k \in \{1,\dots,q\}$, select a subset $\tilde{J}^{(k)} \subset J^{(k)}$ of indices such that*

$$\forall \tilde{i}, \tilde{j} \in \tilde{J}^{(k)}, \qquad \tilde{i} \neq \tilde{j} \implies B_{\rho/2}^{(k)}(\tilde{i}) \cap B_{\rho/2}^{(k)}(\tilde{j}) = \emptyset, \tag{5.59}$$

$$\forall i \in J^{(k)}, \qquad \exists \tilde{i} \in \tilde{J}^{(k)} : i \in B_\rho^{(k)}(\tilde{i}). \tag{5.60}$$

Assign every index in $J^{(k)}$ to the element of $\tilde{J}^{(k)}$ closest to it, using an arbitrary method to break ties. That is, writing $j \rightsquigarrow \tilde{j}$ for the assignment of j to $\tilde{j}$,

$$\tilde{j} \in \operatorname*{arg\,min}_{\tilde{j}' \in \tilde{J}^{(k)}} d(j, \tilde{j}'), \tag{5.61}$$

for all $j \in J^{(k)}$ and $\tilde{j} \in \tilde{J}^{(k)}$ such that $j \rightsquigarrow \tilde{j}$. Define $\tilde{I} \coloneqq \bigcup_{1 \leq k \leq q} \tilde{J}^{(k)}$ and define the auxiliary sparsity pattern $\tilde{S}_\rho \subset \tilde{I} \times \tilde{I}$ by

$$\tilde{S}_\rho \coloneqq \left\{ (\tilde{i}, \tilde{j}) \in \tilde{I} \times \tilde{J} \,\middle|\, \exists i \rightsquigarrow \tilde{i}, j \rightsquigarrow \tilde{j} : d(i,j) \leq \rho \right\}. \tag{5.62}$$

Define the sparsity pattern $S_\rho \subset I \times I$ as

$$S_\rho \coloneqq \left\{ (i,j) \in I \times I \,\middle|\, \exists \tilde{i}, \tilde{j} \in \tilde{I} : i \rightsquigarrow \tilde{i}, j \rightsquigarrow \tilde{j}, (\tilde{i}, \tilde{j}) \in \tilde{S}_\rho \right\} \tag{5.63}$$

and call the elements of $\tilde{J}^{(k)}$ supernodes. *Color each $\tilde{j} \in \tilde{J}^{(k)}$ in one of $p^{(k)}$ colors such that no $\tilde{i}, \tilde{j} \in \tilde{J}^{(k)}$ with $(\tilde{i}, \tilde{j}) \in \tilde{S}_\rho$ have the same color. For $i \in J^{(k)}$ write $\mathrm{node}(i)$ for the $\tilde{i} \in \tilde{J}^{(k)}$ such that $i \rightsquigarrow \tilde{i}$ and write $\mathrm{color}(\tilde{i})$ for the color of $\tilde{i}$. Define the supernodal multicolor ordering $\prec_\rho$ by reordering the elements of I such that*

1. *$i \prec_\rho j$ for $i \in J^{(k)}$, $j \in J^{(l)}$, and $k < l$;*
2. *within each level $J^{(k)}$, we order the elements of supernodes colored in the same color consecutively, i.e., given $i, j \in J^{(k)}$ such that $\mathrm{color}(\mathrm{node}(i)) \neq \mathrm{color}(\mathrm{node}(j))$, $i \prec_\rho j \implies i' \prec_\rho j'$ for $\mathrm{color}(\mathrm{node}(i')) = \mathrm{color}(\mathrm{node}(i))$, and $\mathrm{color}(\mathrm{node}(j')) = \mathrm{color}(\mathrm{node}(j))$; and*
3. *the elements of each supernode appear consecutively, i.e., given $i, j \in J^{(k)}$ such that $\mathrm{node}(i) \neq \mathrm{node}(j)$, $i \prec_\rho j \implies i' \prec_\rho j'$ for $\mathrm{node}(i') = \mathrm{node}(i)$, and $\mathrm{node}(j') = \mathrm{node}(j)$.*

Starting from a hierarchical ordering and sparsity pattern, the modified ordering and sparsity pattern can be obtained efficiently.

LEMMA 5.26. *In the setting of Examples* 5.1 *and* 5.2, *given* $\{(i,j) \mid d(i,j) \leq \rho\}$, *there exist constants* C *and* $p_{\max}$ *depending only on the dimension* d *and the cost of computing* $d(\cdot,\cdot)$ *such that the ordering and sparsity pattern presented in Construction* 5.25 *can be constructed with* $p^{(k)} \leq p_{\max}$, *for each* $1 \leq k \leq q$, *in computational complexity* $Cq\rho^d N$.

Proof. The aggregation into supernodes can be done via a greedy algorithm by keeping track of all nodes that are not already within distance $\rho/2$ of a supernode and removing them one-at-a-time. We can then go through ρ-neighbourhoods and remove points within distance $\rho/2$ from our list of candidates for future supernodes. To create the coloring, we use the greedy graph coloring of [45] on the undirected graph G with vertices $\tilde{J}^{(k)}$ and edges $\left\{(\tilde{i},\tilde{j}) \in \tilde{S}_\rho \,\middle|\, \tilde{i},\tilde{j} \in \tilde{J}^{(k)}\right\}$. Defining $\deg(G)$ as the maximum number of edges connected to any vertex of G, the computational complexity of greedy graph coloring is bounded above by $\deg(G)\#\left(J^{(k)}\right)$ and the number of colors used by $\deg(G)+1$. A sphere-packing argument shows that $\deg(G)$ is at most a constant depending only on the dimension d, which yields the result. ∎

THEOREM 5.27. *In the setting of Examples* 5.1 *and* 5.2, *there exists a constant* C *depending only on* $d, s, \|\mathcal{L}\|, \|\mathcal{L}^{-1}\|$, h, *and* δ *such that, given the ordering* $\prec_\rho$ *and sparsity pattern* S_ρ *defined as in Construction* 5.25 *with* $\rho \geq C\log(N/\epsilon)$, *the incomplete Cholesky factor* L *obtained from Algorithm* 2.2 *has accuracy*

$$\|LL^T - \Theta\|_{\mathrm{Fro}} \leq \epsilon. \tag{5.64}$$

Furthermore, Algorithm 2.2 *has complexity of at most* $CN\rho^{2d}\log^2 N$ *in time and at most* $CN\rho^d \log N$ *in space.*

Proof. The triangle inequality implies that $S_\rho \subset \{(i,j) \mid d(i,j) \leq 2\rho\}$ and hence the bound on the complexity of Algorithm 2.2 follows from Theorem 5.24. The approximation property of the incomplete factors follows from the last part of Theorem 5.23 and a stability result for the incomplete Cholesky factorization with the supernodal multicolor ordering and sparsity pattern detailed in section SM3. ∎

This allows us to prove the main theorem presented in the introduction.

Proof of Theorem 2.1. Theorem 2.1 follows from Theorem 5.27 since rescaling the weights of the measurements to 1 increases bounds on errors by at most a multiplicative polynomial factor in N. By increasing the constant, this factor can be subsumed in the N-dependence of ρ. ∎

We have now established the results on exponential decay of the Cholesky factors of Θ and the accuracy of Algorithm 2.2. Before proceeding to the next section, we will quickly establish a result on low-rank approximation of the Cholesky factors.

THEOREM 5.28 (approximate PCA). *In the setting of Theorem* 3.1, *take* $\rho = \infty$ *and let* $L^{(k)}$ *be the matrix formed by the leading* k *columns of the Cholesky factors of* Θ *in the maximin ordering. Let* $l[i_k]$ *be as in* (2.7). *Then there exists a constant* C *depending only on* $\|\mathcal{L}\|$, $\|\mathcal{L}^{-1}\|$, d, *and* s *such that*

$$\left\|\Theta - L^{(k)}L^{(k),\top}\right\| \leq C l_{i_{k+1}}^{2s-d}. \tag{5.65}$$

Proof. Write $I = I_1 \cup I_2$ with $I_1 \coloneqq \{i_1,\ldots,i_k\}$ and $I_2 \coloneqq I \backslash I_1$. By Lemma 5.4, the approximation error made by keeping only the first k columns of the Cholesky factor-

ization is equal to the Schur complement $\Theta_{2,2}-\Theta_{2,1}\Theta_{1,1}^{-1}\Theta_{1,2}$. Consider the implicit hierarchy of the maximin ordering as in Figure 3.5 with $h = 1/2$ and let $p \in \{1, \dots, q\}$ be such that $2^{-p} \leq l[k]/l[1] \leq 2^{-p+1}$. Write $I = I_a \cup I_b$ with $I_a \coloneqq I^{(p)}$ and $I_b \coloneqq I \setminus I^{(p)}$. The variational property (5.22) implies that $\Theta_{2,2}-\Theta_{2,1}\Theta_{1,1}^{-1}\Theta_{1,2} \leq \Theta_{b,b}-\Theta_{b,a}\Theta_{a,a}^{-1}\Theta_{a,b}$. Theorem 5.16 (with $h = 1/2$ obtained from the implicit hierarchy of Figure 3.5) implies that $\Theta_{b,b} - \Theta_{b,a}\Theta_{a,a}^{-1}\Theta_{a,b} \leq C(\frac{1}{2})^{2s(p-1)-d}$. (The extra multiplicative $(\frac{1}{2})^{-d}$ term arises because the measurement functions are scaled by $h^{kd/2}$ in Example 5.1 with $h = \frac{1}{2}$.) We conclude the proof using $2^{-p-1} \leq l[k+1]/l[1] \leq 2^{-p+1}$. ∎

6. Extensions and byproducts.

6.1. The cases $s \leq d/2$ or $s \notin \mathbb{N}$. Theorem 3.1 requires that $s > d/2$ to ensure that the elements of $H^s(\Omega)$ are continuous (by the Sobolev embedding theorem) and that pointwise evaluations of the Green's function are well defined. The accuracy estimate of Theorem 3.1 can be extended to $s \leq d/2$ by replacing pointwise evaluations of the Green's function by local averages and using variants of the Haar prewavelets of Example 5.2 instead of variants of the subsampled Diracs of Example 5.1 to decompose Θ as in (3.12). Numerical experiments also suggest that the exponential decay of Cholesky factors still holds for $s \leq d/2$ if the local averages of Example 5.2 are subsampled as in Example 5.1, whereas the low-rank approximation becomes suboptimal. As illustrated in Table 4.5, for Matérn kernels we observe no difference (in accuracy vs. complexity) between integer and noninteger values of s.

6.2. Sparse factorization of $A = \Theta^{-1}$. Let $LL^\top = \Theta$ be the Cholesky factorization of the covariance matrix Θ. Writing $P^\updownarrow$ for the order-reversing permutation,

$$P^\updownarrow \Theta^{-1} P^\updownarrow = P^\updownarrow L^{-\top} L^{-1} P^\updownarrow = \left(P^\updownarrow L^{-\top} P^\updownarrow\right)\left(P^\updownarrow L^{-1} P^\updownarrow\right). \tag{6.1}$$

Since $P^\updownarrow L^{-\top} P^\updownarrow$ is lower triangular, it is the Cholesky factor of Θ^{-1} in the reverse elimination ordering. Furthermore, since $L^{-\top} = AL$ and both A and L are exponentially decaying, the Cholesky factors of A are also exponentially decaying if the Gaussian elimination is performed using the reverse of section 2's ordering. In fact, the following, stronger, theorem holds.

THEOREM 6.1. *In the setting of Theorem* 3.1*, let*

$$\mathring{S}_\rho \coloneqq \{(i,j) \in I \times I \,|\, \operatorname{dist}(\operatorname{supp}(\phi_i), \operatorname{supp}(\phi_j)) \leq \rho \min(l[i], l[j])\}, \tag{6.2}$$

let L be the Cholesky factor of A in the reverse ordering, and define

$$L_{ij}^{\mathring{S}_\rho} \coloneqq \begin{cases} L_{ij} & \text{for } (i,j) \in \mathring{S}_\rho, \\ 0 & \text{else.} \end{cases} \tag{6.3}$$

Then there exists a constant C depending only on d, Ω, s, $\|\mathcal{L}\|$, $\|\mathcal{L}^{-1}\|$, and δ such that for $\rho \geq C \log(N/\epsilon)$, we have $\left\|PAP - L^{\mathring{S}_\rho} L^{\mathring{S}_\rho,\top}\right\|_{\mathrm{Fro}} \leq \epsilon$.

Using this result and the fact that $\#\mathring{S}_\rho$ has $\mathcal{O}(\rho^d + 1)$ nonzero entries per column, one can prove that using Algorithm 2.2 with a supernodal ordering as described in Construction 5.25 yields an ϵ-approximate Cholesky factorization of A in computational complexity $\mathcal{O}(N \log(N/\epsilon)^{2d})$ in time and $\mathcal{O}(N \log(N/\epsilon)^d)$ in space. The matrix A is *essentially* a discretized elliptic partial differential operator, and analogous results can be obtained in the setting where A is obtained as a discretization of $\mathcal{L}$ with regular finite elements and Θ is the inverse of that discretized operator. Numerical experiments suggest that exponential decay properties also hold for discretized

second-order elliptic equations in two or three dimensions (where $s = 1 \leq d/2$) when using subsampling as in Example 5.1; see [74, section 3.1] for a special case of this result on regular meshes. Thus, by computing the incomplete Cholesky factorization, we obtain a direct solver for general elliptic PDEs with complexity $\mathcal{O}(N \log(N/\epsilon)^{2d})$ in time and $\mathcal{O}(N \log(N/\epsilon)^{d})$ in space. To the best of our knowledge, this is the best asymptotic complexity reported for such a solver in the literature (for elliptic PDEs with rough coefficients and rigorous a priori estimates of complexity vs. accuracy). It is not surprising that we obtain a fast solver for elliptic PDEs because our work is based on the fast solvers introduced in [62, 63], which in turn can be shown to be a blockwise version of the Cholesky factorization in nonstandard form introduced by [31], where the inverses of diagonal blocks are computed using iterative methods. By instead applying the Cholesky factorization in nonstandard form, the logarithmic factor in the complexity of the gamblet transform can be improved. However, the error estimates of [62] and [63] improve significantly upon those in [31] by establishing that exponential accuracy can be obtained with a finite number of vanishing moments even for rough coefficients. The present work further extends the results on Cholesky factorization to the setting of multiresolution schemes based on subsampling (without any vanishing moments). For such multiresolution basis the nonstandard form just reduces to computing an ordinary incomplete Cholesky factorization with the smaller sparsity pattern $\mathring{S}_\rho$, thus greatly simplifying the implementation. We note that by using direct inversion methods similar to [53] it would be possible in principle to directly compute ϵ-approximations of the Cholesky factors of Θ^{-1} from $\mathcal{O}(N \log(N/\epsilon)^{d})$ entries of Θ at computational cost of $\mathcal{O}(N \log(N/\epsilon)^{2d})$, but we defer a more detailed investigation to future work.

7. Comparison to related work.

7.1. $\mathcal{H}$-matrix approximations from sparse Cholesky factorization.

The $\mathcal{H}$-matrix data structure [37] uses low-rank approximations for blocks $\Theta_{\bar{I}\bar{J}}$ ($\bar{I}, \bar{J} \subset I$) fulfilling the admissibility condition

$$\min\big(\operatorname{diam}\{x_i\}_{i\in\bar{I}}, \operatorname{diam}\{x_i\}_{i\in\bar{J}}\big) \leq \eta \operatorname{dist}\big(\{x_i\}_{i\in\bar{I}}, \{x_i\}_{i\in\bar{J}}\big). \tag{7.1}$$

The approximation property of the incomplete Cholesky factorization in maximin ordering (Theorem 3.1) directly implies bounds on the spectral decay of admissible blocks in the $\mathcal{H}$-matrix framework, as can be seen from the representation

$$\Theta = LL^\top \iff \Theta = \sum_{i=1}^{N} L_{:i} \otimes L_{:i} \tag{7.2}$$

of the Cholesky factorization of Θ. If L is sparse according to the sparsity pattern obtained in section 2, then $L_{:i} \otimes L_{:i}$ can contribute to the rank of the submatrix $\Theta_{\bar{I}\bar{J}}$ only if

$$2\rho l[i] \geq \operatorname{dist}\left(\{x_j\}_{j\in\bar{I}}, \{x_j\}_{j\in\bar{J}}\right) \text{ and } \max\left(\operatorname{dist}\left(x_i, \{x_j\}_{j\in\bar{I}}\right), \operatorname{dist}\left(x_i, \{x_j\}_{j\in\bar{J}}\right)\right) \leq \rho l[i]. \tag{7.3}$$

The number of $i \in I$ satisfying (7.3) is at most $C(\eta, d)\rho^d \log N$, which recovers (up to constants) the same rank bounds as obtained in [7] for second-order elliptic PDEs with rough coefficients. However, the converse is not true and most hierarchical matrix representations cannot be written in terms of a sparse Cholesky factorization of Θ. For example, adding a diagonal matrix to Θ does not affect the ranks of admissible blocks, but it diminishes the screening effect and thus the approximation property of the incomplete Cholesky factorization as obtained in section 2 (see subsection 4.3).

7.2. Comparison to Cholesky factorization in wavelet bases. [31] computes sparse Cholesky factorizations of (discretized) differential/integral operators represented in a wavelet basis. Using a *fine-to-coarse* elimination ordering, they establish that the resulting Cholesky factors decay polynomially with an exponent matching the number of vanishing moments of the underlying wavelet basis.

For differential operators, this coincides algorithmically with the Cholesky factorization described in subsection 6.2 and the gamblet transform of [62] and [63], whose estimates guarantee exponential decay. In particular, Gines, Beylkin, and Dunn [31] numerically observe a uniform bound on $\mathrm{cond}(B^{(k)})$ which they relate to the approximate sparsity of their proposed Cholesky factorization.

For integral operators, [31] uses a *fine-to-coarse* ordering and we use a *coarse-to-fine* ordering. While their results rely on the approximate sparsity of the integral operator represented in the wavelet basis, our approximation remains accurate for multiresolution bases (e.g., the maximin ordering in section 2) in which Θ is dense, which avoids the $\mathcal{O}(N^2)$ complexity of a basis transform (or the implementation of adaptive quadrature rules to mitigate this cost).

7.3. Vanishing moments. Let $\mathcal{P}^{s-1}(\tau)$ denote the set of polynomials of order at most $s-1$ that are supported on $\tau \subset \Omega$. [62] and [63] show that (5.18) and (5.17) hold when $\mathcal{L}$ is an elliptic partial differential operator of order s (as described in subsection 2.1) and the measurements are local polynomials of order up to $s-1$ (i.e., $\phi_{i,\alpha} = 1_{\tau_i} p_\alpha$ with $p_\alpha \in \mathcal{P}^{s-1}(\tau_i)$). Using these $\phi_{i,\alpha}$ as measurements is equivalent to using wavelets ϕ_i satisfying the vanishing moment condition

$$[\phi_i, p] = 0 \quad \forall i \in I, p \in \mathcal{P}^{s-1}. \tag{7.4}$$

The requirement for vanishing moments has three important consequences. First, it requires that the order of the operator be known a priori, so that a suitable number of vanishing moments can be ensured. Second, ensuring a suitable number of vanishing moments greatly increases the complexity of the implementation. Third, in order to provide vanishing moments, the measurements ϕ_i, $i \in J^{(k)}$, have to be obtained from weighted averages over domains of size of order h^k. Therefore, even computing the first entry of the matrix Θ in the multiresolution basis will have complexity $\mathcal{O}(N^2)$, since it requires taking an average over almost all of $I \times I$. One of the main analytical results of this paper is to show that these vanishing moment conditions and local averages are not necessary for higher-order operators (which, in particular, enables the generalization of the gamblet transform to hierarchies of measurements defined as in Examples 5.1 and 5.2).

7.4. Comparison to multiresolution approximation (M-RA). In spatial statistics, the method most closely related to ours is the M-RA of [48], where a Gaussian process is approximated by a sum, at different scales, of predictive processes described in [5]. Following the intuition of the screening effect, these processes are assumed to be block-independent with respect to a domain decomposition at the respective scale, allowing for near-linear computational complexity. Although the specific multiresolution scheme and its accuracy are a function of the specific choice of basis functions and of the *knots* to be conditioned upon at each scale, no systematic strategy and no theoretical error bounds are provided for best accuracy. We suspect that no scheme relying on block-sparsity assumptions can also guarantee exponential accuracy in near-linear computational complexity, though we note that the *taper-M-RA* introduced by [49], independently of and after the first version of the present

article, does not impose conditional block-independence and could therefore be made exponentially accurate. While our present work and that of [48] are both motivated by a hierarchical exploitation of the screening effect, we identify a concrete and simple algorithm that has a guaranteed exponential accuracy for a wide range of kernel matrices.

8. Conclusions. We have shown that the dense covariance matrices obtained from a wide range of covariance functions associated to smooth Gaussian processes have almost sparse Cholesky factors. Using this property, these matrices can be inverted in near-linear computational complexity just by applying zero fill-in incomplete Cholesky factorization with an a priori ordering and sparsity pattern. Sparse Cholesky factorization of sparse matrices is by now a classical field, but we are not aware of prior work on the sparse factorization of dense matrices, other than for the purpose of preconditioning. While our algorithm is subject to the curse of high dimensionality like other hierarchy-based methods, it is able to exploit low dimensionality in the data without any user intervention. Our results are motivated by the probabilistic interpretation of Cholesky factorization and proved rigorously by using and generalizing recent results on operator-adapted wavelets. By reversing the elimination order, we also obtain a fast direct solver for elliptic PDEs whose rigorous a priori accuracy-vs.-complexity estimates advance the current state of the art for general elliptic PDEs.

Acknowledgments. We would like to thank C. Oates and P. Schröder for helpful discussions, and C. Scovel for many helpful comments and suggestions on an earlier version of this paper (June 2017, arXiv:1706.02205). We would like to thank the two anonymous referees for their constructive feedback, which helped to improve the paper.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards Applied Mathematics Series 55, U.S. Government Printing Office, Washington, D.C., 1964.

[2] S. Ambikasaran and E. Darve, *An $\mathcal{O}(N \log N)$ fast direct solver for partial hierarchically semi-separable matrices*, J. Sci. Comput., 57 (2013), pp. 477–501, https://doi.org/10.1007/s10915-013-9714-z.

[3] S. Ambikasaran, D. Foreman-Mackey, L. Greengard, D. W. Hogg, and M. O'Neil, *Fast direct methods for Gaussian processes*, IEEE Trans. Pattern Anal. Mach. Intell., 38 (2016), pp. 252–265, https://doi.org/10.1109/TPAMI.2015.2448083.

[4] F. R. Bach and M. I. Jordan, *Kernel independent component analysis*, J. Mach. Learn. Res., 3 (2003), pp. 1–48.

[5] S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang, *Gaussian predictive process models for large spatial data sets*, J. R. Stat. Soc. Ser. B Stat. Methodol., 70 (2008), pp. 825–848, https://doi.org/10.1111/j.1467-9868.2008.00663.x.

[6] M. Bebendorf, *Hierarchical Matrices*, Lect. Notes Comput. Sci. Eng. 63, Springer, New York, 2008, https://doi.org/10.1007/978-3-540-77147-0.

[7] M. Bebendorf and W. Hackbusch, *Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients*, Numer. Math., 95 (2003), pp. 1–28, https://doi.org/10.1007/s00211-002-0445-6.

[8] M. Bebendorf and S. Rjasanow, *Adaptive low-rank approximation of collocation matrices*, Computing, 70 (2003), pp. 1–24, https://doi.org/10.1007/s00607-002-1469-6.

[9] M. Benzi, *Localization in matrix computations: Theory and applications*, in Exploiting Hidden Structure in Matrix Computations: Algorithms and Applications: Cetraro, Italy 2015, M. Benzi and V. Simoncini, eds., Springer, New York, 2016, pp. 211–317, https://doi.org/10.1007/978-3-319-49887-4_4.

[10] M. Benzi and V. Simoncini, *Decay bounds for functions of Hermitian matrices with banded or Kronecker structure*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1263–1282, https://doi.org/10.1137/151006159.

[11] M. Benzi and M. Tůma, *Orderings for factorized sparse approximate inverse preconditioners*, SIAM J. Sci. Comput., 21 (2000), pp. 1851–1868, https://doi.org/10.1137/S1064827598339372.

[12] G. Beylkin, R. Coifman, and V. Rokhlin, *Fast wavelet transforms and numerical algorithms.* I, Comm. Pure Appl. Math., 44 (1991), pp. 141–183, https://doi.org/10.1002/cpa.3160440202.

[13] S. Börm, *Approximation of solution operators of elliptic partial differential equations by $\mathcal{H}$- and $\mathcal{H}^2$-matrices*, Numer. Math., 115 (2010), pp. 165–193, https://doi.org/10.1007/s00211-009-0278-7.

[14] S. Börm, *Efficient Numerical Methods for Non-Local Operators: $\mathcal{H}^2$-Matrix Compression, Algorithms and Analysis*, EMS Tracts Math. 14, European Mathematical Society, Zürich, Switzerland, 2010, https://doi.org/10.4171/091.

[15] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004, https://doi.org/10.1017/CBO9780511804441.

[16] A. Brandt, *Multi-level adaptive techniques (MLAT) for partial differential equations: Ideas and software*, in Mathematical Software III, Publ. Math. Res. Center 39, Academic Press, New York, 1977, pp. 277–318.

[17] D. L. Brown, J. Gedicke, and D. Peterseim, *Numerical homogenization of heterogeneous fractional Laplacians*, Multiscale Model. Simul., 16 (2018), pp. 1305–1332, https://doi.org/10.1137/17M1147305.

[18] S. Chandrasekaran, M. Gu, and T. Pals, *A fast ULV decomposition solver for hierarchically semiseparable representations*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 603–622, https://doi.org/10.1137/S0895479803436652.

[19] Y. Daon and G. Stadler, *Mitigating the influence of the boundary on PDE-based covariance operators*, Inverse Probl. Imaging, 12 (2018), pp. 1083–1102, https://doi.org/10.3934/ipi.2018045.

[20] S. Dekel and D. Leviatan, *The Bramble–Hilbert lemma for convex domains*, SIAM J. Math. Anal., 35 (2004), pp. 1203–1212, https://doi.org/10.1137/S0036141002417589.

[21] S. Demko, W. F. Moss, and P. W. Smith, *Decay rates for inverses of band matrices*, Math. Comp., 43 (1984), pp. 491–499, https://doi.org/10.2307/2008290.

[22] R. P. Fedorenko, *A relaxation method of solution of elliptic difference equations*, Ž. Vyčisl. Mat. i Mat. Fiz., 1 (1961), pp. 922–927.

[23] M. Feischl and D. Peterseim, *Sparse Compression of Expected Solution Operators*, SIAM J. Numer. Anal., 58 (2020), pp. 3144–3164.

[24] S. Fine and K. Scheinberg, *Efficient SVM training using low-rank kernel representations*, J. Mach. Learn. Res., 2 (2001), pp. 243–264.

[25] J. Fitzsimons, D. Granziol, K. Cutajar, M. Osborne, M. Filippone, and S. Roberts, *Entropic trace estimates for log determinants*, in Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, New York, 2017, pp. 323–338, https://doi.org/10.1007/978-3-319-71249-9_20.

[26] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, *Spectral grouping using the Nyström method*, IEEE Trans. Pattern Anal. Mach. Intell., 26 (2004), pp. 214–225, https://doi.org/10.1109/TPAMI.2004.1262185.

[27] R. Furrer, M. G. Genton, and D. Nychka, *Covariance tapering for interpolation of large spatial datasets*, J. Comput. Graph. Statist., 15 (2006), pp. 502–523, https://doi.org/10.1198/106186006X132178.

[28] A. George, *Nested dissection of a regular finite element mesh*, SIAM J. Numer. Anal., 10 (1973), pp. 345–363, https://doi.org/10.1137/0710032.

[29] A. George and J. W. H. Liu, *The evolution of the minimum degree ordering algorithm*, SIAM Rev., 31 (1989), pp. 1–19, https://doi.org/10.1137/1031001.

[30] J. R. Gilbert and R. E. Tarjan, *The analysis of a nested dissection algorithm*, Numer. Math., 50 (1987), pp. 377–404, https://doi.org/10.1007/BF01396660.

[31] D. Gines, G. Beylkin, and J. Dunn, *LU factorization of non-standard forms and direct multiresolution solvers*, Appl. Comput. Harmon. Anal., 5 (1998), pp. 156–201, https://doi.org/10.1006/acha.1997.0227.

[32] T. Gneiting and M. Schlather, *Stochastic models that separate fractal dimension and the Hurst effect*, SIAM Rev., 46 (2004), pp. 269–282, https://doi.org/10.1137/S0036144501394387.

[33] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulations*, J. Comput. Phys., 73 (1987), pp. 325–348, https://doi.org/10.1016/0021-9991(87)90140-9.

[34] J. Guinness, *Permutation and grouping methods for sharpening Gaussian process approximations*, Technometrics, 60 (2018), pp. 415–429, https://doi.org/10.1080/00401706.2018.1437476.

[35] P. Guttorp and T. Gneiting, *Studies in the history of probability and statistics.* XLIX. *On the Matérn correlation family*, Biometrika, 93 (2006), pp. 989–995, https://doi.org/10.1093/biomet/93.4.989.

[36] W. Hackbusch, *A fast iterative method for solving Poisson's equation in a general region*, in Numerical Treatment of Differential Equations, Lecture Notes in Math. 631, Springer, New York, 1978, pp. 51–62.

[37] W. Hackbusch, *A sparse matrix arithmetic based on $\mathcal{H}$-matrices.* I. *Introduction to $\mathcal{H}$-matrices*, Computing, 62 (1999), pp. 89–108, https://doi.org/10.1007/s006070050015.

[38] W. Hackbusch, *Multi-Grid Methods and Applications*, Springer Ser. Comput. Math. 4, Springer, New York, 2013, https://doi.org/10.1007/978-3-662-02427-0.

[39] W. Hackbusch and S. Börm, *Data-sparse approximation by adaptive $\mathcal{H}^2$-matrices*, Computing, 69 (2002), pp. 1–35, https://doi.org/10.1007/s00607-002-1450-4.

[40] W. Hackbusch and B. N. Khoromskij, *A sparse $\mathcal{H}$-matrix arithmetic.* II. *Application to multi-dimensional problems*, Computing, 64 (2000), pp. 21–47.

[41] K. L. Ho and L. Ying, *Hierarchical interpolative factorization for elliptic operators: Integral equations*, Comm. Pure Appl. Math., 69 (2016), pp. 1314–1353, https://doi.org/10.1002/cpa.21577.

[42] T. Hofmann, B. Schölkopf, and A. J. Smola, *Kernel methods in machine learning*, Ann. Statist., 36 (2008), pp. 1171–1220, https://doi.org/10.1214/009053607000000677.

[43] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1994, https://doi.org/10.1017/CBO9780511840371.

[44] T. Y. Hou and P. Zhang, *Sparse operator compression of higher-order elliptic operators with rough coefficients*, Res. Math. Sci., 4 (2017), 24, https://doi.org/10.1186/s40687-017-0113-1.

[45] T. Husfeldt, *Graph colouring algorithms*, in Topics in Chromatic Graph Theory, Encyclopedia Math. Appl. 156, Cambridge University Press, Cambridge, 2015, pp. 277–303, https://doi.org/10.1017/CBO9781139519793.016.

[46] T. Iwashita and M. Shimasaki, *Block red-black ordering: A new ordering strategy for parallelization of ICCG method*, Int. J. Parallel Program., 31 (2003), pp. 55–75, https://doi.org/10.1023/A:1021738303840.

[47] S. Jaffard, *Propriétés des matrices "bien localisées" près de leur diagonale et quelques applications*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 461–476.

[48] M. Katzfuss, *A multi-resolution approximation for massive spatial datasets*, J. Amer. Statist. Assoc., 112 (2017), pp. 201–214, https://doi.org/10.1080/01621459.2015.1123632.

[49] M. Katzfuss and W. Gong, *A class of multi-resolution approximations for large spatial datasets*, Statist. Sinica, 30 (2020), pp. 2203–2226.

[50] R. Kornhuber, D. Peterseim, and H. Yserentant, *An analysis of a class of variational multiscale methods based on subspace decomposition*, Math. Comp., 87 (2018), pp. 2765–2774, https://doi.org/10.1090/mcom/3302.

[51] I. Krishtal, T. Strohmer, and T. Wertz, *Localization of matrix factorizations*, Found. Comput. Math., 15 (2015), pp. 931–951, https://doi.org/10.1007/s10208-014-9196-x.

[52] S. Li, M. Gu, C. J. Wu, and J. Xia, *New efficient and robust HSS Cholesky factorization of SPD matrices*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 886–904, https://doi.org/10.1137/110851110.

[53] L. Lin, C. Yang, J. C. Meza, J. Lu, L. Ying, and W. E, *SelInv—An algorithm for selected inversion of a sparse symmetric matrix*, ACM Trans. Math. Software, 37 (2011), 40, https://doi.org/10.1145/1916461.1916464.

[54] F. Lindgren, H. Rue, and J. Lindström, *An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach*, J. R. Stat. Soc. Ser. B Stat. Methodol., 73 (2011), pp. 423–498, https://doi.org/10.1111/j.1467-9868.2011.00777.x.

[55] R. J. Lipton, D. J. Rose, and R. E. Tarjan, *Generalized nested dissection*, SIAM J. Numer. Anal., 16 (1979), pp. 346–358, https://doi.org/10.1137/0716027.

[56] J. W. H. Liu, E. G. Ng, and B. W. Peyton, *On finding supernodes for sparse matrix computations*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 242–252, https://doi.org/10.1137/0614019.

[57] A. Målqvist and D. Peterseim, *Localization of elliptic multiscale problems*, Math. Comp., 83 (2014), pp. 2583–2603, https://doi.org/10.1090/S0025-5718-2014-02868-8.

[58] P.-G. Martinsson, *Compressing rank-structured matrices via randomized sampling*, SIAM J. Sci. Comput., 38 (2016), pp. A1959–A1986, https://doi.org/10.1137/15M1016679.

[59] B. Matérn, *Spatial Variation: Stochastic Models and Their Application to Some Problems in Forest Surveys and Other Sampling Investigations*, Meddelanden Fran Statens Skogsforskningsinstitut, Stockholm, 1960.

[60] J. A. Meijerink and H. A. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (1977), pp. 148–162, https://doi.org/10.2307/2005786.

[61] C. A. Micchelli and T. J. Rivlin, *A survey of optimal recovery*, in Optimal Estimation in Approximation Theory, Plenum, New York, 1977, pp. 1–54, https://doi.org/10.1007/978-1-4684-2388-4_1.

[62] H. Owhadi, *Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games*, SIAM Rev., 59 (2017), pp. 99–149, https://doi.org/10.1137/15M1013894.

[63] H. Owhadi and C. Scovel, *Universal Scalable Robust Solvers from Computational Information Games and Fast Eigenspace Adapted Multiresolution Analysis*,arXiv:1703.10761v2, 2017.

[64] H. Owhadi and C. Scovel, *Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization: From a Game Theoretic Approach to Numerical Approximation and Algorithm Design*, Cambridge Monogr. Appl. Comput. Math. 35, Cambridge University Press, Cambridge, 2019, https://doi.org/10.1017/9781108594967.

[65] H. Owhadi, L. Zhang, and L. Berlyand, *Polyharmonic homogenization, rough polyharmonic splines and sparse super-localization*, ESAIM Math. Model. Numer. Anal., 48 (2014), pp. 517–552, https://doi.org/10.1051/m2an/2013118.

[66] J. Quiñonero-Candela and C. E. Rasmussen, *A unifying view of sparse approximate Gaussian process regression*, J. Mach. Learn. Res., 6 (2005), pp. 1939–1959.

[67] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Adapt. Comput. Mach. Learn., MIT Press, Cambridge, MA, 2006, https://doi.org/10.7551/mitpress/3206.001.0001.

[68] L. Roininen, J. M. J. Huttunen, and S. Lasanen, *Whittle–Matérn priors for Bayesian statistical inversion with applications in electrical impedance tomography*, Inverse Probl. Imaging, 8 (2014), pp. 561–586, https://doi.org/10.3934/ipi.2014.8.561.

[69] L. Roininen, M. S. Lehtinen, S. Lasanen, M. Orispää, and M. Markkanen, *Correlation priors*, Inverse Probl. Imaging, 5 (2011), pp. 167–184, https://doi.org/10.3934/ipi.2011.5.167.

[70] L. Roininen, P. Piiroinen, and M. Lehtinen, *Constructing continuous stationary covariances as limits of the second-order stochastic difference equations*, Inverse Probl. Imaging, 7 (2013), pp. 611–647, https://doi.org/10.3934/ipi.2013.7.611.

[71] E. Rothberg and A. Gupta, *An efficient block-oriented approach to parallel sparse Cholesky factorization*, SIAM J. Sci. Comput., 15 (1994), pp. 1413–1439, https://doi.org/10.1137/0915085.

[72] A. K. Saibaba, A. Alexanderian, and I. C. F. Ipsen, *Randomized matrix-free trace and log-determinant estimators*, Numer. Math., 137 (2017), pp. 353–395, https://doi.org/10.1007/s00211-017-0880-z.

[73] H. Sang and J. Z. Huang, *A full scale approximation of covariance functions for large spatial data sets*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 74 (2012), pp. 111–132, https://doi.org/10.1111/j.1467-9868.2011.01007.x.

[74] J. Schröder, U. Trottenberg, and K. Witsch, *On fast Poisson solvers and applications*, in Numerical Treatment of Differential Equations, Lecture Notes in Math. 631, Springer, New York, 1978, pp. 153–187, https://doi.org/10.1007/BFb0067471.

[75] A. Schwaighofer and V. Tresp, *Transductive and inductive methods for approximate Gaussian process regression*, in Proceedings of Advances in Neural Information Processing Systems 15 (NIPS 2002), S. Becker, S. Thrun, and K. Obermayer, eds., 2003, pp. 977–984.

[76] J. R. Shewchuk, *An Introduction to the Conjugate Gradient Method Without the Agonizing Pain*, Technical report, Department of Computer Science, Carnegie-Mellon University, 1994; also available online from https://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf.

[77] A. J. Smola and P. L. Bartlett, *Sparse greedy Gaussian process regression*, in Proceedings of Advances in Neural Information Processing Systems 13 (NIPS 2000), 2001, pp. 619–625.

[78] E. Snelson and Z. Ghahramani, *Sparse Gaussian processes using pseudo-inputs*, in Proceedings of Advances in Neural Information Processing Systems 18, Y. Weiss, P. B. Schölkopf, and J. C. Platt, eds., MIT Press, Cambridge, MA, 2006, pp. 1257–1264.

[79] M. L. Stein, 2010 *Rietz Lecture: When does the screening effect hold?*, Ann. Statist., 39 (2011), pp. 2795–2819, https://doi.org/10.1214/11-AOS909.

[80] P. Whittle, *On stationary processes in the plane*, Biometrika, 41 (1954), pp. 434–449, https://doi.org/10.1093/biomet/41.3-4.434.
[81] P. Whittle, *Stochastic processes in several dimensions*, Bull. Inst. Internat. Statist., 40 (1963), pp. 974–994.
[82] C. K. I. Williams and M. Seeger, *Using the Nyström method to speed up kernel machines*, in Proceedings of Advances in Neural Information Processing Systems 13, T. K. Leen, T. G. Dietterich, and V. Tresp, eds., MIT Press, Cambridge, MA, 2001, pp. 682–688.
[83] J. Xia, S. Chandrasekaran, M. Gu, and X. S. Li, *Fast algorithms for hierarchically semiseparable matrices*, Numer. Linear Algebra Appl., 17 (2010), pp. 953–976, https://doi.org/10.1002/nla.691.
[84] F. Zhang, ed., *The Schur Complement and Its Applications*, Numerical Methods and Algorithms 4, Springer, New York, 2005, https://doi.org/10.1007/b105056.