Foundations of Computational Mathematics
https://doi.org/10.1007/s10208-022-09556-w

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Learning Elliptic Partial Differential Equations with Randomized Linear Algebra

Nicolas Boullé[1] · **Alex Townsend**[2]

Received: 31 January 2021 / Revised: 18 November 2021 / Accepted: 20 November 2021

Abstract
Given input–output pairs of an elliptic partial differential equation (PDE) in three dimensions, we derive the first theoretically rigorous scheme for learning the associated Green's function G. By exploiting the hierarchical low-rank structure of G, we show that one can construct an approximant to G that converges almost surely and achieves a relative error of $\mathcal{O}(\Gamma_\epsilon^{-1/2}\log^3(1/\epsilon)\epsilon)$ using at most $\mathcal{O}(\epsilon^{-6}\log^4(1/\epsilon))$ input–output training pairs with high probability, for any $0 < \epsilon < 1$. The quantity $0 < \Gamma_\epsilon \leq 1$ characterizes the quality of the training dataset. Along the way, we extend the randomized singular value decomposition algorithm for learning matrices to Hilbert–Schmidt operators and characterize the quality of covariance kernels for PDE learning.

Keywords Data-driven discovery of PDEs · Randomized SVD · Green's function · Hilbert–Schmidt operators · Low-rank approximation

Mathematics Subject Classification 65N80 · 35J08 · 35R30 · 60G15 · 65F55

Communicated by Arieh Iserles.

This work is supported by the EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling (EP/L015803/1) in collaboration with Simula Research Laboratory and by the National Science Foundation Grants DMS-1818757, DMS-1952757, and DMS-2045646.

✉ Nicolas Boullé
boulle@maths.ox.ac.uk

Alex Townsend
townsend@cornell.edu

[1] Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK

[2] Department of Mathematics, Cornell University, Ithaca, NY 14853, USA

Published online: 18 January 2022

1 Introduction

Can one learn a differential operator from pairs of solutions and righthand sides? If so, how many pairs are required? These two questions have received significant research attention [17,31,34,43]. From data, one hopes to eventually learn physical laws of nature or conservation laws that elude scientists in the biological sciences [63], computational fluid dynamics [49], and computational physics [45]. The literature contains many highly successful practical schemes based on deep learning techniques [38,48]. However, the challenge remains to understand when and why deep learning is effective theoretically. This paper describes the first theoretically justified scheme for discovering scalar-valued elliptic partial differential equations (PDEs) in three variables from input–output data and provides a rigorous learning rate. While our novelties are mainly theoretical, we hope to motivate future practical choices in PDE learning.

We suppose that there is an unknown second-order uniformly elliptic linear PDE operator[1] $\mathcal{L} : \mathcal{H}^2(D) \cap \mathcal{H}_0^1(D) \to L^2(D)$ with a bounded domain $D \subset \mathbb{R}^3$ with Lipschitz smooth boundary [16], which takes the form

$$(\mathcal{L}u(x) = -\nabla \cdot (A(x)\nabla u) + c(x) \cdot \nabla u + d(x)u, \quad x \in D, \quad u|_{\partial D} = 0. \tag{1}$$

Here, for every $x \in D$, we have that $A(x) \in \mathbb{R}^{3\times 3}$ is a symmetric positive definite matrix with bounded coefficient functions so that[2] $A_{ij} \in L^\infty(D)$, $c \in L^r(D)$ with $r \geq 3$, $d \in L^s(D)$ for $s \geq 3/2$, and $d(x) \geq 0$ [28]. We emphasize that the regularity requirements on the variable coefficients are quite weak.

The goal of PDE learning is to discover the operator $\mathcal{L}$ from $N \geq 1$ input–output pairs, i.e., $\{(f_j, u_j)\}_{j=1}^N$, where $\mathcal{L}u_j = f_j$ and $u_j|_{\partial D} = 0$ for $1 \leq j \leq N$. There are two main types of PDE learning tasks: (1) Experimentally determined input–output pairs, where one must do the best one can with the predetermined information and (2) algorithmically determined input–output pairs, where the data-driven learning algorithm can select $f_1, \ldots, f_N$ for itself. In this paper, we focus on the PDE learning task where we have algorithmically determined input–output pairs. In particular, we suppose that the functions $f_1, \ldots, f_N$ are generated at random and are drawn from a Gaussian process (GP) (see Sect. 2.3). To keep our theoretical statements manageable, we restrict our attention to PDEs of the form:

$$\mathcal{L}u = -\nabla \cdot (A(x)\nabla u), \quad x \in D, \quad u|_{\partial D} = 0. \tag{2}$$

Lower-order terms in Eq. (1) should cause few theoretical problems [3], though our algorithm and our bounds get far more complicated.

[1] Here, $L^2(D)$ is the space of square-integrable functions defined on D, $\mathcal{H}^k(D)$ is the space of k times weakly differentiable functions in the L^2-sense, and $\mathcal{H}_0^1(D)$ is the closure of $\mathcal{C}_c^\infty(D)$ in $\mathcal{H}^1(D)$. Here, $\mathcal{C}_c^\infty(D)$ is the space of infinitely differentiable compactly supported functions on D. Roughly speaking, $\mathcal{H}_0^1(D)$ are the functions in $\mathcal{H}^1(D)$ that are zero on the boundary of D.

[2] For $1 \leq r \leq \infty$, we denote by $L^r(D)$ the space of functions defined on the domain D with finite L^r norm, where $\|f\|_r = (\int_D |f|^r \mathrm{d}x)^{1/r}$ if $r < \infty$, and $\|f\|_\infty = \inf\{C > 0 : |f(x)| \leq C \text{ for almost every } x \in D\}$.

The approach that dominates the PDE learning literature is to directly learn $\mathcal{L}$ by either (1) learning parameters in the PDE [4,64], (2) using neural networks to approximate the action of the PDE on functions [45–49], or (3) deriving a model by composing a library of operators with sparsity considerations [9,35,52,53,59,60]. Instead of trying to learn the unbounded, closed operator $\mathcal{L}$ directly, we follow [6,17,18] and discover the Green's function associated with $\mathcal{L}$. That is, we attempt to learn the function $G : D \times D \to \mathbb{R}^+ \cup \{\infty\}$ such that [16]

$$u_j(x) = \int_D G(x, y) f_j(y) \mathrm{d}y, \qquad x \in D, \qquad 1 \leq j \leq N. \tag{3}$$

Seeking G, as opposed to $\mathcal{L}$, has several theoretical benefits:

1. The integral operator in Eq. (3) is compact [15], while $\mathcal{L}$ is only closed [14]. This allows G to be rigorously learned by input–output pairs $\{(f_j, u_j)\}_{j=1}^N$, as its range can be approximated by finite-dimensional spaces (see Theorem 3).
2. It is known that G has a hierarchical low-rank structure [3, Theorem 2.8]: for $0 < \epsilon < 1$, there exists a function $G_k(x, y) = \sum_{j=1}^k g_j(x) h_j(y)$ with $k = \mathcal{O}(\log^4(1/\epsilon))$ such that [3, Theorem 2.8]

$$\|G - G_k\|_{L^2(X\times Y)} \leq \epsilon \|G\|_{L^2(X\times \hat{Y})},$$

   where $X, Y \subseteq D$ are sufficiently separated domains, and $Y \subseteq \hat{Y} \subseteq D$ denotes a larger domain than Y (see Theorem 4 for the definition). The further apart X and Y, the faster the singular values of G decay. Moreover, G also has an off-diagonal decay property [19,25]:

$$G(x, y) \leq \frac{c}{\|x - y\|_2} \|G\|_{L^2(D\times D)}, \qquad x \neq y \in D,$$

   where c is a constant independent of x and y. Exploiting these structures of G leads to a rigorous algorithm for constructing a global approximant to G (see Sect. 4).
3. The function G is smooth away from its diagonal, allowing one to efficiently approximate it [19].

Once a global approximation $\tilde{G}$ has been constructed for G using input–output pairs, given a new righthand side f one can directly compute the integral in Eq. (3) to obtain the corresponding solution u to Eq. (1). Usually, numerically computing the integral in Eq. (3) must be done with sufficient care as G possesses a singularity when $x = y$. However, our global approximation $\tilde{G}$ has an hierarchical structure and is constructed as 0 near the diagonal. Therefore, for each fixed $x \in D$, we simply recommend that $\int_D \tilde{G}(x, y) f_j(y) \mathrm{d}y$ is partitioned into the panels that corresponds to the hierarchical decomposition, and then discretized each panel with a quadrature rule.

1.1 Main Contributions

There are two main contributions in this paper: (1) the generalization of the randomized singular value decomposition (SVD) algorithm for learning matrices from matrix-vector products to Hilbert–Schmidt (HS) operators and (2) a theoretical learning rate for discovering Green's functions associated with PDEs of the form Eq. (2). These contributions are summarized in Theorem 1 and 3.

Theorem 1 says that, with high probability, one can recover a near-best rank k HS operator using $k + p$ operator-function products, for a small integer p. In the bound of the theorem, a quantity, denoted by $0 < \gamma_k \leq 1$, measures the quality of the input–output training pairs (see Sects. 3.1 and 3.4). We then combine Theorem 1 with the theory of Green's functions for elliptic PDEs to derive a theoretical learning rate for PDEs.

In Theorem 3, we show that Green's functions associated with uniformly elliptic PDEs in three dimensions can be recovered using $N = \mathcal{O}(\epsilon^{-6}\log^4(1/\epsilon))$ input–output pairs $(f_j, u_j)_{j=1}^N$ to within an accuracy of $\mathcal{O}(\Gamma_\epsilon^{-1/2}\log^3(1/\epsilon)\epsilon)$ with high probability, for $0 < \epsilon < 1$. Our learning rate associated with uniformly elliptic PDEs in three variables is therefore $\mathcal{O}(\epsilon^{-6}\log^4(1/\epsilon))$. The quantity $0 < \Gamma_\epsilon \leq 1$ (defined in Sect. 4.4.2) measures the quality of the GP used to generate the random functions $\{f_j\}_{j=1}^N$ for learning G. We emphasize that the number of training pairs is small only if the GP's quality is high. The probability bound in Theorem 3 implies that the constructed approximation is close to G with high probability and converges almost surely to the Green's function as $\epsilon \to 0$.

1.2 Organization of Paper

The paper is structured as follows. In Sect. 2, we briefly review HS operators and GPs. We then generalize the randomized SVD algorithm to HS operators in Sect. 3. Next, in Sect. 4, we characterize the learning rate for PDEs of the form of Eq. (2) (see Theorem 3). Finally, we conclude and discuss potential further directions in Sect. 5.

2 Background Material

We begin by reviewing quasimatrices (see Sect. 2.1), HS operators (see Sect. 2.2), and GPs (see Sect. 2.3).

2.1 Quasimatrices

Quasimatrices are an infinite-dimensional analogue of tall-skinny matrices [57]. Let $D_1, D_2 \subseteq \mathbb{R}^d$ be two domains with $d \geq 1$ and denote by $L^2(D_1)$ the space of square-integrable functions defined on D_1. Many of results in this paper are easier to state using quasimatrices. We say that $\mathbf{\Omega}$ is a $D_1 \times k$ quasimatrix, if $\mathbf{\Omega}$ is a matrix with k

columns where each column is a function in $L^2(D_1)$. That is,

$$\boldsymbol{\Omega} = \big[\omega_1 \,|\, \cdots \,|\, \omega_k\big], \qquad \omega_j \in L^2(D_1).$$

Quasimatrices are useful to define analogues of matrix operations for HS operators [11,56–58]. For example, if $\mathscr{F} : L^2(D_1) \to L^2(D_2)$ is a HS operator, then we write $\mathscr{F}\boldsymbol{\Omega}$ to denote the quasimatrix obtained by applying $\mathscr{F}$ to each column of $\boldsymbol{\Omega}$. Moreover, we write $\boldsymbol{\Omega}^*\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}\boldsymbol{\Omega}^*$ to mean the following:

$$\boldsymbol{\Omega}^*\boldsymbol{\Omega} = \begin{bmatrix} \langle \omega_1, \omega_1 \rangle & \cdots & \langle \omega_1, \omega_k \rangle \\ \vdots & \ddots & \vdots \\ \langle \omega_k, \omega_1 \rangle & \cdots & \langle \omega_k, \omega_k \rangle \end{bmatrix}, \qquad \boldsymbol{\Omega}\boldsymbol{\Omega}^* = \sum_{j=1}^{k} \omega_j(x)\omega_j(y),$$

where $\langle \cdot, \cdot \rangle$ is the $L^2(D_1)$ inner-product. Many operations for rectangular matrices in linear algebra can be generalized to quasimatrices [57].

2.2 Hilbert–Schmidt Operators

HS operators are an infinite-dimensional analogue of matrices acting on vectors. Since $L^2(D_1)$ is a separable Hilbert space, there is a complete orthonormal basis $\{e_j\}_{j=1}^{\infty}$ for $L^2(D_1)$. We call $\mathscr{F} : L^2(D_1) \to L^2(D_2)$ a HS operator [23, Ch. 4] with HS norm $\|\mathscr{F}\|_{\mathrm{HS}}$ if $\mathscr{F}$ is linear and

$$\|\mathscr{F}\|_{\mathrm{HS}} := \left(\sum_{j=1}^{\infty} \|\mathscr{F} e_j\|_{L^2(D_2)}^2 \right)^{1/2} < \infty.$$

The archetypical example of an HS operator is an HS integral operator $\mathscr{F} : L^2(D_1) \to L^2(D_2)$ defined by

$$(\mathscr{F} f)(x) = \int_{D_1} G(x, y) f(y) \mathrm{d}y, \qquad f \in L^2(D_1), \quad x \in D_2,$$

where $G \in L^2(D_2 \times D_1)$ is the kernel of $\mathscr{F}$ and $\|\mathscr{F}\|_{\mathrm{HS}} = \|G\|_{L^2(D_2 \times D_1)}$. Since HS operators are compact operators, they have an SVD [23, Theorem 4.3.1]. That is, there exists a nonnegative sequence $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ and an orthonormal basis $\{q_j\}_{j=1}^{\infty}$ for $L^2(D_2)$ such that for any $f \in L^2(D_1)$ we have

$$\mathscr{F} f = \sum_{\substack{j=1 \\ \sigma_j > 0}}^{\infty} \sigma_j \langle e_j, f \rangle q_j, \tag{4}$$

where the equality holds in the $L^2(D_2)$ sense. Note that we use the complete SVD, which includes singular functions associated with the kernel of $\mathscr{F}$. Moreover, one finds

that $\|\mathscr{F}\|_{\mathrm{HS}}^2 = \sum_{j=1}^{\infty} \sigma_j^2$, which shows that the HS norm is an infinite-dimensional analogue of the Frobenius matrix norm $\|\cdot\|_{\mathrm{F}}$. In the same way that truncating the SVD after k terms gives a best rank k matrix approximation, truncating Eq. (4) gives a best approximation in the HS norm. That is, [23, Theorem 4.4.7]

$$\|\mathscr{F} - \mathscr{F}_k\|_{\mathrm{HS}}^2 = \sum_{j=k+1}^{\infty} \sigma_j^2, \qquad \mathscr{F}_k f = \sum_{j=1}^{k} \sigma_j \langle e_j, f \rangle q_j, \quad f \in L^2(D_1).$$

In this paper, we are interested in constructing an approximation to G in Eq. (3) from input–output pairs $\{(f_j, u_j)\}_{j=1}^N$ such that $u_j = \mathscr{F} f_j$.

Throughout this paper, the HS operator denoted by $\boldsymbol{\Omega}\boldsymbol{\Omega}^*\mathscr{F} : L^2(D_1) \to L^2(D_2)$ is given by $\boldsymbol{\Omega}\boldsymbol{\Omega}^*\mathscr{F} f = \sum_{j=1}^{k} \langle \omega_j, \mathscr{F} f \rangle \omega_j$. If we consider the operator $\boldsymbol{\Omega}^*\mathscr{F} : L^2(D_1) \to \mathbb{R}^k$, then $\|\boldsymbol{\Omega}^*\mathscr{F}\|_{\mathrm{HS}}^2 = \sum_{j=1}^{\infty} \|\mathscr{F} e_j\|_2^2$. Similarly, for $\mathscr{F}\boldsymbol{\Omega} : \mathbb{R}^k \to L^2(D_2)$ we have $\|\mathscr{F}\boldsymbol{\Omega}\|_{\mathrm{HS}}^2 = \sum_{j=1}^{k} \|\mathscr{F}\tilde{e}_j\|_{L^2(D_2)}^2$, where $\{\tilde{e}_j\}_{j=1}^k$ is an orthonormal basis of $\mathbb{R}^k$. Moreover, if $\boldsymbol{\Omega}$ has full column rank then $\mathbf{P}_{\boldsymbol{\Omega}}\mathscr{F} = \boldsymbol{\Omega}(\boldsymbol{\Omega}^*\boldsymbol{\Omega})^{\dagger}\boldsymbol{\Omega}^*\mathscr{F}$ is the orthogonal projection of the range of $\mathscr{F}$ onto the column space of $\boldsymbol{\Omega}$. Here, $(\boldsymbol{\Omega}^*\boldsymbol{\Omega})^{\dagger}$ is the pseudo-inverse of $\boldsymbol{\Omega}^*\boldsymbol{\Omega}$.

2.3 Gaussian Processes

A GP is an infinite-dimensional analogue of a multivariate Gaussian distribution and a function drawn from a GP is analogous to a randomly generated vector. If $K : D \times D \to \mathbb{R}$ is a continuous symmetric positive semidefinite kernel, where $D \subseteq \mathbb{R}^d$ is a domain, then a GP is a stochastic process $\{X_t,\ t \geq 0\}$ such that for every finite set of indices $t_1, \ldots, t_n \geq 0$ the vector of random variables $(X_{t_1}, \ldots, X_{t_n})$ is a multivariate Gaussian distribution with mean $(0, \ldots, 0)$ and covariance $K_{ij} = K(t_i, t_j)$ for $1 \leq i, j \leq n$. We denote a GP with mean $(0, \ldots, 0)$ and covariance K by $\mathcal{GP}(0, K)$.

Since K is a continuous symmetric positive semidefinite kernel, it has nonnegative eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and there is an orthonormal basis of eigenfunctions $\{\psi_j\}_{j=1}^{\infty}$ of $L^2(D)$ such that [23, Theorem 4.6.5]:

$$K(x, y) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x) \psi_j(y), \quad \int_D K(x, y) \psi_j(y) \mathrm{d}y = \lambda_j \psi_j(x), \quad x, y \in D, \tag{5}$$

where the infinite sum is absolutely and uniformly convergent [39]. In addition, we define the trace of the covariance kernel K by $\mathrm{Tr}(K) := \sum_{j=1}^{\infty} \lambda_j < \infty$. The eigendecomposition of K gives an algorithm for generating functions from $\mathcal{GP}(0, K)$. In particular, if $\omega \sim \sum_{j=1}^{\infty} \sqrt{\lambda_j} c_j \psi_j$, where the coefficients $\{c_j\}_{j=1}^{\infty}$ are independent and identically distributed standard Gaussian random variables, then $\omega \sim \mathcal{GP}(0, K)$ [26,33]. We also have

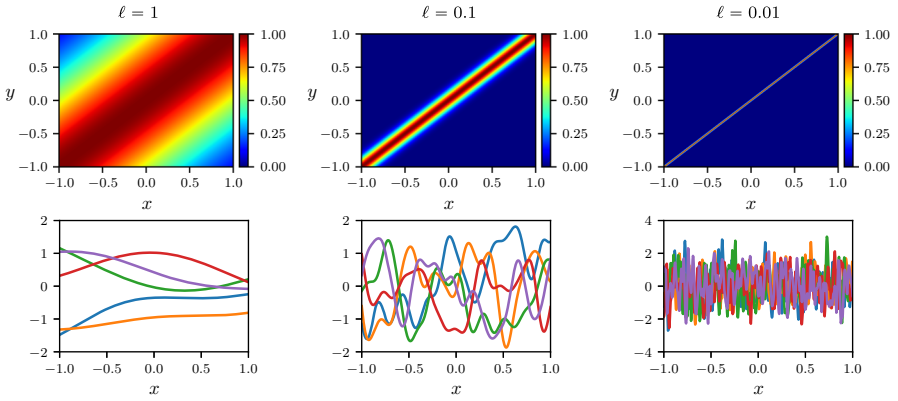

Fig. 1 Squared-exponential covariance kernel K_{SE} with parameter $\ell = 1, 0.1, 0.01$ (top row) and five functions sampled from $\mathcal{GP}(0, K_{\mathrm{SE}})$ (bottom row)

$$\mathbb{E}\Big[\|\omega\|^2_{L^2(D)}\Big] = \sum_{j=1}^{\infty} \lambda_j \mathbb{E}\Big[c_j^2\Big] \|\psi_j\|^2_{L^2(D)} = \sum_{j=1}^{\infty} \lambda_j = \int_D K(y, y)\, \mathrm{d}y < \infty,$$

where the last equality is analogous to the fact that the trace of a matrix is equal to the sum of its eigenvalues. In this paper, we restrict our attention to GPs with positive definite covariance kernels so that the eigenvalues of K are strictly positive.

In Fig. 1, we display the squared-exponential kernel defined as $K_{\mathrm{SE}}(x, y) = \exp(-|x - y|^2/(2\ell^2))$ for $x, y \in [-1, 1]$ [50, Chapt. 4] with parameters $\ell = 1, 0.1, 0.01$ together with sampled functions from $\mathcal{GP}(0, K_{\mathrm{SE}})$. We observe that the functions become more oscillatory as the length-scale parameter ℓ decreases and hence the numerical rank of the kernel increases or, equivalently, the associated eigenvalues $\{\lambda_j\}$ decay more slowly to zero.

3 Low-Rank Approximation of Hilbert–Schmidt Operators

In a landmark paper, Halko, Martinsson, and Tropp proved that one could learn the column space of a finite matrix—to high accuracy and with a high probability of success—by using matrix-vector products with standard Gaussian random vectors [22]. We now set out to generalize this from matrices to HS operators. Alternative randomized low-rank approximation techniques such as the generalized Nyström method [42] might also be generalized in a similar manner. Since the proof is relatively long, we state our final generalization now.

Theorem 1 *Let $D_1, D_2 \subseteq \mathbb{R}^d$ be domains with $d \geq 1$ and $\mathscr{F} : L^2(D_1) \to L^2(D_2)$ be a HS operator. Select a target rank $k \geq 1$, an oversampling parameter $p \geq 2$, and a $D_1 \times (k + p)$ quasimatrix $\mathbf{\Omega}$ such that each column is drawn from $\mathcal{GP}(0, K)$, where $K : D_1 \times D_1 \to \mathbb{R}$ is a continuous symmetric positive definite kernel with eigenvalues*

$\lambda_1 \geq \lambda_2 \geq \cdots > 0$. *If* $\mathbf{Y} = \mathscr{F}\boldsymbol{\Omega}$, *then*

$$\mathbb{E}[\|\mathscr{F} - \mathbf{P}_{\mathbf{Y}}\mathscr{F}\|_{\mathrm{HS}}] \leq \left(1 + \sqrt{\frac{1}{\gamma_k}\frac{k(k+p)}{p-1}}\right)\left(\sum_{j=k+1}^{\infty}\sigma_j^2\right)^{1/2}, \tag{6}$$

where $\gamma_k = k/(\lambda_1 \operatorname{Tr}(\mathbf{C}^{-1}))$ *with* $\mathbf{C}_{ij} = \int_{D_1 \times D_1} v_i(x)K(x,y)v_j(y)\mathrm{d}x\mathrm{d}y$ *for* $1 \leq i, j \leq k$. *Here,* $\mathbf{P}_{\mathbf{Y}}$ *is the orthogonal projection onto the vector space spanned by the columns of* $\mathbf{Y}$, σ_j *is the* j*th singular value of* $\mathscr{F}$, *and* v_j *is the* j*th right singular vector of* $\mathscr{F}$.

Assume further that $p \geq 4$, *then for any* $s, t \geq 1$, *we have*

$$\|\mathscr{F} - \mathbf{P}_{\mathbf{Y}}\mathscr{F}\|_{\mathrm{HS}} \leq \sqrt{1 + t^2 s^2 \frac{3}{\gamma_k}\frac{k(k+p)}{p+1}\sum_{j=1}^{\infty}\frac{\lambda_j}{\lambda_1}}\left(\sum_{j=k+1}^{\infty}\sigma_j^2\right)^{1/2}, \tag{7}$$

with probability $\geq 1 - t^{-p} - [se^{-(s^2-1)/2}]^{k+p}$.

We remark that the term $[se^{-(s^2-1)/2}]^{k+p}$ in the statement of Theorem 1 is bounded by e^{-s^2} for $s \geq 2$ and $k + p \geq 5$. In the rest of the section, we prove this theorem.

3.1 Three Caveats that Make the Generalization Non-Trivial

One might imagine that the generalization of the randomized SVD algorithm from matrices to HS operators is trivial, but this is not the case due to three caveats:

1. The randomized SVD on finite matrices always uses matrix-vector products with standard Gaussian random vectors [22]. However, for GPs, one must always have a continuous kernel K in $\mathcal{GP}(0, K)$, which discretizes to a non-standard multivariate Gaussian distribution. Therefore, we must extend [22, Theorem 10.5] to allow for non-standard multivariate Gaussian distributions. The discrete version of our extension is the following:

Corollary 1 *Let* $\mathbf{A}$ *be a real* $n_2 \times n_1$ *matrix with singular values* $\sigma_1 \geq \cdots \geq \sigma_{\min\{n_1,n_2\}}$. *Choose a target rank* $k \geq 1$ *and an oversampling parameter* $p \geq 2$. *Draw an* $n_1 \times (k+p)$ *Gaussian matrix,* $\boldsymbol{\Omega}$, *with independent columns where each column is from a multivariate Gaussian distribution with mean* $(0, \ldots, 0)^\top$ *and positive definite covariance matrix* $\mathbf{K}$. *If* $\mathbf{Y} = \mathbf{A}\boldsymbol{\Omega}$, *then the expected approximation error is bounded by*

$$\mathbb{E}\left[\|\mathbf{A} - \mathbf{P}_{\mathbf{Y}}\mathbf{A}\|_{\mathrm{F}}\right] \leq \left(1 + \sqrt{\frac{k+p}{p-1}\sum_{j=n_1-k+1}^{n_1}\frac{\lambda_1}{\lambda_j}}\right)\left(\sum_{j=k+1}^{\infty}\sigma_j^2\right)^{1/2}, \tag{8}$$

where $\lambda_1 \geq \cdots \geq \lambda_{n_1} > 0$ *are the eigenvalues of* $\mathbf{K}$ *and* $\mathbf{P}_{\mathbf{Y}}$ *is the orthogonal projection onto the vector space spanned by the columns of* $\mathbf{Y}$. *Assume further that*

$p \geq 4$, then for any $s, t \geq 1$, we have

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{Y}}\mathbf{A}\|_{\mathrm{F}} \leq \left(1 + ts \cdot \sqrt{\frac{3(k+p)}{p+1}\left(\sum_{j=1}^{n_1}\lambda_j\right)\sum_{j=n_1-k+1}^{n_1}\frac{1}{\lambda_j}}\right)\left(\sum_{j=k+1}^{\infty}\sigma_j^2\right)^{1/2},$$

with probability $\geq 1 - t^{-p} - [se^{-(s^2-1)/2}]^{k+p}$.

Choosing a covariance matrix $\mathbf{K}$ with sufficient eigenvalue decay so that $\lim_{n_1\to\infty}\sum_{j=1}^{n_1}\lambda_j < \infty$ allows $\mathbb{E}[\|\mathbf{\Omega}\|_{\mathrm{F}}^2]$ to remain bounded as $n_1 \to \infty$. This is of interest when applying the randomized SVD algorithm to extremely large matrices and is critical for HS operators. A stronger statement of this result [8, Theorem 2] shows that prior information on $\mathbf{A}$ can be incorporated into the covariance matrix to achieve lower approximation error than the randomized SVD with standard Gaussian vectors.

2. We need an additional essential assumption. The kernel in $\mathcal{GP}(0, K)$ is "reasonable" for learning $\mathscr{F}$, where reasonableness is measured by the quantity γ_k in Theorem 1. If the first k right singular functions of the HS operator $v_1, \ldots, v_k$ are spanned by the first $k+m$ eigenfunctions of K $\psi_1, \ldots, \psi_{k+m}$, for some $m \in \mathbb{N}$, then (see Eq. (11) and Lemma 2)

$$\frac{1}{k}\sum_{j=1}^{k}\frac{\lambda_1}{\lambda_j} \leq \frac{1}{\gamma_k} \leq \frac{1}{k}\sum_{j=m+1}^{k+m}\frac{\lambda_1}{\lambda_j}.$$

   In the matrix setting, this assumption always holds with $m = n_1 - k$ (see Corollary 1) and one can have $\gamma_k = 1$ when $\lambda_1 = \cdots = \lambda_{n_1}$ [22, Theorem 10.5].
3. Probabilistic error bounds for the randomized SVD in [22] are derived using tail bounds for functions of standard Gaussian matrices [30, Sect. 5.1]. Unfortunately, we are not aware of tail bounds for non-standard Gaussian quasimatrices. This results in a slightly weaker probability bound than [22, Theorem 10.7].

3.2 Deterministic Error Bound

Apart from the three caveats, the proof of Theorem 1 follows the outline of the argument in [22, Theorem 10.5]. We define two quasimatrices $\mathbf{U}$ and $\mathbf{V}$ containing the left and right singular functions of $\mathscr{F}$ so that the jth column of $\mathbf{V}$ is v_j. We also denote by $\mathbf{\Sigma}$ the infinite diagonal matrix with the singular values of $\mathscr{F}$, i.e., $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$, on the diagonal. Finally, for a fixed $k \geq 1$, we define the $D_1 \times k$ quasimatrix as the truncation of $\mathbf{V}$ after the first k columns and $\mathbf{V}_2$ as the remainder. Similarly, we split $\mathbf{\Sigma}$ into two parts:

$$\mathbf{\Sigma} = \begin{array}{c} \begin{array}{cc} k & \infty \end{array} \\ \begin{pmatrix} \mathbf{\Sigma}_1 & 0 \\ 0 & \mathbf{\Sigma}_2 \end{pmatrix} \end{array} \begin{array}{c} k \\ \infty \end{array}.$$

We are ready to prove an infinite-dimensional analogue of [22, Theorem 9.1] for HS operators.

Theorem 2 (Deterministic error bound) *Let $\mathscr{F} : L^2(D_1) \to L^2(D_2)$ be a HS operator with SVD given in Eq. (4). Let $\boldsymbol{\Omega}$ be a $D_1 \times \ell$ quasimatrix and $\mathbf{Y} = \mathscr{F}\boldsymbol{\Omega}$. If $\boldsymbol{\Omega}_1 = \mathbf{V}_1^*\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_2 = \mathbf{V}_2^*\boldsymbol{\Omega}$, then assuming $\boldsymbol{\Omega}_1$ has full rank, we have*

$$\|\mathscr{F} - \mathbf{P}_{\mathbf{Y}}\mathscr{F}\|_{\mathrm{HS}}^2 \le \|\boldsymbol{\Sigma}_2\|_{\mathrm{HS}}^2 + \|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{HS}}^2,$$

where $\mathbf{P}_{\mathbf{Y}} = \mathbf{Y}(\mathbf{Y}^*\mathbf{Y})^{\dagger}\mathbf{Y}^*$ is the orthogonal projection onto the space spanned by the columns of $\mathbf{Y}$ and $\boldsymbol{\Omega}_1^{\dagger} = (\boldsymbol{\Omega}_1^*\boldsymbol{\Omega}_1)^{-1}\boldsymbol{\Omega}_1^*$.

Proof First, note that because $\mathbf{U}\mathbf{U}^*$ is the orthonormal projection onto the range of $\mathscr{F}$ and $\mathbf{U}$ is a basis for the range, we have

$$\|\mathscr{F} - \mathbf{P}_{\mathbf{Y}}\mathscr{F}\|_{\mathrm{HS}} = \|\mathbf{U}\mathbf{U}^*\mathscr{F} - \mathbf{P}_{\mathbf{Y}}\mathbf{U}\mathbf{U}^*\mathscr{F}\|_{\mathrm{HS}}.$$

By Parseval's theorem [51, Theorem 4.18], we have

$$\|\mathbf{U}\mathbf{U}^*\mathscr{F} - \mathbf{P}_{\mathbf{Y}}\mathbf{U}\mathbf{U}^*\mathscr{F}\|_{\mathrm{HS}} = \|\mathbf{U}^*\mathbf{U}\mathbf{U}^*\mathscr{F} - \mathbf{U}^*\mathbf{P}_{\mathbf{Y}}\mathbf{U}\mathbf{U}^*\mathscr{F}\mathbf{V}\|_{\mathrm{HS}}.$$

Moreover, we have the equality $\|\mathscr{F} - \mathbf{P}_{\mathbf{Y}}\mathscr{F}\|_{\mathrm{HS}} = \|(\mathbf{I} - \mathbf{P}_{\mathbf{U}^*\mathbf{Y}})\mathbf{U}^*\mathscr{F}\mathbf{V}\|_{\mathrm{HS}}$ because the inner product $\langle \sum_{j=1}^{\infty} \alpha_j u_j, \sum_{j=1}^{\infty} \beta u_j \rangle = 0$ if and only if $\sum_{j=1}^{\infty} \alpha_j \beta_j = 0$. We now take $\mathbf{A} = \mathbf{U}^*\mathscr{F}\mathbf{V}$, which is a bounded infinite matrix such that $\|\mathbf{A}\|_{\mathrm{F}} = \|\mathscr{F}\|_{\mathrm{HS}} < \infty$. The statement of the theorem immediately follows from the proof of [22, Theorem 9.1]. □

This theorem shows that the bound on the approximation error $\|\mathscr{F} - \mathbf{P}_{\mathbf{Y}}\mathscr{F}\|_{\mathrm{HS}}$ depends on the singular values of the HS operator and the test matrix $\boldsymbol{\Omega}$.

3.3 Probability Distribution of $\boldsymbol{\Omega}_1$

If the columns of $\boldsymbol{\Omega}$ are independent and identically distributed as $\mathcal{GP}(0, K)$, then the matrix $\boldsymbol{\Omega}_1$ in Theorem 2 is of size $k \times \ell$ with entries that follow a Gaussian distribution. To see this, note that

$$\boldsymbol{\Omega}_1 = \mathbf{V}_1^*\boldsymbol{\Omega} = \begin{pmatrix} \langle v_1, \omega_1 \rangle & \cdots & \langle v_1, \omega_\ell \rangle \\ \vdots & \ddots & \vdots \\ \langle v_k, \omega_1 \rangle & \cdots & \langle v_k, \omega_\ell \rangle \end{pmatrix}, \qquad \omega_j \sim \mathcal{GP}(0, K).$$

If $\omega \sim \mathcal{GP}(0, K)$ with K given in Eq. (5), then we find that $\langle v, \omega \rangle \sim \mathcal{N}(0, \sum_{j=1}^{\infty} \lambda_j \langle v, \psi_j \rangle^2)$ so we conclude that $\boldsymbol{\Omega}_1$ has Gaussian entries with zero mean. Finding the covariances between the entries is more involved.

Lemma 1 *With the same setup as Theorem 2, suppose that the columns of $\boldsymbol{\Omega}$ are independent and identically distributed as $\mathcal{GP}(0, K)$. Then, the matrix $\boldsymbol{\Omega}_1 = \mathbf{V}_1^*\boldsymbol{\Omega}$ in*

Theorem 2 has independent columns and each column is identically distributed as a multivariate Gaussian with positive definite covariance matrix $\mathbf{C}$ *given by*

$$\mathbf{C}_{ij} = \int_{D_1 \times D_1} v_i(x) K(x, y) v_j(y) \mathrm{d}x \mathrm{d}y, \qquad 1 \le i, j \le k, \tag{9}$$

where v_i *is the* i*th column of* $\mathbf{V}_1$.

Proof We already know that the entries are Gaussian with mean 0. Moreover, the columns are independent because $\omega_1, \ldots, \omega_\ell$ are independent. Therefore, we focus on the covariance matrix. Let $1 \le i, i' \le k$, $1 \le j, j' \le \ell$, then since $\mathbb{E}\big[\langle v_i, \omega_j \rangle\big] = 0$ we have

$$\operatorname{cov}(\langle v_i, \omega_j \rangle, \langle v_{i'}, \omega_{j'} \rangle) = \mathbb{E}\left[\langle v_i, \omega_j \rangle \langle v_{i'}, \omega_{j'} \rangle\right] = \mathbb{E}\left[X_{ij} X_{i'j'}\right],$$

where $X_{ij} = \langle v_i, \omega_j \rangle$. Since $\langle v_i, \omega_j \rangle \sim \sum_{n=1}^{\infty} \sqrt{\lambda_n} c_n^{(j)} \langle v_i, \psi_n \rangle$, where $c_n^{(j)} \sim \mathcal{N}(0, 1)$, we have

$$\operatorname{cov}(\langle v_i, \omega_j \rangle, \langle v_{i'}, \omega_{j'} \rangle) = \mathbb{E}\left[\lim_{m_1, m_2 \to \infty} X_{ij}^{m_1} X_{i'j'}^{m_2}\right], \qquad X_{ij}^{m_1} := \sum_{n=1}^{m_1} \sqrt{\lambda_n} c_n^{(j)} \langle v_i, \psi_n \rangle.$$

We first show that $\lim_{m_1, m_2 \to \infty} \left| \mathbb{E}\left[X_{ij}^{m_1} X_{i'j'}^{m_2}\right] - \mathbb{E}\left[X_{ij} X_{i'j'}\right] \right| = 0$. For any $m_1, m_2 \ge 1$, we have by the triangle inequality,

$$\begin{aligned}
&\left| \mathbb{E}\left[X_{ij}^{m_1} X_{i'j'}^{m_2}\right] - \mathbb{E}\left[X_{ij} X_{i'j'}\right] \right| \\
&\quad \le \mathbb{E}\left[\left| X_{ij}^{m_1} X_{i'j'}^{m_2} - X_{ij} X_{i'j'} \right| \right] \\
&\quad \le \mathbb{E}\left[\left| (X_{ij}^{m_1} - X_{ij}) X_{i'j'}^{m_2} \right| \right] + \mathbb{E}\left[\left| X_{ij} (X_{i'j'}^{m_2} - X_{i'j'}) \right| \right] \\
&\quad \le \mathbb{E}\left[\left| X_{ij}^{m_1} - X_{ij} \right|^2 \right]^{\frac{1}{2}} \mathbb{E}\left[\left| X_{i'j'}^{m_2} \right|^2 \right]^{\frac{1}{2}} + \mathbb{E}\left[\left| X_{i'j'} - X_{i'j'}^{m_2} \right|^2 \right]^{\frac{1}{2}} \mathbb{E}\left[\left| X_{ij} \right|^2 \right]^{\frac{1}{2}},
\end{aligned}$$

where the last inequality follows from the Cauchy–Schwarz inequality. We now set out to show that both terms in the last inequality converge to zero as $m_1, m_2 \to \infty$. The terms $\mathbb{E}[|X_{i'j'}^{m_2}|^2]$ and $\mathbb{E}[|X_{ij}|^2]$ are bounded by $\sum_{n=1}^{\infty} \lambda_n < \infty$, using the Cauchy–Schwarz inequality. Moreover, we have

$$\mathbb{E}\left[\left| X_{ij}^{m_1} - X_{ij} \right|^2 \right] = \mathbb{E}\left[\left| \sum_{n=m_1+1}^{\infty} \sqrt{\lambda_n} c_n^{(j)} \langle v_i, \psi_n \rangle \right|^2 \right] \le \sum_{n=m_1+1}^{\infty} \lambda_n \xrightarrow[m_1 \to \infty]{} 0,$$

because $X_{ij} - X_{ij}^{m_1} \sim \mathcal{N}(0, \sum_{n=m_1+1}^{\infty} \lambda_n \langle v_i, \psi_n \rangle^2)$. Therefore, we find that $\operatorname{cov}(X_{ij}, X_{i'j'}) = \lim_{m_1,m_2\to\infty} \mathbb{E}[X_{ij}^{m_1} X_{i'j'}^{m_2}]$ and we obtain

$$\begin{aligned}\operatorname{cov}(X_{ij}, X_{i'j'}) &= \lim_{m_1,m_2\to\infty} \mathbb{E}\left[\sum_{n=1}^{m_1}\sum_{n'=1}^{m_2} \sqrt{\lambda_n\lambda_{n'}} c_n^{(j)} c_{n'}^{(j')} \langle v_i, \psi_n\rangle\langle v_{i'}, \psi_{n'}\rangle\right] \\ &= \lim_{m_1,m_2\to\infty} \sum_{n=1}^{m_1}\sum_{n'=1}^{m_2} \sqrt{\lambda_n\lambda_{n'}} \mathbb{E}[c_n^{(j)} c_{n'}^{(j')}] \langle v_i, \psi_n\rangle\langle v_{i'}, \psi_{n'}\rangle.\end{aligned}$$

The latter expression is zero if $n \neq n'$ or $j \neq j'$ because then $c_n^{(j)}$ and $c_{n'}^{(j')}$ are independent random variables with mean 0. Since $\mathbb{E}[(c_n^{(j)})^2] = 1$, we have

$$\operatorname{cov}(X_{ij}, X_{i'j'}) = \begin{cases} \sum_{n=1}^{\infty} \lambda_n \langle v_i, \psi_n\rangle\langle v_{i'}, \psi_n\rangle, & j = j', \\ 0, & \text{otherwise.}\end{cases}$$

The result follows as the infinite sum is equal to the integral in Eq. (9). To see that $\mathbf{C}$ is positive definite, let $a \in \mathbb{R}^k$, then $a^*\mathbf{C}a = \mathbb{E}[Z_a^2] \geq 0$, where $Z_a \sim \mathcal{N}(0, \sum_{n=1}^{\infty} \lambda_n \langle a_1 v_1 + \cdots + a_k v_k, \psi_n\rangle^2)$. Moreover, $a^*\mathbf{C}a = 0$ implies that $a = 0$ because $v_1, \ldots, v_k$ are orthonormal and $\{\psi_n\}$ is an orthonormal basis of $L^2(D_1)$. □

Lemma 1 gives the distribution of the matrix $\boldsymbol{\Omega}_1$, which is essential to prove Theorem 1 in Sect. 3.6. In particular, $\boldsymbol{\Omega}_1$ has independent columns that are each distributed as a multivariate Gaussian with covariance matrix given in Eq. (9).

3.4 Quality of the Covariance Kernel

To investigate the quality of the kernel, we introduce the Wishart distribution, which is a family of probability distributions over symmetric and nonnegative-definite matrices that often appear in the context of covariance matrices [61]. If $\boldsymbol{\Omega}_1$ is a $k \times \ell$ random matrix with independent columns, where each column is a multivariate Gaussian distribution with mean $(0, \ldots, 0)^\top$ and covariance $\mathbf{C}$, then $\mathbf{A} = \boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^*$ has a Wishart distribution [61]. We write $\mathbf{A} \sim W_k(\ell, \mathbf{C})$. We note that $\|\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^2 = \operatorname{Tr}[(\boldsymbol{\Omega}_1^\dagger)^*\boldsymbol{\Omega}_1^\dagger] = \operatorname{Tr}(\mathbf{A}^{-1})$, where the second equality holds with probability one because the matrix $\mathbf{A} = \boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^*$ is invertible with probability one (see [41, Theorem 3.1.4]). By [41, Theorem 3.2.12] for $\ell - k \geq 2$, we have $\mathbb{E}[\mathbf{A}^{-1}] = \frac{1}{\ell-k-1}\mathbf{C}^{-1}$, $\mathbb{E}[\operatorname{Tr}(\mathbf{A}^{-1})] = \operatorname{Tr}(\mathbf{C}^{-1})/(\ell - k - 1)$, and conclude that

$$\mathbb{E}\left[\|\boldsymbol{\Omega}_1^\dagger\|_{\mathrm{F}}^2\right] = \frac{1}{\gamma_k\lambda_1}\frac{k}{\ell - k - 1}, \qquad \gamma_k := \frac{k}{\lambda_1 \operatorname{Tr}(\mathbf{C}^{-1})}. \tag{10}$$

The quantity γ_k can be viewed as measuring the quality of the covariance kernel K for learning the HS operator $\mathscr{F}$ (see Theorem 1). First, $1 \leq \gamma_k < \infty$ as $\mathbf{C}$ is symmetric positive definite. Moreover, for $1 \leq j \leq k$, the jth largest eigenvalue of $\mathbf{C}$

is bounded by the jth largest eigenvalue of K as $\mathbf{C}$ is a principal submatrix of $\mathbf{V}^* K \mathbf{V}$ [27, Sect. III.5]. Therefore, the following inequality holds,

$$\frac{1}{k} \sum_{j=1}^{k} \frac{\lambda_1}{\lambda_j} \leq \frac{1}{\gamma_k} < \infty, \tag{11}$$

and the harmonic mean of the first k scaled eigenvalues of K is a lower bound for $1/\gamma_k$. In the ideal situation, the eigenfunctions of K are the right singular functions of $\mathscr{F}$, i.e., $\psi_n = v_n$, $\mathbf{C}$ is a diagonal matrix with entries $\lambda_1, \ldots, \lambda_k$, and $\gamma_k = k/(\sum_{j=1}^{k} \lambda_1/\lambda_j)$ is as small as possible.

We now provide a useful upper bound on γ_k in a more general setting.

Lemma 2 *Let $\mathbf{V}_1$ be a $D_1 \times k$ quasimatrix with orthonormal columns and assume that there exists $m \in \mathbb{N}$ such that the columns of $\mathbf{V}_1$ are spanned by the first $k+m$ eigenvectors of the continuous positive definite kernel $K : D_1 \times D_1 \to \mathbb{R}$. Then*

$$\frac{1}{\gamma_k} \leq \frac{1}{k} \sum_{j=m+1}^{k+m} \frac{\lambda_1}{\lambda_j},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ are the eigenvalues of K. This bound is tight in the sense that the inequality can be attained as an equality.

Proof Let $\mathbf{Q} = \left[v_1 \,|\, \cdots \,|\, v_k \,|\, q_{k+1} \,|\, \cdots \,|\, q_{k+m} \right]$ be a quasimatrix with orthonormal columns whose columns form an orthonormal basis for $\mathrm{Span}(\psi_1, \ldots, \psi_{k+m})$. Then, $\mathbf{Q}$ is an invariant space of K and $\mathbf{C}$ is a principal submatrix of $\mathbf{Q}^* K \mathbf{Q}$, which has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{k+m}$. By [27, Theorem 6.46] the k eigenvalues of $\mathbf{C}$, denoted by $\mu_1, \ldots, \mu_k$, are greater than the first $k+m$ eigenvalues of K: $\mu_j \geq \lambda_{m+j}$ for $1 \leq j \leq k$, and the result follows as the trace of a matrix is the sum of its eigenvalues. □

3.5 Probabilistic Error Bounds

As discussed in Sect. 3.1, we need to extend the probability bounds of the randomized SVD to allow for non-standard Gaussian random vectors. The following lemma is a generalization of [22, Theorem A.7].

Lemma 3 *Let $k, \ell \geq 1$ such that $\ell - k \geq 4$ and $\mathbf{\Omega}_1$ be a $k \times \ell$ random matrix with independent columns such that each column has mean $(0, \ldots, 0)^\top$ and positive definite covariance $\mathbf{C}$. For all $t \geq 1$, we have*

$$\mathbb{P}\left\{ \|\mathbf{\Omega}_1^\dagger\|_{\mathrm{F}}^2 > \frac{3 \operatorname{Tr}(\mathbf{C}^{-1})}{\ell - k + 1} \cdot t^2 \right\} \leq t^{-(\ell - k)}.$$

Proof Since $\mathbf{\Omega}_1\mathbf{\Omega}_1^* \sim W_k(\ell, \mathbf{C})$, the reciprocals of its diagonal elements follow a scaled chi-square distribution [41, Theorem 3.2.12], i.e.,

$$\frac{\left((\mathbf{\Omega}_1\mathbf{\Omega}_1^*)^{-1}\right)_{jj}}{\left(\mathbf{C}^{-1}\right)_{jj}} \sim X_j^{-1}, \qquad X_j \sim \chi^2_{\ell-k+1}, \qquad 1 \le j \le k.$$

Let $Z = \|\mathbf{\Omega}_1^\dagger\|_{\mathrm{F}}^2 = \mathrm{Tr}[(\mathbf{\Omega}_1\mathbf{\Omega}_1^*)^{-1}]$ and $q = (\ell - k)/2$. Following the proof of [22, Theorem A.7], we have the inequality

$$\mathbb{P}\left\{|Z| \ge \frac{3\,\mathrm{Tr}(\mathbf{C}^{-1})}{\ell - k + 1} \cdot t^2\right\} \le \left[\frac{3\,\mathrm{Tr}(\mathbf{C}^{-1})}{\ell - k + 1} \cdot t^2\right]^{-q} \mathbb{E}\left[|Z|^q\right], \quad t \ge 1.$$

Moreover, by the Minkowski inequality, we have

$$\left(\mathbb{E}\left[|Z^q|\right]\right)^{1/q} = \left(\mathbb{E}\left[\left|\sum_{j=1}^k [\mathbf{C}^{-1}]_{jj} X_j^{-1}\right|^q\right]\right)^{1/q} \le \sum_{j=1}^k [\mathbf{C}^{-1}]_{jj} \mathbb{E}\left[|X_j^{-1}|^q\right]^{1/q} \le \frac{3\,\mathrm{Tr}(\mathbf{C}^{-1})}{\ell - k + 1},$$

where the last inequality is from [22, Lemma A.9]. The result follows from the argument in the proof of [22, Theorem A.7]. □

Under the assumption of Lemma 2, we find that Lemma 3 gives the following bound:

$$\mathbb{P}\left\{\|\mathbf{\Omega}_1^\dagger\|_{\mathrm{F}} > t \cdot \sqrt{\frac{3}{\ell - k + 1} \sum_{j=m+1}^{k+m} \lambda_j^{-1}}\right\} \le t^{-(\ell-k)}.$$

In particular, in the finite-dimensional case when $\lambda_1 = \cdots = \lambda_n = 1$, we recover the probabilistic bound found in [22, Theorem A.7].

To obtain the probability statement found in Eq. (13), we require control of the tail of the distribution of a Gaussian quasimatrix with non-standard covariance kernel (see Sect. 3.6). In the theory of the randomized SVD, one relies on the concentration of measure results [22, Prop. 10.3]. However, we need to employ a different strategy and instead directly bound the HS norm of $\mathbf{\Omega}_2$. One difficulty is that the norm of this matrix must be controlled for large dimensions n, which leads to a weaker probability bound than [22]. While it is possible to apply Markov's inequality to obtain deviation bounds, we highlight that Lemma 4 provides a Chernoff-type bound, i.e., exponential decay of the tail distribution of $\|\mathbf{\Omega}_2\|_{\mathrm{HS}}$, which is crucial to approximate Green's functions (see Sect. 4.4.3).

Lemma 4 *With the same notation as in Theorem 2, let $\ell \ge k \ge 1$. For all $s \ge 1$, we have*

$$\mathbb{P}\left\{\|\mathbf{\Omega}_2\|_{\mathrm{HS}}^2 > \ell s^2\,\mathrm{Tr}(K)\right\} \le \left[s e^{-(s^2-1)/2}\right]^\ell.$$

Proof We first remark that

$$\|\mathbf{\Omega}_2\|_{\mathrm{HS}}^2 \leq \|\mathbf{\Omega}\|_{\mathrm{HS}}^2 = \sum_{j=1}^{\ell} Z_j, \qquad Z_j := \|\omega_j\|_{L^2(D_1)}^2, \tag{12}$$

where the Z_j are independent and identically distributed (i.i.d) because $\omega_j \sim \mathcal{GP}(0, K)$ are i.i.d. For $1 \leq j \leq \ell$, we have (c.f. Sect. 2.3),

$$\omega_j = \sum_{m=1}^{\infty} c_m^{(j)} \sqrt{\lambda_m} \psi_m,$$

where $c_m^{(j)} \sim \mathcal{N}(0, 1)$ are i.i.d for $m \geq 1$ and $1 \leq j \leq \ell$. First, since the series in Eq. (12) converges absolutely, we have

$$Z_j = \sum_{m=1}^{\infty} (c_m^{(j)})^2 \lambda_m = \lim_{N \to \infty} \sum_{m=1}^{N} X_m, \qquad X_m = (c_m^{(j)})^2 \lambda_m,$$

where the X_m are independent random variables and $X_m \sim \lambda_m \chi^2$ for $1 \leq m \leq N$. Here, χ^2 denotes the chi-squared distribution [40, Chapt. 4.3].

Let $N \geq 1$ and $0 < \theta < 1/(2\,\mathrm{Tr}(K))$, we can bound the moment generating function of $\sum_{m=1}^{N} X_m$ as

$$\begin{aligned}\mathbb{E}\left[e^{\theta \sum_{m=1}^{N} X_m}\right] = \prod_{m=1}^{N} \mathbb{E}\left[e^{\theta X_m}\right] = \prod_{m=1}^{N} (1 - 2\theta\lambda_m)^{-1/2} &\leq \left(1 - 2\theta \sum_{m=1}^{N} \lambda_m\right)^{-1/2} \\ &\leq (1 - 2\theta\,\mathrm{Tr}(K))^{-1/2},\end{aligned}$$

because X_m/λ_m are independent random variables that follow a chi-squared distribution. Using the monotone convergence theorem, we have

$$\mathbb{E}\left[e^{\theta Z_j}\right] \leq (1 - 2\theta\,\mathrm{Tr}(K))^{-1/2}.$$

Let $\tilde{s} \geq 0$ and $0 < \theta < 1/(2\,\mathrm{Tr}(K))$, by the Chernoff bound [10, Theorem 1], we obtain

$$\begin{aligned}\mathbb{P}\left\{\|\mathbf{\Omega}_2\|_{\mathrm{HS}}^2 > \ell(1 + \tilde{s})\,\mathrm{Tr}(K)\right\} &\leq e^{-(1+\tilde{s})\,\mathrm{Tr}(K)\ell\theta}\,\mathbb{E}\left[e^{\theta Z_j}\right]^{\ell} \\ &= e^{-(1+\tilde{s})\,\mathrm{Tr}(K)\ell\theta}(1 - 2\theta\,\mathrm{Tr}(K))^{-\ell/2}.\end{aligned}$$

We can minimize this upper bound over $0 < \theta < 1/(2\,\mathrm{Tr}(K))$ by choosing $\theta = \tilde{s}/(2(1 + \tilde{s})\,\mathrm{Tr}(K))$, which gives

$$\mathbb{P}\left\{\|\mathbf{\Omega}_2\|_{\mathrm{HS}}^2 > \ell(1 + \tilde{s})\,\mathrm{Tr}(K)\right\} \leq (1 + \tilde{s})^{\ell/2} e^{-\ell\tilde{s}/2}.$$

Choosing $s = \sqrt{1+\tilde{s}} \geq 1$ concludes the proof. □

Lemma 4 can be refined further to take into account the interaction between the Hilbert–Schmidt operator $\mathscr{F}$ and the covariance kernel K (see [8, Lemma 7]).

3.6 Randomized SVD Algorithm for HS Operators

We first prove an intermediary result, which generalizes [22, Prop. 10.1] to HS operators. Note that one may obtain sharper bounds using a suitably chosen covariance kernels that yields a lower approximation error [8].

Lemma 5 *Let $\mathbf{\Sigma}_2$, $\mathbf{V}_2$, and $\mathbf{\Omega}$ be defined as in Theorem 2, and $\mathbf{T}$ be an $\ell \times k$ matrix, where $\ell \geq k \geq 1$. Then,*

$$\mathbb{E}\left[\|\mathbf{\Sigma}_2\mathbf{V}_2^*\mathbf{\Omega}\mathbf{T}\|_{\mathrm{HS}}^2\right] \leq \lambda_1 \|\mathbf{\Sigma}_2\|_{\mathrm{HS}}^2 \|\mathbf{T}\|_{\mathrm{F}}^2,$$

where λ_1 is the first eigenvalue of K.

Proof Let $\mathbf{T} = \mathbf{U}_\mathbf{T}\mathbf{D}_\mathbf{T}\mathbf{V}_\mathbf{T}^*$ be the SVD of $\mathbf{T}$. If $\{v_{\mathbf{T},i}\}_{i=1}^k$ are the columns of $\mathbf{V}_\mathbf{T}$, then

$$\mathbb{E}\left[\|\mathbf{\Sigma}_2\mathbf{V}_2^*\mathbf{\Omega}\mathbf{T}\|_{\mathrm{HS}}^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[\|\mathbf{\Sigma}_2\mathbf{\Omega}_2\mathbf{U}_\mathbf{T}\mathbf{D}_\mathbf{T}\mathbf{V}_\mathbf{T}^* v_{\mathbf{T},i}\|_2^2\right],$$

where $\mathbf{\Omega}_2 = \mathbf{V}_2^*\mathbf{\Omega}$. Therefore, we have

$$\mathbb{E}\left[\|\mathbf{\Sigma}_2\mathbf{\Omega}_2\mathbf{T}\|_{\mathrm{HS}}^2\right] = \sum_{i=1}^{k} ((\mathbf{D}_\mathbf{T})_{ii})^2 \mathbb{E}\left[\|\mathbf{\Sigma}_2\mathbf{\Omega}_2\mathbf{U}_\mathbf{T}(:,i)\|_2^2\right].$$

Moreover, using the monotone convergence theorem for non-negative random variables, we have

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{\Sigma}_2\mathbf{\Omega}_2\mathbf{U}_\mathbf{T}(:,i)\|_2^2\right] &= \mathbb{E}\left[\sum_{n=1}^{\infty}\sum_{j=1}^{\ell} \sigma_{k+n}^2 |\mathbf{\Omega}_2(n,j)|^2 \mathbf{U}_\mathbf{T}(j,i)^2\right] \\ &= \sum_{n=1}^{\infty}\sum_{j=1}^{\ell} \sigma_{k+n}^2 \mathbf{U}_\mathbf{T}(j,i)^2 \mathbb{E}\left[|\mathbf{\Omega}_2(n,j)|^2\right],\end{aligned}$$

where $\sigma_{k+1}, \sigma_{k+2}, \ldots$ are the diagonal elements of $\mathbf{\Sigma}_2$. Then, the quasimatrix $\mathbf{\Omega}_2$ has independent columns and, using Lemma 1, we have

$$\mathbb{E}\left[|\mathbf{\Omega}_2(n,j)|^2\right] = \int_{D_1\times D_1} v_{k+n}(x) K(x,y) v_{k+n}(y) \mathrm{d}x\mathrm{d}y,$$

where v_{k+n} is the nth column of $\mathbf{V}_2$. Then, $\mathbb{E}\left[|\boldsymbol{\Omega}_2(n,j)|^2\right] \leq \lambda_1$, as $\mathbb{E}\left[|\boldsymbol{\Omega}_2(n,j)|^2\right]$ is written as a Rayleigh quotient. Finally, we have

$$\mathbb{E}\left[\|\boldsymbol{\Sigma}_2\mathbf{V}_2^*\boldsymbol{\Omega}\mathbf{T}\|_{\mathrm{HS}}^2\right] \leq \lambda_1 \sum_{i=1}^{k}((\mathbf{D_T})_{ii})^2 \sum_{j=1}^{\ell}\mathbf{U_T}(j,i)^2 \sum_{n=1}^{\infty}\sigma_{k+n}^2 = \lambda_1\|\mathbf{T}\|_{\mathrm{F}}^2\|\boldsymbol{\Sigma}_2\|_{\mathrm{HS}}^2,$$

by orthonormality of the columns on $\mathbf{U_T}$. □

We are now ready to prove Theorem 1, which shows that the randomized SVD can be generalized to HS operators.

Proof of Theorem 1 Let $\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2$ be the quasimatrices defined in Theorem 2. The $k \times (k+p)$ matrix $\boldsymbol{\Omega}_1$ has full rank with probability one and by Theorem 2, we have

$$\begin{aligned}\mathbb{E}\left[\|(\mathbf{I}-\mathbf{P_Y})\mathscr{F}\|_{\mathrm{HS}}\right] &\leq \mathbb{E}\left[\left(\|\boldsymbol{\Sigma}_2\|_{\mathrm{HS}}^2+\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{HS}}^2\right)^{1/2}\right] \leq \|\boldsymbol{\Sigma}_2\|_{\mathrm{HS}} + \mathbb{E}\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{HS}} \\ &\leq \|\boldsymbol{\Sigma}_2\|_{\mathrm{HS}} + \mathbb{E}\left[\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\|_{\mathrm{HS}}^2\right]^{1/2}\mathbb{E}\left[\|\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{F}}^2\right]^{1/2},\end{aligned}$$

where the last inequality follows from Cauchy–Schwarz inequality. Then, using Lemma 5 and Eq. (10), we have

$$\mathbb{E}\left[\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\|_{\mathrm{HS}}^2\right] \leq \lambda_1(k+p)\|\boldsymbol{\Sigma}_2\|_{\mathrm{HS}}^2, \qquad \text{and} \qquad \mathbb{E}\left[\|\boldsymbol{\Omega}_1\|_{\mathrm{F}}^2\right] \leq \frac{1}{\gamma_k\lambda_1}\frac{k}{p-1}.$$

where γ_k is defined in Sect. 3.4. The observation that $\|\boldsymbol{\Sigma}_2\|_{\mathrm{HS}}^2 = \sum_{j=k+1}^{\infty}\sigma_j^2$ concludes the proof of Eq. (6).

For the probabilistic bound in Eq. (7), we note that by Theorem 2 we have,

$$\|\mathscr{F} - \mathbf{P_Y}\mathscr{F}\|_{\mathrm{HS}}^2 \leq \|\boldsymbol{\Sigma}_2\|_{\mathrm{HS}}^2 + \|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{HS}}^2 \leq (1+\|\boldsymbol{\Omega}_2\|_{\mathrm{HS}}^2\|\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{F}}^2)\|\boldsymbol{\Sigma}_2\|_{\mathrm{HS}}^2,$$

where the second inequality uses the submultiplicativity of the HS norm. The bound follows from bounding $\|\boldsymbol{\Omega}_1^{\dagger}\|_{\mathrm{F}}^2$ and $\|\boldsymbol{\Omega}_2\|_{\mathrm{HS}}^2$ using Lemma 3 and 4, respectively. □

4 Recovering the Green's Function from Input–Output Pairs

It is known that the Green's function associated with Eq. (2) always exists, is unique, is a nonnegative function $G : D \times D \to \mathbb{R}^+ \cup \{\infty\}$ such that

$$u(x) = \int_D G(x,y)f(y)\mathrm{d}y, \qquad f \in \mathcal{C}_c^{\infty}(D),$$

and for each $y \in \Omega$ and any $r > 0$, we have $G(\cdot, y) \in \mathcal{H}^1(D \setminus B_r(y)) \cap \mathcal{W}_0^{1,1}(D)$ [19].[3] Since the PDE in Eq. (2) is self-adjoint, we also know that for almost every $x, y \in D$, we have $G(x, y) = G(y, x)$ [19].

We now state Theorem 3, which shows that if $N = \mathcal{O}(\epsilon^{-6} \log^4(1/\epsilon))$ and one has N input–output pairs $\{(f_j, u_j)\}_{j=1}^N$ with algorithmically selected f_j, then the Green's function associated with $\mathcal{L}$ in Eq. (2) can be recovered to within an accuracy of $\mathcal{O}(\Gamma_\epsilon^{-1/2} \log^3(1/\epsilon)\epsilon)$ with high probability. Here, the quantity $0 < \Gamma_\epsilon \leq 1$ measures the quality of the random input functions $\{f_j\}_{j=1}^N$ (see Sect. 4.4.2).

Theorem 3 *Let $0 < \epsilon < 1$, $D \subset \mathbb{R}^3$ be a bounded Lipschitz domain, and $\mathcal{L}$ given in Eq. (2). If G is the Green's function associated with $\mathcal{L}$, then there is a randomized algorithm that constructs an approximation $\tilde{G}$ of G using $\mathcal{O}(\epsilon^{-6} \log^4(1/\epsilon))$ input–output pairs such that, as $\epsilon \to 0$, we have*

$$\|G - \tilde{G}\|_{L^2(D\times D)} = \mathcal{O}\left(\Gamma_\epsilon^{-1/2} \log^3(1/\epsilon)\epsilon\right) \|G\|_{L^2(D\times D)}, \tag{13}$$

with probability $\geq 1 - \mathcal{O}(\epsilon^{\log(1/\epsilon)-6})$. The term Γ_ϵ is defined by Eq. (25).

Our algorithm that leads to the proof of Theorem 3 relies on the extension of the randomized SVD to HS operator (see Sect. 3) and a hierarchical partition of the domain of G into "well-separated" domains.

4.1 Recovering the Green's Function on Admissible Domains

Roughly speaking, as $\|x - y\|_2$ increases G becomes smoother about (x, y), which can be made precise using so-called admissible domains [1,2,21]. Let $\operatorname{diam} X := \sup_{x,y\in X} \|x - y\|_2$ be the diameter of X, $\operatorname{dist}(X, Y) := \inf_{x\in X, y\in Y} \|x - y\|_2$ be the shortest distance between X and Y, and $\rho > 0$ be a fixed constant. If $X, Y \subset \mathbb{R}^3$ are bounded domains, then we say that $X \times Y$ is an admissible domain if $\operatorname{dist}(X, Y) \geq \rho \max\{\operatorname{diam} X, \operatorname{diam} Y\}$; otherwise, we say that $X \times Y$ is non-admissible. There is a weaker definition of admissible domains as $\operatorname{dist}(X, Y) \geq \rho \min\{\operatorname{diam} X, \operatorname{diam} Y\}$ [21, p. 59], but we do not consider it.

4.1.1 Approximation Theory on Admissible Domains

It turns out that the Green's function associated with Eq. (2) has rapidly decaying singular values when restricted to admissible domains. Roughly speaking, if $X, Y \subset D$ are such that $X \times Y$ is an admissible domain, then G is well-approximated by a function of the form [3]

$$G_k(x, y) = \sum_{j=1}^{k} g_j(x) h_j(y), \qquad (x, y) \in X \times Y, \tag{14}$$

[3] Here, $B_r(y) = \{z \in \mathbb{R}^3 : \|z - y\|_2 < r\}$, $\mathcal{W}^{1,1}(D)$ is the space of weakly differentiable functions in the L^1-sense, and $\mathcal{W}_0^{1,1}(D)$ is the closure of $\mathcal{C}_c^\infty(D)$ in $\mathcal{W}^{1,1}(D)$.

for some functions $g_1, \ldots, g_k \in L^2(X)$ and $h_1, \ldots, h_k \in L^2(Y)$. This is summarized in Theorem 4, which is a corollary of [3, Theorem 2.8].

Theorem 4 *Let G be the Green's function associated with Eq. (2) and $\rho > 0$. Let $X, Y \subset D$ such that $\operatorname{dist}(X, Y) \geq \rho \max\{\operatorname{diam} X, \operatorname{diam} Y\}$. Then, for any $0 < \epsilon < 1$, there exists $k \leq k_\epsilon := \lceil c(\rho, \operatorname{diam} D, \kappa_C)\rceil \lceil \log(1/\epsilon)\rceil^4 + \lceil \log(1/\epsilon)\rceil$ and an approximant, G_k, of G in the form given in Eq. (14) such that*

$$\|G - G_k\|_{L^2(X\times Y)} \leq \epsilon \|G\|_{L^2(X\times \hat{Y})}, \qquad \hat{Y} := \{y \in D,\ \operatorname{dist}(y, Y) \leq \frac{\rho}{2} \operatorname{diam} Y\},$$

where $\kappa_C = \lambda_{\max}/\lambda_{\min}$ is the spectral condition number of the coefficient matrix $A(x)$ in Eq. (2)[4] *and c is a constant that only depends on ρ, $\operatorname{diam} D$, κ_C.*

Proof In [3, Theorem 2.8], it is shown that if $Y = \tilde{Y} \cap D$ and $\tilde{Y}$ is convex, then there exists $k \leq c_{\rho/2}^3 \lceil \log(1/\epsilon)\rceil^4 + \lceil \log(1/\epsilon)\rceil$ and an approximant, G_k, of G such that

$$\|G(x, \cdot) - G_k(x, \cdot)\|_{L^2(Y)} \leq \epsilon \|G(x, \cdot)\|_{L^2(\hat{Y})}, \qquad x \in X, \tag{15}$$

where $\hat{Y} := \{y \in D,\ \operatorname{dist}(y, Y) \leq \frac{\rho}{2} \operatorname{diam} Y\}$ and $c_{\rho/2}$ is a constant that only depends on ρ, $\operatorname{diam} Y$, and κ_C. As remarked by [3], $\tilde{Y}$ can be included in a convex of diameter $\operatorname{diam} D$ that includes D to obtain the constant $c(\rho, \operatorname{diam} D, \kappa_C)$. The statement follows by integrating the error bound in Eq. (15) over X. □

Since the truncated SVD of G on $X \times Y$ gives the best rank $k_\epsilon \geq k$ approximation to G, Theorem 4 also gives bounds on singular values:

$$\left(\sum_{j=k_\epsilon+1}^{\infty} \sigma_{j,X\times Y}^2\right)^{1/2} \leq \|G - G_k\|_{L^2(X\times Y)} \leq \epsilon \|G\|_{L^2(X\times \hat{Y})}, \tag{16}$$

where $\sigma_{j,X\times Y}$ is the jth singular value of G restricted to $X \times Y$. Since $k_\epsilon = \mathcal{O}(\log^4(1/\epsilon))$, we conclude that the singular values of G restricted to admissible domains $X \times Y$ rapidly decay to zero.

4.1.2 Randomized SVD for Admissible Domains

Since G has rapidly decaying singular values on admissible domains $X \times Y$, we use the randomized SVD for HS operators to learn G on $X \times Y$ with high probability (see Sect. 3).

We start by defining a GP on the domain Y. Let $\mathcal{R}_{Y\times Y} K$ be the restriction[5] of the covariance kernel K to the domain $Y \times Y$, which is a continuous symmetric positive definite kernel so that $\mathcal{GP}(0, \mathcal{R}_{Y\times Y} K)$ defines a GP on Y. We choose a target rank $k \geq 1$, an oversampling parameter $p \geq 2$, and form a quasimatrix $\boldsymbol{\Omega} = \left[f_1 \mid \cdots \mid f_{k+p}\right]$ such that $f_j \in L^2(Y)$ and $f_j \sim \mathcal{GP}(0, \mathcal{R}_{Y\times Y} K)$ are identically distributed and

[4] Here, $\lambda_{\max}$ is defined as $\sup_{x\in D} \lambda_{\max}(A(x))$ and $\lambda_{\min} = \inf_{x\in D} \lambda_{\min}(A(x)) > 0$.

[5] We denote the restriction operator by $\mathcal{R}_{Y\times Y} : L^2(D \times D) \to L^2(Y \times Y)$.

independent. We then extend by zero each column of $\mathbf{\Omega}$ from $L^2(Y)$ to $L^2(D)$ by $\mathcal{R}_Y^*\mathbf{\Omega} = \left[\mathcal{R}_Y^* f_1 \mid \cdots \mid \mathcal{R}_Y^* f_{k+p}\right]$, where $\mathcal{R}_Y^* f_j \sim \mathcal{GP}(0, \mathcal{R}_{Y\times Y}^* \mathcal{R}_{Y\times Y} K)$. The zero extension operator $\mathcal{R}_Y^* : L^2(Y) \to L^2(D)$ is the adjoint of $\mathcal{R}_Y : L^2(D) \to L^2(Y)$.

Given the training data, $\mathbf{Y} = \left[u_1 \mid \cdots \mid u_{k+p}\right]$ such that $\mathcal{L}u_j = \mathcal{R}_Y^* f_j$ and $u_j|_{\partial D} = 0$, we now construct an approximation to G on $X \times Y$ using the randomized SVD (see Sect. 3). Following Theorem 1, we have the following approximation error for $t \geq 1$ and $s \geq 2$:

$$\|G - \tilde{G}_{X\times Y}\|_{L^2(X\times Y)}^2 \leq \left(1 + t^2 s^2 \frac{3}{\gamma_{k,X\times Y}} \frac{k(k+p)}{p+1} \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_1}\right) \left(\sum\nolimits_{j=k+1}^{\infty} \sigma_{j,X\times Y}^2\right)^{1/2}, \tag{17}$$

with probability greater than $1 - t^{-p} - e^{-s^2(k+p)}$. Here, $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ are the eigenvalues of K, $\tilde{G}_{X\times Y} = \mathbf{P}_{\mathcal{R}_X\mathbf{Y}} \mathcal{R}_X \mathscr{F} \mathcal{R}_Y^*$ and $\mathbf{P}_{\mathcal{R}_X\mathbf{Y}} = \mathcal{R}_X\mathbf{Y}((\mathcal{R}_X\mathbf{Y})^* \mathcal{R}_X\mathbf{Y})^\dagger (\mathcal{R}_X\mathbf{Y})^*$ is the orthogonal projection onto the space spanned by the columns of $\mathcal{R}_X\mathbf{Y}$. Moreover, $\gamma_{k,X\times Y}$ is a measure of the quality of the covariance kernel of $\mathcal{GP}(0, \mathcal{R}_{Y\times Y}^* \mathcal{R}_{Y\times Y} K)$ (see Sect. 3.4) and, for $1 \leq i, j \leq k$, defined as $\gamma_{k,X\times Y} = k/(\lambda_1 \operatorname{Tr}(\mathbf{C}_{X\times Y}^{-1}))$, where

$$[\mathbf{C}_{X\times Y}]_{ij} = \int_{D\times D} \mathcal{R}_Y^* v_{i,X\times Y}(x) K(x,y) \mathcal{R}_Y^* v_{j,X\times Y}(y) \mathrm{d}x \mathrm{d}y,$$

and $v_{1,X\times Y}, \ldots, v_{k,X\times Y} \in L^2(Y)$ are the first k right singular functions of G restricted to $X \times Y$.

Unfortunately, there is a big problem with the formula $\tilde{G}_{X\times Y} = \mathbf{P}_{\mathcal{R}_X\mathbf{Y}} \mathcal{R}_X \mathscr{F} \mathcal{R}_Y^*$. It cannot be formed because we only have access to input–output data, so we have no mechanism for composing $\mathbf{P}_{\mathcal{R}_X\mathbf{Y}}$ on the left of $\mathcal{R}_X \mathscr{F} \mathcal{R}_Y^*$. Instead, we note that since the partial differential operator in Eq. (2) is self-adjoint, $\mathscr{F}$ is self-adjoint, and G is itself symmetric. That means we can use this to write down a formula for $\tilde{G}_{Y\times X}$ instead. That is,

$$\tilde{G}_{Y\times X} = \tilde{G}_{X\times Y}^* = \mathcal{R}_Y \mathscr{F} \mathcal{R}_X^* \mathbf{P}_{\mathcal{R}_X\mathbf{Y}},$$

where we used the fact that $\mathbf{P}_{\mathcal{R}_X\mathbf{Y}}$ is also self-adjoint. This means we can construct $\tilde{G}_{Y\times X}$ by asking for more input–output data to assess the quasimatrix $\mathscr{F}(\mathcal{R}_X^* \mathcal{R}_X\mathbf{Y})$. Of course, to compute $\tilde{G}_{X\times Y}$, we can swap the roles of X and Y in the above argument.

With a target rank of $k = k_\epsilon = \lceil c(\rho, \operatorname{diam} D, \kappa_C)\rceil \lceil \log(1/\epsilon)\rceil^4 + \lceil \log(1/\epsilon)\rceil$ and an oversampling parameter of $p = k_\epsilon$, we can combine Theorem 4 and Eq. (16) and (17) to obtain the bound

$$\|G - \tilde{G}_{X\times Y}\|_{L^2(X\times Y)}^2 \leq \left(1 + t^2 s^2 \frac{6k_\epsilon}{\gamma_{k_\epsilon,X\times Y}} \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_1}\right) \epsilon^2 \|G\|_{L^2(X\times \hat{Y})}^2,$$

with probability greater than $1 - t^{-k_\epsilon} - e^{-2s^2 k_\epsilon}$. A similar approximation error holds for $\tilde{G}_{Y\times X}$ without additional evaluations of $\mathscr{F}$. We conclude that our algorithm requires

$N_{\epsilon,X\times Y} = 2(k_\epsilon + p) = \mathcal{O}\big(\log^4(1/\epsilon)\big)$ input–output pairs to learn an approximant to G on $X \times Y$ and $Y \times X$.

4.2 Ignoring the Green's Function on Non-Admissible Domains

When the Green's function is restricted to non-admissible domains, its singular values may not decay. Instead, to learn G we take advantage of the off-diagonal decay property of G. It is known that for almost every $x \neq y \in D$ then

$$G(x, y) \leq \frac{c_{\kappa_C}}{\|x - y\|_2} \|G\|_{L^2(D\times D)}, \tag{18}$$

where c_{κ_C} is an implicit constant that only depends on κ_C (see [19, Theorem 1.1]).[6]

If $X \times Y$ is a non-admissible domain, then for any $(x, y) \in X \times Y$, we find that

$$\|x - y\|_2 \leq \operatorname{dist}(X, Y) + \operatorname{diam}(X) + \operatorname{diam}(Y) < (2 + \rho) \max\{\operatorname{diam} X, \operatorname{diam} Y\},$$

because $\operatorname{dist}(X, Y) < \rho \max\{\operatorname{diam} X, \operatorname{diam} Y\}$. This means that $x \in B_r(y) \cap D$, where $r = (2 + \rho) \max\{\operatorname{diam} X, \operatorname{diam} Y\}$. Using Eq. (18), we have

$$\begin{aligned}\int_X G(x, y)^2 dx &\leq \int_{B_r(y)\cap D} G(x, y)^2 \mathrm{d}x \leq c_{\kappa_C}^2 \|G\|_{L^2(D\times D)}^2 \int_{B_r(y)} \|x - y\|_2^{-2} \mathrm{d}x \\ &\leq 4\pi c_{\kappa_C}^2 r \|G\|_{L^2(D\times D)}^2.\end{aligned}$$

Noting that $\operatorname{diam}(Y) \leq r/(2 + \rho)$ and $\int_Y 1 \mathrm{d}y \leq 4\pi(\operatorname{diam}(Y)/2)^3/3$, we have the following inequality for non-admissible domains $X \times Y$:

$$\|G\|_{L^2(X\times Y)}^2 \leq \frac{2\pi^2}{3(2 + \rho)^3} c_{\kappa_C}^2 r^4 \|G\|_{L^2(D\times D)}^2, \tag{19}$$

where $r = (2 + \rho) \max\{\operatorname{diam} X, \operatorname{diam} Y\}$. We conclude that the Green's function restricted to a non-admissible domain has a relatively small norm when the domain itself is small. Therefore, in our approximant $\tilde{G}$ for G, we ignore G on non-admissible domains by setting $\tilde{G}$ to be zero.

4.3 Hierarchical Admissible Partition of Domain

We now describe a hierarchical partitioning of $D \times D$ so that many subdomains are admissible domains, and the non-admissible domains are all small. For ease of notion, we may assume—without loss of generality—that $\operatorname{diam} D = 1$ and $D \subset [0, 1]^3$; otherwise, one should shift and scale D. Moreover, partitioning $[0, 1]^3$ and restricting the partition to D is easier than partitioning D directly (Fig. 2). For the definition of admissible domains, we find it convenient to select $\rho = 1/\sqrt{3}$.

[6] Note that we have normalized [19, Eq. 1.8] to highlight the dependence on $\|G\|_{L^2(D\times D)}$.

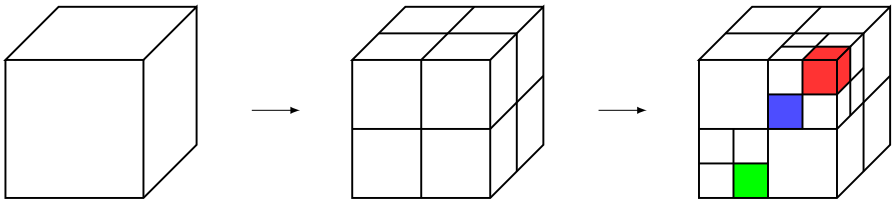

Fig. 2 Two levels of hierarchical partitioning of $[0, 1]^3$. The blue and green domains are admissible, while the blue and red domains are non-admissible (Color figure online)

Let $I = [0, 1]^3$. The hierarchical partitioning for n levels is defined recursively as:

- $I_{1\times 1\times 1} := I_1 \times I_1 \times I_1 = [0, 1]^3$ is the root for level $L = 0$.
- At a given level $0 \le L \le n - 1$, if $I_{j_1\times j_2\times j_3} := I_{j_1} \times I_{j_2} \times I_{j_3}$ is a node of the tree, then it has 8 children defined as

$$\{I_{2j_1+n_j(1)} \times I_{2j_2+n_j(2)} \times I_{2j_3+n_j(3)} \mid n_j \in \{0, 1\}^3\}.$$

Here, if $I_j = [a, b]$, $0 \le a < b \le 1$, then $I_{2j} = \left[a, \frac{a+b}{2}\right]$ and $I_{2j+1} = \left[\frac{a+b}{2}, b\right]$. The set of non-admissible domains can be given by this unwieldy expression

$$P_{\text{non-adm}} = \bigcup_{\substack{\bigwedge_{i=1}^3 |j_i-\tilde{j}_i|\le 1 \\ 2^n\le j_1,j_2,j_3\le 2^{n+1}-1 \\ 2^n\le \tilde{j}_1,\tilde{j}_2,\tilde{j}_3\le 2^{n+1}-1}} I_{j_1\times j_2\times j_3} \times I_{\tilde{j}_1\times\tilde{j}_2\times\tilde{j}_3}, \tag{20}$$

where $\wedge$ is the logical "and" operator. The set of admissible domains is given by

$$P_{\text{adm}} = \bigcup_{L=1}^{n} \Lambda(P_{\text{non-adm}}(L-1))\backslash P_{\text{non-adm}}(L)), \tag{21}$$

where $P_{\text{non-adm}}(L)$ is the set of non-admissible domain for a hierarchical level of L and

$$\Lambda(P_{\text{non-adm}}(L-1)) = \bigcup_{\substack{I_{j_1\times j_2\times j_3}\times I_{\tilde{j}_1\times\tilde{j}_2\times\tilde{j}_3} \\ \in P_{\text{non-adm}}(L-1)}} \bigcup_{n_j,n_{\tilde{j}}\in\{0,1\}^3} I_{\times_{i=1}^3 2j_i+n_j(i)} \times I_{\times_{i=1}^3 2\tilde{j}_i+n_{\tilde{j}}(i)}.$$

Using Eqs. (20)–(21), the number of admissible and non-admissible domains are precisely $|P_{\text{non-adm}}| = (3 \times 2^n - 2)^3$ and $|P_{\text{adm}}| = \sum_{\ell=1}^{n} 2^6(3 \times 2^{L-1} - 2)^3 - (3 \times 2^L - 2)^3$. In particular, the size of the partition at the hierarchical level $0 \le L \le n$ is equal to 8^L and the tree has a total of $(8^{n+1} - 1)/7$ nodes (see Fig. 3).

Finally, the hierarchical partition of $D \times D$ can be defined via the partition $P = P_{\text{adm}} \cup P_{\text{non-adm}}$ of $[0, 1]^3$ by doing the following:

$$D \times D = \bigcup_{\tau\times\sigma\in P} (\tau \cap D) \times (\sigma \cap D).$$

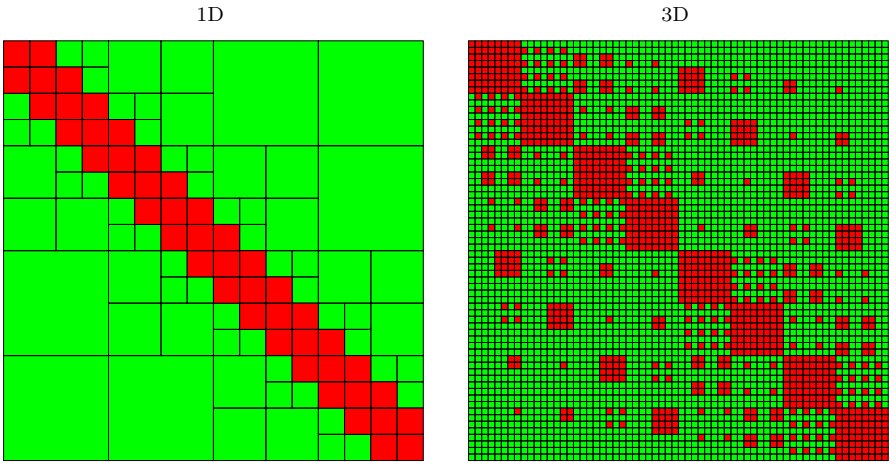

Fig. 3 For illustration purposes, we include the hierarchical structure of the Green's functions in 1D after 4 levels (left) and in 3D after 2 levels (right). The hierarchical structure in 3D is complicated as this is physically a 6-dimensional tensor that has been rearranged so it can be visualized (Color figure online)

The sets of admissible and non-admissible domains of $D \times D$ are denoted by P_{adm} and $P_{\text{non-adm}}$ in the next sections.

4.4 Recovering the Green's Function on the Entire Domain

We now show that we can recover G on the entire domain $D \times D$.

4.4.1 Global Approximation on the Non-Admissible Set

Let n_ϵ be the number of levels in the hierarchical partition $D \times D$ (see Sect. 4.3). We want to make sure that the norm of the Green's function on all non-admissible domains is small so that we can safely ignore that part of G (see Sect. 4.2). As one increases the hierarchical partitioning levels, the volume of the non-admissible domains get smaller (see Fig. 4).

Let $X \times Y \in P_{\text{non-adm}}$ be a non-admissible domain, the two domains X and Y have diameter bounded by $\sqrt{3}/2^{n_\epsilon}$ because they are included in cubes of side length $1/2^{n_\epsilon}$ (see Sect. 4.3). Combining this with Eq. (19) yields

$$\|G\|^2_{L^2(X\times Y)} \leq 2\pi^2(6+\sqrt{3})c^2_{\kappa_C} 2^{-4n_\epsilon} \|G\|^2_{L^2(D\times D)}.$$

Therefore, the L^2-norm of G on the non-admissible domain $P_{\text{non-adm}}$ satisfies

$$\|G\|^2_{L^2(P_{\text{non-adm}})} = \sum_{X\times Y\in P_{\text{non-adm}}} \|G\|^2_{L^2(X\times Y)} \leq 54\pi^2(6+\sqrt{3})c^2_{\kappa_C} 2^{-n_\epsilon} \|G\|^2_{L^2(D\times D)},$$

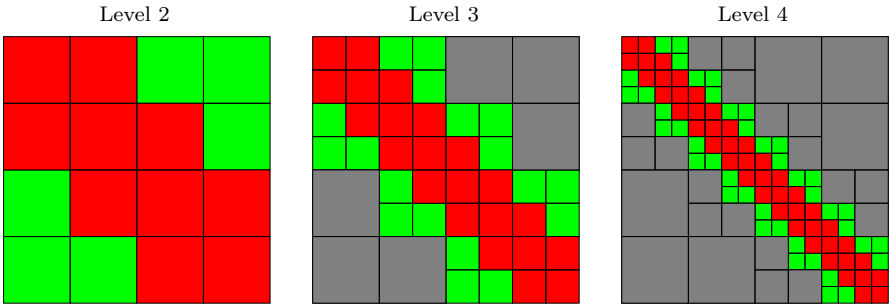

Fig. 4 For illustration purposes, we include the hierarchical structure of the Green function in 1D. The green blocks are admissible domains at that level, the gray blocks are admissible at a higher level, and the red blocks are the non-admissible domains at that level. The area of the non-admissible domains decreases at deeper levels (Color figure online)

where we used $|P_{\text{non-adm}}| = (3 \times 2^{n_\epsilon} - 2)^3 \leq 27(2^{3n_\epsilon})$. This means that if we select n_ϵ to be

$$n_\epsilon = \left\lceil \log_2(54\pi^2(6+\sqrt{3})c_{\kappa_C}^2) + 2\log_2(1/\epsilon) \right\rceil \sim 2\log_2(1/\epsilon), \tag{22}$$

then we guarantee that $\|G\|_{L^2(P_{\text{non-adm}})} \leq \epsilon \|G\|_{L^2(D\times D)}$. We can safely ignore G on non-admissible domains—by taking the zero approximant—while approximating G to within ϵ.

4.4.2 Learning Rate of the Green's Function

Following Sect. 4.1.2, we can construct an approximant $\tilde{G}_{X\times Y}$ to the Green's function on an admissible domain $X \times Y$ of the hierarchical partitioning using the HS randomized SVD algorithm, which requires $N_{\epsilon,X\times Y} = \mathcal{O}(\log^4(1/\epsilon))$ input–output training pairs (see Sect. 4.1.2). Therefore, the number of training input–output pairs needed to construct an approximant to G on all admissible domains is given by

$$N_\epsilon = \sum_{X\times Y \in P_{\text{adm}}} N_{\epsilon,X\times Y} = \mathcal{O}\left(|P_{\text{adm}}| \log^4(1/\epsilon)\right),$$

where $|P_{\text{adm}}|$ denotes the total number of admissible domains at the hierarchical level n_ϵ, which is given by Eq. (22). Then, we have (see Sect. 4.3):

$$|P_{\text{adm}}| = \sum_{\ell=1}^{n_\epsilon} 2^6(3 \times 2^{\ell-1} - 2)^3 - (3 \times 2^\ell - 2)^3 \leq 6^3 2^{3n_\epsilon}, \tag{23}$$

and, using Eq. (22), we obtain $|P_{\text{adm}}| = \mathcal{O}(1/\epsilon^6)$. This means that the total number of required input–output training pairs to learn G with high probability is bounded by

$$N_\epsilon = \mathcal{O}\left(\epsilon^{-6} \log^4(1/\epsilon)\right).$$

4.4.3 Global Approximation Error

We know that with $N_\epsilon = \mathcal{O}(\epsilon^{-6}\log^4(1/\epsilon))$ input–output training pairs, we can construct an accurate approximant to G on each admissible and non-admissible domain. Since the number of admissible and non-admissible domains depends on ϵ, we now check that this implies a globally accurate approximant that we denote by $\tilde{G}$.

Since $\tilde{G}$ is zero on non-admissible domains and $P_{\text{adm}} \cap P_{\text{non-adm}}$ has measure zero, we have

$$\|G - \tilde{G}\|^2_{L^2(D\times D)} \le \epsilon^2 \|G\|^2_{L^2(D\times D)} + \sum_{X\times Y \in P_{\text{adm}}} \|G - \tilde{G}\|^2_{L^2(X\times Y)}. \tag{24}$$

Following Sect. 4.4.2, if $X \times Y$ is admissible then the approximation error satisfies

$$\|G - \tilde{G}_{X\times Y}\|^2_{L^2(X\times Y)} \le 12t^2 s^2 \frac{k_\epsilon}{\gamma_{k_\epsilon, X\times Y}} \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_1} \epsilon^2 \|G\|^2_{L^2(X\times \hat{Y})},$$

with probability greater than $1 - t^{-k_\epsilon} - e^{-2s^2 k_\epsilon}$. Here, $\hat{Y} = \{y \in D,\ \text{dist}(y, Y) \le \text{diam}\, Y/2\sqrt{3}\}$ (see Theorem 4 with $\rho = 1/\sqrt{3}$). To measure the worst $\gamma_{k_\epsilon, X\times Y}$, we define

$$\Gamma_\epsilon = \min\{\gamma_{k_\epsilon, X\times Y} : X \times Y \in P_{\text{adm}}\}. \tag{25}$$

From Eq. (11), we know that $0 < \Gamma_\epsilon \le 1$ and that $1/\Gamma_\epsilon$ is greater than the harmonic mean of the first k_ϵ scaled eigenvalues of the covariance kernel K, i.e.,

$$\frac{1}{\Gamma_\epsilon} \ge \frac{1}{k_\epsilon} \sum_{j=1}^{k_\epsilon} \frac{\lambda_1}{\lambda_j}, \tag{26}$$

Now, one can see that $X \times \hat{Y}$ is included in at most $5^3 = 125$ neighbors including itself. Assuming that all the probability bounds hold on the admissible domains, this implies that

$$\begin{aligned}\sum_{X\times Y\in P_{\text{adm}}} \|G - \tilde{G}\|^2_{L^2(X\times Y)} &\le \sum_{X\times Y\in P_{\text{adm}}} \|G - \tilde{G}\|^2_{L^2(X\times Y)} \le 12t^2 s^2 \frac{k_\epsilon}{\lambda_1 \Gamma_\epsilon} \text{Tr}(K)\epsilon^2 \sum_{X\times Y\in P_{\text{adm}}} \|G\|^2_{L^2(X\times \hat{Y})} \\ &\le 1500 t^2 s^2 \frac{k_\epsilon}{\lambda_1 \Gamma_\epsilon} \text{Tr}(K)\epsilon^2 \|G\|^2_{L^2(D\times D)}.\end{aligned}$$

We then choose $t = e$ and $s = k_\epsilon^{1/4}$ so that the approximation bound on each admissible domain holds with probability of failure less than $2e^{-\sqrt{k_\epsilon}}$. Finally, using Eq. (24) we conclude that as $\epsilon \to 0$, the approximation error on $D \times D$ satisfies

$$\|G - \tilde{G}\|_{L^2(D\times D)} = \mathcal{O}\left(\Gamma_\epsilon^{-1/2} \log^3(1/\epsilon)\epsilon\right) \|G\|_{L^2(D\times D)},$$

with probability $\geq (1-2e^{-\sqrt{k_\epsilon}})^{6^3 2^{3n_\epsilon}} = 1-\mathcal{O}(\epsilon^{\log(1/\epsilon)-6})$, where n_ϵ is given by Eq. (22). We conclude that the approximant $\tilde{G}$ is a good approximation to G with very high probability.

5 Conclusions and Discussion

This paper rigorously learns the Green's function associated with a PDE rather than the partial differential operator (PDO). By extending the randomized SVD to HS operators, we can identify a learning rate associated with elliptic PDOs in three dimensions and bound the number of input–output training pairs required to recover a Green's function approximately. One practical outcome of this work is a measure for the quality of covariance kernels, which may be used to design efficient kernels for PDE learning tasks.

There are several possible future extensions of these results related to the recovery of hierarchical matrices, the study of other partial differential operators, and practical deep learning applications, which we discuss further in this section.

5.1 Fast and Stable Reconstruction of Hierarchical Matrices

We described an algorithm for reconstructing Green's function on admissible domains of a hierarchical partition of $D \times D$ that requires performing the HS randomized SVD $\mathcal{O}(\epsilon^{-6})$ times. We want to reduce it to a factor that is $\mathcal{O}(\text{polylog}(1/\epsilon))$.

For $n \times n$ hierarchical matrices, there are several existing algorithms for recovering the matrix based on matrix-vector products [5,32,36,37]. There are two main approaches: (1) The "bottom-up" approach: one begins at the lowest level of the hierarchy and moves up and (2) The "top-down" approach: one updates the approximant by peeling off the off-diagonal blocks and going down the hierarchy. The bottom-up approach requires $\mathcal{O}(n)$ applications of the randomized SVD algorithm [36]. There are lower complexity alternatives that only require $\mathcal{O}(\log(n))$ matrix-vector products with random vectors [32]. However, the algorithm in [32] is not yet proven to be theoretically stable as errors from low-rank approximations potentially accumulate exponentially, though this is not observed in practice. For symmetric positive semi-definite matrices, it may be possible to employ a sparse Cholesky factorization [54,55]. This leads us to formulate the following challenge:

> **Algorithmic challenge:** Design a provably stable algorithm that can recover an $n \times n$ hierarchical matrix using $\mathcal{O}(\log(n))$ matrix-vector products with high probability?

If one can design such an algorithm and it can be extended to HS operators, then the $\mathcal{O}(\epsilon^{-6}\log^4(1/\epsilon))$ term in Theorem 3 may improve to $\mathcal{O}(\text{polylog}(1/\epsilon))$. This means that the learning rate of partial differential operators of the form of Eq. (2) will be a polynomial in $\log(1/\epsilon)$ and grow sublinearly with respect to $1/\epsilon$.

5.2 Extension to Other Partial Differential Operators

Our learning rate for elliptic PDOs in three variables (see Sect. 4) depends on the decay of the singular values of the Green's function on admissible domains [3]. We expect that one can also find the learning rate for other PDOs.

It is known that the Green's functions associated to elliptic PDOs in two dimensions exist and satisfy the following pointwise estimate [12]:

$$|G(x,y)| \leq C\left(\frac{1}{\gamma R^2} + \log\left(\frac{R}{\|x-y\|_2}\right)\right), \quad \|x-y\|_2 \leq R := \frac{1}{2}\max(d_x, d_y), \tag{27}$$

where $d_x = \operatorname{dist}(x, \partial D)$, γ is a constant depending on the size of the domain D, and C is an implicit constant. One can conclude that $G(x,\cdot)$ is locally integrable for all $x \in D$ with $\|G(x,\cdot)\|_{L^p(B_r(x)\cap D)} < \infty$ for $r > 0$ and $1 \leq p < \infty$. We believe that the pointwise estimate in Eq. (27) implies the off-diagonal low-rank structure of G here, as suggested in [3]. Therefore, we expect that the results in this paper can be extended to elliptic PDOs in two variables.

PDOs in four or more variables are far more challenging since we rely on the following bound on the Green's function on non-admissible domains [19]:

$$G(x,y) \leq \frac{c(d,\kappa_C)}{\lambda_{\min}}\|x-y\|_2^{2-d}, \qquad x \neq y \in D,$$

where $D \subset \mathbb{R}^d$, $d \geq 3$ is the dimension, and c is a constant depending only on d and κ_C. This inequality implies that the L^p-norm of G on non-admissible domains is finite when $0 \leq p < d/(d-2)$. However, for a dimension $d \geq 4$, we have $p < 2$ and one cannot ensure that the L^2 norm of G is finite. Therefore, the Green's function may not be compatible with the HS randomized SVD.

It should also be possible to characterize the learning rate for elliptic PDOs with lower order terms (under reasonable conditions) [13,24,28] and many parabolic operators [29] as the associated Green's functions have similar regularity and pointwise estimates. The main task is to extend [3, Theorem 2.8] to construct separable approximations of the Green's functions on admissible domains. In contrast, we believe that deriving a theoretical learning rate for hyperbolic PDOs remains a significant research challenge for many reasons. The first roadblock is that the Green's function associated with hyperbolic PDOs do not necessarily lie in $L^2(D \times D)$. For example, the Green's function associated with the wave equation in three variables, i.e., $Ł = \partial_t^2 - \nabla^2$, is not square-integrable as

$$G(x,t,y,s) = \frac{\delta(t-s-\|x-y\|_2)}{4\pi\|x-y\|_2}, \qquad (x,t),(y,s) \in \mathbb{R}^3 \times [0,\infty),$$

where $\delta(\cdot)$ is the Dirac delta function.

5.3 Connection with Neural Networks

There are many possible connections between this work and neural networks (NNs) from practical and theoretical viewpoints. The proof of Theorem 3 relies on the construction of a hierarchical partition of the domain $D \times D$ and the HS randomized SVD algorithm applied on each admissible domain. This gives an algorithm for approximating Green's functions with high probability. However, there are more practical approaches that currently do not have theoretical guarantees [17,18].

A promising opportunity is to design a NN that can learn and approximate Green's functions using input–output training pairs $\{(f_j, u_j)\}_{j=1}^N$ [6]. Once a neural network $\mathcal{N}$ has been trained such that $\|\mathcal{N} - G\|_{L^2} \leq \epsilon \|G\|_{L^2}$, the solution to $Łu = f$ can be obtained by computing the following integral:

$$u(x) = \int_D \mathcal{N}(x, y) f(y) \mathrm{d}y.$$

Therefore, this may give an efficient computational approach for discovering operators since a NN is only trained once. Incorporating a priori knowledge of the Green's function into the network architecture design could be particularly beneficial. One could also wrap the selection of the kernel in the GP for generating random functions and training data into a Bayesian framework.

Finally, we wonder how many parameters in a NN are needed to approximate a Green's function associated with elliptic PDOs within a tolerance of $0 < \epsilon < 1$. Can one exploit the off-diagonal low-rank structure of Green's functions to reduce the number of parameters? We expect the recent work on the characterization of ReLU NNs' approximation power is useful [20,44,62]. The use of NNs with high approximation power such as rational NNs might also be of interest to approximate the singularities of the Green's function near the diagonal [7].

Acknowledgements We want to thank Max Jenquin and Tianyi Shi for discussions. We also thank Matthew Colbrook, Abinand Gopal, Daniel Kressner, and Yuji Nakatsukasa for their feedback and suggestions on the paper. We are indebted to Christopher Earls for telling us about the idea of using Green's functions and Gaussian processes for PDE learning. We are grateful to Joel Tropp, whose suggestions led to sharper bounds for the randomized SVD, and the anonymous referees for their comments which improved the quality of the paper.

References

1. Ballani, J., Kressner, D.: Matrices with hierarchical low-rank structures. In: Exploiting Hidden Structure in Matrix Computations: Algorithms and Applications, pp. 161–209. Springer (2016)

2. Bebendorf, M.: Hierarchical Matrices: A Means to Efficiently Solve Elliptic Boundary Value Problems. Lecture Notes in Computational Science and Engineering. Springer-Verlag (2008)
3. Bebendorf, M., Hackbusch, W.: Existence of $\mathscr{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients. Numer. Math. **95**(1), 1–28 (2003)
4. Bonito, A., Cohen, A., DeVore, R., Petrova, G., Welper, G.: Diffusion coefficients estimation for elliptic partial differential equations. SIAM J. Math. Anal. **49**(2), 1570–1592 (2017)
5. Boukaram, W., Turkiyyah, G., Keyes, D.: Randomized GPU algorithms for the construction of hierarchical matrices from matrix-vector operations. SIAM J. Sci. Comput. **41**(4), C339–C366 (2019)
6. Boullé, N., Earls, C.J., Townsend, A.: Data-driven discovery of physical laws with human-understandable deep learning. arXiv preprint arXiv:2105.00266 (2021)
7. Boullé, N., Nakatsukasa, Y., Townsend, A.: Rational neural networks. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 14243–14253 (2020)
8. Boullé, N., Townsend, A.: A generalization of the randomized singular value decomposition. arXiv preprint arXiv:2105.13052 (2021)
9. Brunton, S.L., Proctor, J.L., Kutz, J.N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Natl. Acad. Sci. USA **113**(15) (2016)
10. Chernoff, H.: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann. Math. Stat. pp. 493–507 (1952)
11. de Boor, C.: An alternative approach to (the teaching of) rank, basis, and dimension. Lin. Alg. Appl. **146**, 221–229 (1991)
12. Dong, H., Kim, S.: Green's matrices of second order elliptic systems with measurable coefficients in two dimensional domains. Trans. Am. Math. Soc. **361**(6), 3303–3323 (2009)
13. Dong, H., Kim, S.: Green's function for nondivergence elliptic operators in two dimensions. SIAM J. Math. Anal. 53(4), 4637–4656 (2021). https://doi.org/10.1137/20M1323618
14. Edmunds, D.E., Evans, W.D.: Spectral theory and differential operators. Oxford University Press (2018)
15. Edmunds, D.E., Kokilashvili, V.M., Meskhi, A.: Bounded and compact integral operators. Springer Science & Business Media (2013)
16. Evans, L.C.: Partial Differential Equations. American Mathematical Society, Providence, R.I. (2010)
17. Feliu-Faba, J., Fan, Y., Ying, L.: Meta-learning pseudo-differential operators with deep neural networks. J. Comput. Phys. **408**, 109309 (2020)
18. Gin, C.R., Shea, D.E., Brunton, S.L., Kutz, J.N.: DeepGreen: Deep Learning of Green's Functions for Nonlinear Boundary Value Problems. arXiv preprint arXiv:2101.07206 (2020). https://doi.org/10.1038/s41598-021-00773-xl
19. Grüter, M., Widman, K.O.: The Green function for uniformly elliptic equations. Manuscripta Math. **37**(3), 303–342 (1982)
20. Gühring, I., Kutyniok, G., Petersen, P.: Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. Anal. Appl. **18**(05), 803–859 (2020)
21. Hackbusch, W.: Hierarchical Matrices: Algorithms and Analysis. Springer (2015)
22. Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. **53**(2), 217–288 (2011)
23. Hsing, T., Eubank, R.: Theoretical foundations of functional data analysis, with an introduction to linear operators. John Wiley & Sons (2015)
24. Hwang, S., Kim, S.: Green's function for second order elliptic equations in non-divergence form. Potential Anal. **52**(1), 27–39 (2020)
25. Kang, K., Kim, S.: Global pointwise estimates for Green's matrix of second order elliptic systems. J. Differ. Equ. **249**(11), 2643–2662 (2010)
26. Karhunen, K.: Über lineare methoden in der wahrscheinlichkeitsrechnung. Ann. Acad. Science Fenn., Ser. A. I. **37**, 3–79 (1946)
27. Kato, T.: Perturbation Theory for Linear Operators. Springer Science & Business Media (2013)
28. Kim, S., Sakellaris, G.: Green's function for second order elliptic equations with singular lower order coefficients. Commun. Partial. Differ. Equ. **44**(3), 228–270 (2019)
29. Kim, S., Xu, L.: Green's function for second order parabolic equations with singular lower order coefficients. Commun. Pure Appl. Anal. **21**(1), 1–21 (2022). https://doi.org/10.3934/cpaa.20211641
30. Ledoux, M.: The concentration of measure phenomenon. Math. Surveys. Monog. 89. AMS, Providence, RI (2001)

31. Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., Anandkumar, A.: Fourier neural operator for parametric partial differential equations. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=c8P9NQVtmnO
32. Lin, L., Lu, J., Ying, L.: Fast construction of hierarchical matrix representation from matrix–vector multiplication. J. Comput. Phys. **230**(10), 4071–4087 (2011)
33. Loève, M.: Fonctions aleatoire de second ordre. Rev. Sci. **84**, 195–206 (1946)
34. Long, Z., Lu, Y., Ma, X., Dong, B.: PDE-NET: Learning PDEs from data. In: International Conference on Machine Learning, pp. 3208–3216. PMLR (2018)
35. Maddu, S., Cheeseman, B.L., Sbalzarini, I.F., Müller, C.L.: Stability selection enables robust learning of partial differential equations from limited noisy data. arXiv preprint arXiv:1907.07810 (2019)
36. Martinsson, P.G.: A fast randomized algorithm for computing a hierarchically semiseparable representation of a matrix. SIAM J. Matrix Anal. Appl. **32**(4), 1251–1274 (2011)
37. Martinsson, P.G.: Compressing rank-structured matrices via randomized sampling. SIAM J. Sci. Comput. **38**(4), A1959–A1986 (2016)
38. Meng, X., Li, Z., Zhang, D., Karniadakis, G.E.: PPINN: Parareal physics-informed neural network for time-dependent PDEs. Comput. Methods Appl. Mech. Eng. **370**, 113250 (2020)
39. Mercer, J.: Functions of positive and negative type, and their connection the theory of integral equations. Philos. T. R. Soc. A **209**(441-458), 415–446 (1909)
40. Mood, A.M., Graybill, F.A., Boes, D.C.: Introduction to the Theory of Statistics, 3rd edn. McGraw-Hill (1974)
41. Muirhead, R.J.: Aspects of multivariate statistical theory. John Wiley & Sons (2009)
42. Nakatsukasa, Y.: Fast and stable randomized low-rank matrix approximation. arXiv preprint arXiv:2009.11392 (2020)
43. Pang, G., Yang, L., Karniadakis, G.E.: Neural-net-induced Gaussian process regression for function approximation and PDE solution. J. Comput. Phys. **384**, 270–288 (2019)
44. Petersen, P., Voigtlaender, F.: Optimal approximation of piecewise smooth functions using deep ReLU neural networks. Neural Netw. **108**, 296–330 (2018)
45. Raissi, M.: Deep hidden physics models: Deep learning of nonlinear partial differential equations. J. Mach. Learn. Res. **19**(1), 932–955 (2018)
46. Raissi, M., Karniadakis, G.E.: Hidden physics models: Machine learning of nonlinear partial differential equations. J. Comput. Phys. **357**, 125–141 (2018)
47. Raissi, M., Perdikaris, P., Karniadakis, G.E.: Multistep neural networks for data-driven discovery of nonlinear dynamical systems. arXiv preprint arXiv:1801.01236 (2018)
48. Raissi, M., Perdikaris, P., Karniadakis, G.E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. **378**, 686–707 (2019)
49. Raissi, M., Yazdani, A., Karniadakis, G.E.: Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. Science **367**(6481), 1026–1030 (2020)
50. Rasmussen, C.E., Williams, C.: Gaussian processes for machine learning. MIT Press (2006)
51. Rudin, W.: Real and complex analysis, 3rd edn. McGraw-Hill (1986)
52. Rudy, S.H., Brunton, S.L., Proctor, J.L., Kutz, J.N.: Data-driven discovery of partial differential equations. Sci. Adv. **3**(4), e1602614 (2017)
53. Schaeffer, H.: Learning partial differential equations via data discovery and sparse optimization. Proc. Math. Phys. Eng. Sci. **473**(2197), 20160446 (2017)
54. Schäfer, F., Owhadi, H.: Sparse recovery of elliptic solvers from matrix-vector products. arXiv preprint arXiv:2110.05351 (2021)
55. Schäfer, F., Sullivan, T.J., Owhadi, H.: Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity. Multiscale Model. Sim. **19**(2), 688–730 (2021)
56. Stewart, G.W.: Matrix Algorithms: Volume 1: Basic Decompositions. SIAM (1998)
57. Townsend, A., Trefethen, L.N.: Continuous analogues of matrix factorizations. P. Roy. Soc. A **471**(2173), 20140585 (2015)
58. Trefethen, L.N., Bau III, D.: Numerical linear algebra. SIAM (1997)
59. Voss, H.U., Timmer, J., Kurths, J.: Nonlinear dynamical system identification from uncertain and indirect measurements. Int. J. Bifurc. Chaos Appl. Sci. Eng. **14**(06), 1905–1933 (2004)
60. Wang, Z., Huan, X., Garikipati, K.: Variational system identification of the partial differential equations governing the physics of pattern-formation: inference under varying fidelity and noise. Comput. Methods Appl. Mech. Eng. **356**, 44–74 (2019)

61. Wishart, J.: The generalised product moment distribution in samples from a normal multivariate population. Biometrika pp. 32–52 (1928)
62. Yarotsky, D.: Error bounds for approximations with deep ReLU networks. Neural Netw. **94**, 103–114 (2017)
63. Yazdani, A., Lu, L., Raissi, M., Karniadakis, G.E.: Systems biology informed deep learning for inferring parameters and hidden dynamics. PLoS Comput. Biol. **16**(11), e1007575 (2020)
64. Zhao, H., Storey, B.D., Braatz, R.D., Bazant, M.Z.: Learning the physics of pattern formation from images. Phys. Rev. Lett. **124**(6), 060201 (2020)