Journal of Machine Learning Research 23 (2022) 1-34 Submitted 4/22; Published 8/22

Learning Green's functions associated with time-dependent partial differential equations

Nicolas Boullé BOULLE@MATHS.OX.AC.UK
Mathematical Institute
University of Oxford
Oxford, OX2 6GG, UK

Seick Kim KIMSEICK@YONSEI.AC.KR
Department of Mathematics
Yonsei University
Seoul, 03722, ROK

Tianyi Shi TS777@CORNELL.EDU
Center for Applied Mathematics
Cornell University
Ithaca, NY 14853, USA

Alex Townsend TOWNSEND@CORNELL.EDU
Department of Mathematics
Cornell University
Ithaca, NY 14853, USA

Editor: Michael Mahoney

Abstract

Neural operators are a popular technique in scientific machine learning to learn a mathematical model of the behavior of unknown physical systems from data. Neural operators are especially useful to learn solution operators associated with partial differential equations (PDEs) from pairs of forcing functions and solutions when numerical solvers are not available or the underlying physics is poorly understood. In this work, we attempt to provide theoretical foundations to understand the amount of training data needed to learn time-dependent PDEs. Given input-output pairs from a parabolic PDE in any spatial dimension $n \geq 1$, we derive the first theoretically rigorous scheme for learning the associated solution operator, which takes the form of a convolution with a Green's function G. Until now, rigorously learning Green's functions associated with time-dependent PDEs has been a major challenge in the field of scientific machine learning because G may not be square-integrable when $n > 1$, and time-dependent PDEs have transient dynamics. By combining the hierarchical low-rank structure of G together with randomized numerical linear algebra, we construct an approximant to G that achieves a relative error of $\mathcal{O}(\Gamma_\epsilon^{-1/2}\epsilon)$ in the L^1-norm with high probability by using at most $\mathcal{O}(\epsilon^{-\frac{n+2}{2}}\log(1/\epsilon))$ input-output training pairs, where Γ_ϵ is a measure of the quality of the training dataset for learning G, and $\epsilon > 0$ is sufficiently small.

Keywords: Scientific machine learning, data-driven model, time-dependent PDE, Green's function

1. Introduction

Machine learning, numerical analysis, and scientific computing are successfully combining in the field of scientific machine learning to integrate data and prior knowledge of physical laws to solve inverse problems using deep learning (Karniadakis et al., 2021). The flexibility of neural network architectures and exceptional generalization errors, make neural networks ideal for scientific machine learning. On the other hand, it is challenging to mathematically justify the success of deep learning in this context.

A central topic in scientific machine learning is to discover partial differential equations (PDEs), which are mathematical models describing the relations between variables of a system and their spatial and temporal derivatives, directly from simulations or experimental data. This leads to a wide range of applications in weather forecasting and climate science (Rasp et al., 2018; Zanna and Bolton, 2020), biology (Alber et al., 2019; Raissi et al., 2020), and physics (Karniadakis et al., 2021; Kochkov et al., 2021; Kutz, 2017; Chen and Gu, 2021). Traditionally, PDEs are derived from mechanistic insights using conservation laws, minimum energy principles, or empirical observations (Evans, 2010). With the rapid development of deep learning and the vast collection of experimental results from sensors, we are beginning an exciting new era of uncovering unknown PDE models directly from data. Still, learning time-dependent PDEs is challenging because of transient dynamics.

We aim to provide the first theoretical results to characterize how much training data is needed to learn a time-dependent PDE, and close a theoretical gap with recent data-driven methods (Boullé et al., 2022; Gin et al., 2021; Li et al., 2021; Lu et al., 2021). Our main result, described in Section 1.3, exploits and draws connections with standard tools from numerical analysis, such as approximation theory and numerical linear algebra. While we exclusively focus on theory, the insights provided by this work will be of interest to a broader audience in scientific machine learning and motivate future empirical works and novel physics-informed neural network architectures.

1.1 Parabolic Partial Differential Equations

Throughout this paper, we consider a class of time-dependent PDEs called parabolic partial differential operators. A parabolic partial differential operator defined on a bounded spatial domain $\Omega \subset \mathbb{R}^n$ for some $n \geq 1$ with Lipschitz smooth boundary takes the form:

$$\mathcal{P}u \coloneqq u_t - \nabla \cdot (A(x,t)\nabla u) = f(x,t), \quad x \in \Omega,\, t \in [0,T], \quad 0 < T < \infty. \tag{1}$$

Here, for every $x \in \Omega$ and $t \in [0,T]$, the matrix $A(x,t) \in \mathbb{R}^{n\times n}$ is symmetric positive definite with bounded coefficient functions in $L^\infty(\mathcal{U})$, where $\mathcal{U} \coloneqq \Omega \times [0,T]$, and satisfies the uniform parabolicity condition (see Eq. (3)). Here, $L^\infty(\mathcal{U})$ is the space of measurable functions defined on $\mathcal{U}$ that have a bounded essential supremum. In this manuscript, we also consider two other L^p spaces, the space of absolutely integrable functions, $L^1(\mathcal{U})$, and squared-integrable functions, $L^2(\mathcal{U})$. We emphasize that the regularity requirements on the parabolic PDE are very weak. The function f in Eq. (1) is called the forcing term of the PDE while u is the corresponding system's response or solution. Parabolic PDEs model a wide variety of time-dependent phenomena, including heat conduction, particle diffusion, and option pricing.

The goal of most PDE learning tasks is to learn the solution operator that maps forcing terms to responses, given training data $\{(f_j, u_j)\}_{j=1}^N$ (Boullé et al., 2022; Gin et al., 2021; Kovachki et al., 2021b; Li et al., 2020a,b, 2021; Lu et al., 2021; Wang et al., 2021). Associated with the parabolic operator $\mathcal{P}$ in Eq. (1) is a Green's function $G : \mathcal{U} \times \mathcal{U} \to \mathbb{R}^+$, which is a kernel for the solution operator (Cho et al., 2012). In particular, the solution operator is an integral operator of the form (Evans, 2010):

$$u(x,t) = \int_0^T \int_\Omega G(x,t,y,s) f(y,s) \, \mathrm{d}y \, \mathrm{d}s, \quad (x,t) \in \mathcal{U},$$

where u is the solution to Eq. (1) given the forcing term f. Our goal is to recover G as accurately as possible from forcing functions $f_1, \ldots, f_N$ and their corresponding solutions $u_1, \ldots, u_N$, as well as the evaluation of the adjoint of $\mathcal{P}$. Since we are learning a classical mathematical object, we can gain a mechanistic understanding of the unknown parabolic PDE, and theoretical and practical performance guarantees.

1.2 Challenges and Contributions

In this paper, we derive a rigorous probabilistic algorithm to learn the Green's function G associated with Eq. (1) from random input-output data (f, u) and characterize the number of training pairs needed to learn G to within a given tolerance ϵ with high probability. Since Green's functions associated with Eq. (1) may not be squared-integrable when $n > 1$, we perform our analysis using the L^1-norm and obtain a rigorous learning rate for G in that norm. We summarize the challenges that we face and our main contributions:

Low-rank structure of parabolic Green's functions. It is known that Green's functions associated with elliptic PDEs in dimension $n \geq 3$ have a low-rank structure on well-separated domains and can be approximated by separable functions (Bebendorf and Hackbusch, 2003). This property motivates the use of hierarchical matrices as a way to store, approximate, and compute the inverse of finite element stiffness matrices and discretized Green's functions in quasi-optimal complexity (Bebendorf and Rjasanow, 2003; Bebendorf, 2008; Börm et al., 2003; Hackbusch, 1999; Hackbusch and Khoromskij, 2000; Hackbusch et al., 2004). The low-rank structure of the Green's function is heavily used in numerical solvers for elliptic PDEs, preconditioners for iterative solvers, computing Schur complements (Bebendorf and Hackbusch, 2003), and rigorously learning Green's functions from input-output pairs (Boullé and Townsend, 2022a). While related works (Greengard and Lin, 2000; Greengard and Strain, 1990; Jiang et al., 2015; Li and Greengard, 2007, 2009) exploited the compressibility of the heat kernel to build fast and accurate numerical methods for the evaluation of heat potentials, there is a lack of theoretical results that are analogous to those found in Bebendorf and Hackbusch (2003). In particular, our first contribution is to prove that Green's functions associated with parabolic PDEs admit a low-rank structure on well-separated domains for any spatial dimension, extending (Bebendorf and Hackbusch, 2003). Our analysis is based on the existence of Poincaré and Cacciopolli-type inequalities satisfied by the solutions of parabolic PDEs. We find that for the most efficient hierarchical partitioning of a domain, the time variable must be treated differently from spatial variables, leading to a careful hierarchical partition of the spatio-temporal domain. This result enables the approximation of the entire solution operator associated with a parabolic

PDE by a hierarchical matrix, leading to efficient numerical solvers. Additionally, inspired by the hierarchical structures in elliptic PDEs, several neural network (NN) architectures using a wavelet transform are proposed to learn the solution operators of PDEs across different scales (Feliu-Faba et al., 2020; Gupta et al., 2021), and our analysis suggests this is potentially a good idea for parabolic PDEs too.

Analysis in the $\mathbf{L^1}$-norm. Green's functions associated with parabolic PDEs may not be squared-integrable when $n > 1$, presenting an additional challenge compared to elliptic PDEs with $1 \leq n \leq 3$. For example, consider the forced heat equation in dimension $n > 1$ with zero homogeneous Dirichlet boundary conditions and zero initial conditions, i.e.,

$$\frac{\partial u}{\partial t} - \nabla^2 u = f(x,t), \quad u(x,0) = 0, \quad u(0,t) = 0, \quad (x,t) \in \mathbb{R}^n \times \mathbb{R},$$

The associated Green's function is given by (Evans, 2010, Sec. 2.3.1)

$$G(x,t,y,s) = \frac{\Theta(t-s)}{(4\pi(t-s))^{n/2}} \exp\left(-\frac{1}{4}\frac{|x-y|^2}{t-s}\right), \quad (x,t) \neq (y,s) \in \mathbb{R}^n \times \mathbb{R}, \tag{2}$$

where $\Theta(\cdot)$ is the Heaviside step function that takes the value of one for positive inputs and zero otherwise, and $|\cdot|$ is the Euclidean norm on Ω. Due to the type of the time singularity along the diagonal as s approaches t, G is not a squared-integrable function. However, G does have a bounded L^1-norm. This is a very significant theoretical challenge for rigorously learning the corresponding solution operator as L^1 is not a Hilbert space, contrary to L^2. Almost all the techniques employed in the elliptic case (Boullé and Townsend, 2022a) exploit analogues of matrix results to Hilbert–Schmidt (HS) operators in infinite dimensions, such as the Eckart–Young–Mirsky theorem for best low-rank approximation in the Frobenius norm (Eckart and Young, 1936), and do not generalize to the L^1-norm. The best low-rank approximation problem for matrices in the entrywise ℓ_1-norm is significantly more complicated than that in the Frobenius norm and is, in general, NP-hard (Gillis and Vavasis, 2018; Song et al., 2017). We address this issue by approximating well-separated blocks of the Green's function in the L^2-norm, and then express the final approximation error in the L^1-norm using Moser's local maximum estimate (Lieberman, 1996). This theory may motivate the use of NN architectures that allow for representing maps with singularities that are not square-integrable in deep learning, such as rational NNs (Boullé et al., 2020).

Quality of the training data. Several deep learning techniques for learning solution operators associated with PDEs assume random training and testing forcing functions f in Eq. (1), which are drawn from a Gaussian process (GP) $\mathcal{GP}(0, K)$ with a carefully designed covariance kernel K (Boullé et al., 2022; Gin et al., 2021; Kovachki et al., 2021b; Li et al., 2020a, 2021; Lu et al., 2021; Wang et al., 2021). For example, the GreenLearning, DeepGreen, and DeepONet methods use a squared-exponential kernel, i.e., $K(x,x') = \exp(-|x-x'|^2/(2\ell^2))$, where the length-scale parameter ℓ determines the smoothness of the forcing terms (Boullé et al., 2022; Gin et al., 2021; Lu et al., 2021). In contrast, the Neural Operator approach employs Green's functions associated with Helmholtz equations as covariance kernel, where the Helmholtz frequency varies depending on the problem considered (Li et al., 2020a, 2021). We emphasize that the choice of the covariance kernel is important in PDE learning applications and can be used to enforce prior knowledge about

the PDE to obtain higher accuracy. Our main theoretical result contains a term that characterizes the quality of the random training data, i.e., the covariance kernel of the GP from which forcing terms are sampled (see Section 1.3). This is a step towards understanding the success of state-of-the-art PDE learning techniques and better determine the optimal covariance kernel to minimize the size of the training dataset.

Finally, we regard our work here as giving important theoretical insights and we are not proposing that our rigorous learning algorithm should replace deep learning techniques in practice. Instead, we hope this paper can benefit the state-of-the-art PDE learning techniques by suggesting different optimization algorithms based on an L^1 loss, improving the quality of training datasets, and designing "physics-informed" NN architectures that represent the singularity and low-rank structure present in Green's functions.

1.3 Main Theoretical Results

Our first result (see Theorem 9) shows that Green's functions associated with parabolic operators of the form of Eq. (1) satisfy similar separable approximation properties to Green's functions of elliptic operators (Bebendorf and Hackbusch, 2003, Thm. 2.8) on admissible spatio-temporal domains $Q_X \times Q_Y \subset \mathcal{U} \times \mathcal{U}$. The notion of admissibility (see Section 2.2) ensures that the approximation results only apply to domains $Q_X \times Q_Y$ that do not contain the singular points of the Green's function. For any $0 < \epsilon < 1$ sufficiently small and $k \leq k_\epsilon = \mathcal{O}(\lceil \log \frac{1}{\epsilon} \rceil^{n+3})$, we show that there exists a (low-rank) separable approximation of the form

$$G_k(x,t,y,s) = \sum_{i=1}^{k} u_i(x,t) v_i(y,s), \quad (x,t) \in Q_X,\ (y,s) \in Q_Y,$$

such that

$$\|G - G_k\|_{L^2(Q_X \times Q_Y)} \leq \epsilon \|G\|_{L^2(\hat{Q}_X \times Q_Y)}.$$

Here, $\hat{Q}_X \subset \mathcal{U}$ denotes a domain slightly larger than Q_X.

Throughout this paper, we make the following assumption that allows us to evaluate the adjoint of the parabolic operator to construct an approximant to the Green's function. In practical applications, it may not be possible to evaluate the adjoint, as backward parabolic equations are usually not well-posed (John, 1955; Miranker, 1961). However, numerical experiments suggest that deep NNs can approximate the solution operators associated with non-symmetric problems when the training data contains sufficient prior knowledge of the operator (Boullé et al., 2022; Li et al., 2021; Lu et al., 2021).

Assumption 1. *We assume that we can evaluate the action of the adjoint $\mathcal{P}^*$ of the parabolic operator $\mathcal{P}$, defined as*

$$\mathcal{P}^* u = -u_t - \nabla \cdot (A(x,t)^\top \nabla u),$$

where $A^\top$ is the transpose of the coefficient matrix A.

Our main theoretical result, stated later in Theorem 10, constructs a rigorous probabilistic scheme for learning Green's functions of parabolic operators in spatial dimension $n \geq 1$ within a relative error of $\mathcal{O}(\Gamma_\epsilon^{-1/2}\epsilon)$, with high probability, using at most $\mathcal{O}(\epsilon^{-\frac{n+2}{2}} \log(1/\epsilon))$

input-output training pairs, where $\epsilon > 0$ is a sufficiently small parameter. The factor Γ_ϵ is defined later in Section 4.4 and quantifies the quality of the forcing terms for learning G. This result provides an upper bound for the intrinsic learning rate of parabolic operators. Our theoretical construction relies on the separable approximation result for Green's functions associated with parabolic PDEs described earlier, a careful hierarchical partition of the spatio-temporal domain into well-separated blocks, and the randomized singular value decomposition (SVD) for HS operators (Boullé and Townsend, 2022a,b).

1.4 Related Works

The approaches that dominate the PDE learning literature consist of discovering coefficients of the PDE (Brunton et al., 2016; Rudy et al., 2017; Udrescu and Tegmark, 2020; Udrescu et al., 2020; Zhang and Lin, 2018), building reduced-order models to significantly speed up standard numerical solvers (Qian et al., 2020, 2021), and directly approximating the PDE solution operator by an artificial NN (Boullé et al., 2022; Gin et al., 2021; Kovachki et al., 2021b; Li et al., 2020a,b, 2021; Lu et al., 2021; Wang et al., 2021). Several black-box deep learning techniques are proposed to approximate the solution operator, which maps forcing terms f to observations of the associated system's response u such that $\mathcal{P}(u) = f$, where $\mathcal{P}$ is the partial differential operator. These methods mainly differ in their choice of the NN architecture used to approximate the solution map. For example, Fourier Neural Operator (Li et al., 2021) uses a Fourier transform at each layer, DeepONet (Lu et al., 2021) contains a concatenation of 'trunk' and 'branch' networks to enforce additional structure, and GreenLearning (Boullé et al., 2022) relies on rational NNs (Boullé et al., 2020) to learn Green's functions.

On the theoretical side, most of the research has focused on the approximation theory of infinite-dimensional operators by NNs, such as the generalization of the universal approximation theorem (Cybenko, 1989) to shallow and deep NNs (Chen and Chen, 1995; Lu et al., 2021) as well as error estimates for Fourier Neural Operators and DeepONets with respect to the network width and depth (Kovachki et al., 2021b,a; Lanthaler et al., 2022). Other approaches aim to approximate the matrix of the discretized Green's functions associated with elliptic PDEs from matrix-vector multiplications by exploiting sparsity patterns or hierarchical structure of the matrix (Lin et al., 2011; Schäfer and Owhadi, 2021). In addition, de Hoop et al. (2021) derived convergence rates for learning linear self-adjoint operators based on the assumption that the target operator is diagonal in the basis of the Gaussian prior. More recently and closely related to this work, Boullé and Townsend (2022a) derived an intrinsic "learning rate" for elliptic PDEs using ideas from randomized linear algebra and low-rank approximation theory to characterize the number of training data needed to approximate the associated solution operator or Green's functions.

However, fundamental challenges of the field concern the interpretability of the discovered model to uncover new physical understanding, and performance guarantees with theoretical results. These are challenging, especially when the underlying mathematical model is time-dependent and has short-lived transient dynamics.

1.5 Organization of the Paper

The paper is organized as follows. We first introduce some definitions and our notation in Section 2. Then, we prove a low-rank approximation property of Green's functions associated with parabolic operators in Section 3 on separable domains. We exploit this low rank structure to bound the number of input-output pairs needed to learn Green's functions in Section 4 using the randomized SVD combined with a hierarchical partition of the temporal domain. We conclude in Section 5 with a discussion of the results and future challenges.

2. Background and the Randomized Singular Value Decomposition

This section introduces our notation and background on low-rank functions, admissible domains, and the randomized SVD for HS operators.

2.1 Definitions and Notation

Throughout this paper, $\Omega \subset \mathbb{R}^n$ denotes a bounded domain in spatial dimension $n \geq 1$ satisfying the *uniform interior cone condition* (Gilbarg and Trudinger, 2001, Chapt. 7.7), which is defined as follows.

Definition 1 (Uniform interior cone condition)**.** *We say that Ω satisfies an interior cone condition if there exists an angle $\theta \in (0, \pi/2)$ and a radius $r > 0$ such that for every $x \in \Omega$ there exists a unit vector ξ_x such that the cone*

$$C(x, \xi_x, \theta, r) = \{x + \lambda y : y \in \mathbb{R}^n,\ \|y\|_2 = 1,\ y \cdot \xi_x \geq \cos(\theta),\ \lambda \in [0, r]\}$$

is contained in Ω. Here, '$\cdot$' denotes the standard dot product in $\mathbb{R}^n$.

We note that every bounded domain with a Lipschitz smooth boundary satisfies an interior cone condition.

We consider parabolic PDEs of the form Eq. (1) on the domain $\mathcal{U} = \Omega \times [0, T]$, where $T > 0$. We also assume that the symmetric coefficient matrix $A(x, t) \in \mathbb{R}^{n \times n}$ satisfies the uniform parabolicity condition, i.e., there exist two positive constants $\lambda, \Lambda > 0$ such that

$$\lambda |\xi|^2 \leq A(x, t)\xi \cdot \xi \leq \Lambda |\xi|^2, \quad \xi \in \mathbb{R}^n, \tag{3}$$

where $|\cdot|$ denotes the discrete ℓ_2-norm. This means that the matrix A is uniformly positive definite with eigenvalues in the interval $[\lambda, \Lambda]$ By the Cauchy–Schwarz inequality, we have the following inequality:

$$|A(x, t)\xi \cdot \psi| \leq \Lambda |\xi||\psi|, \quad \xi, \psi \in \mathbb{R}^n.$$

Under these conditions, we can find a nonnegative Green's function $G(x, t, y, s)$ defined on the domain $\{(x, t, y, s) \in \mathcal{U} \times \mathcal{U},\ (x, t) \neq (y, s)\}$ by the following relationship (Cho et al., 2012):

$$\begin{aligned} \mathcal{P}G(x, t, y, s) &= \delta(y - x)\delta(s - t), && (x, t) \in \mathcal{U}, \\ G(\cdot, t, y, s) &= 0, && \text{on } \partial\Omega,\ t \in (0, T), \\ G(\cdot, 0, y, s) &= 0, && \text{in } \Omega, \end{aligned}$$

where $\mathcal{P}$ is the parabolic operator defined in Eq. (1) acting on functions in the variables (x,t) and $\delta(\cdot)$ is the Kronecker delta function.

In this manuscript, we usually work on domains of the form $Q = (\Omega \cap D) \times I$, where D is a bounded convex domain in $\mathbb{R}^n$ and I is an open bounded interval of $\mathbb{R}^+$, and consider the following function spaces:

1. The Banach space $L^1(Q)$, with norm $\|u\|_{L^1(Q)} = \int_Q |u| \,\mathrm{d}x \,\mathrm{d}t$.
2. The Hilbert space $L^2(Q)$, with inner product $\langle u, v \rangle_{L^2(Q)} = \int_Q uv \,\mathrm{d}x \,\mathrm{d}t$.
3. The Hilbert space $W_2^{1,0}(Q)$, with inner product $\langle u, v \rangle_{W_2^{1,0}(Q)} = \int_Q (uv + \nabla u \cdot \nabla v) \,\mathrm{d}x \,\mathrm{d}t$, consisting of all functions $u \in L^2(\mathcal{U})$ with squared-integrable weak derivatives.
4. The Banach space $V_2(Q)$, defined as

$$V_2(Q) \coloneqq \left\{ u \in W_2^{1,0}(Q),\, \|u\|_{V_2(Q)} = \operatorname*{ess\,sup}_{t \in I} \|u(\cdot,t)\|_{L^2(\Omega \cap D)} + \|\nabla u\|_{L^2(Q)} < \infty \right\}.$$

The approximation error between the learned and exact Green's functions are expressed in the $L^1(\mathcal{U} \times \mathcal{U})$-norm as Green's functions associated with parabolic PDEs are usually not squared-integrable when $n > 1$ (see Section 1.2).

2.2 Admissible Domains and Low-rank Functions

We learn Green's functions on subdomains of $\mathcal{U} \times \mathcal{U}$ satisfying an admissibility condition so that the subdomains do not contain the singular points of the Green's functions located along the diagonal $(x,t) = (y,s)$. While the definition of strong admissible (or well-separated) domains is standard for Green's functions associated with elliptic PDEs (Ballani and Kressner, 2016; Bebendorf and Hackbusch, 2003; Bebendorf, 2008; Hackbusch, 2015), we need to adapt the definition for parabolic PDEs slightly. Let $\beta > 0$ be a constant, and consider the following metrics on $\mathcal{U} \times \mathcal{U}$:

$$m(x,t,y,s) = \max\left(\|x - y\|_\infty, \sqrt{|t-s|/\beta} \right), \quad (x,t) \in \mathcal{U},\, (y,s) \in \mathcal{U}, \tag{5}$$

where the spatial and temporal variables are treated differently. The choice of the metric m is related to the exponential term appearing in the Green's function of the heat equation (cf. Eq. (2)) since Green's functions associated with parabolic PDEs satisfy similar Gaussian bounds (Cho et al., 2012). Let $Q_X, Q_Y \subset \mathcal{U}$ be two non-empty domains, we can define the *diameter* of Q_X and the *distance* between Q_X and Q_Y using the metric m as

$$\operatorname{diam} Q_X = \sup_{(x,t),(y,s) \in Q_X} m(x,t,y,s), \qquad \operatorname{dist}(Q_X, Q_Y) = \inf_{(x,t) \in Q_X, (y,s) \in Q_Y} m(x,t,y,s). \tag{6}$$

We use these quantities to define a partition of $\mathcal{U} \times \mathcal{U}$ so that the spatial and temporal domains scale similarly as we approach the singularity of the Green's function (see Section 4.1). Finally, one can combine the notions of diameter and distance for spatio-temporal domains to introduce an admissibility condition, similar to the elliptic case.

Definition 2 (Admissible domains)**.** *For a fixed parameter $\rho > 0$, we say that two non-empty domains $Q_X, Q_Y \subset \mathcal{U}$ are admissible if*

$$\operatorname{dist}(Q_X, Q_Y) \geq \rho \max\{\operatorname{diam} Q_X, \operatorname{diam} Q_Y\}. \tag{7}$$

Otherwise, we say that they are non-admissible.

Fig. 1 illustrates admissible and non-admissible subdomains of $\mathcal{U} = [0,1] \times [0,1]$. In particular, the spatial component of $\mathcal{U}$ is partitioned into four subdomains while the temporal component is partitioned into two. This ensures that all the subdomains have a similar diameter according to Eq. (6). We use a similar strategy when constructing a partition of $\mathcal{U} \times \mathcal{U}$ into admissible and non-admissible subdomains (see Section 4.1).

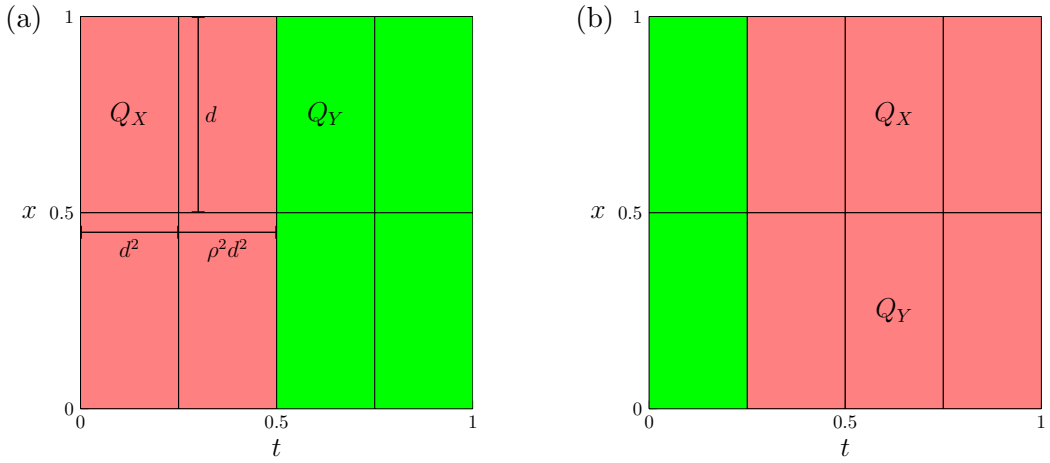

Figure 1: Admissible and non-admissible subdomains of the spatial-temporal domain $\mathcal{U} = [0,1] \times [0,1]$. The spatial domain is partitioned into two while the temporal domain is partitioned into four. In particular, a subdomain $Q_X = D_X \times I_X$ has diameter $d := \operatorname{diam}(Q_X) = \operatorname{diam}(D_X) = \operatorname{diam}(I_X)^{1/2}$. Panels (a) and (b) highlights in green (resp. red) the admissible domains (resp. non-admissible) with Q_X with diameter $d = 1/2$ and $\rho = \beta = 1$ (see Eq. (5) and Theorem 2). Specifically, the subdomain Q_Y is admissible with Q_X in (a) and non-admissible in (b).

On admissible domains $Q_X \times Q_Y \subset \mathcal{U} \times \mathcal{U}$, we aim to construct approximants to the Green's function G associated with Eq. (1). For a given accuracy $0 < \epsilon < 1$, we say that G is of rank k if there exists an integer $k = k(\epsilon)$ and a separable approximation of the form

$$G_k(x,t,y,s) = \sum_{i=1}^{k} u_i(x,t) v_i(y,s), \quad (x,t,y,s) \in Q_X \times Q_Y,$$

such that $\|G - G_k\|_{L^2(Q_X \times Q_Y)} \leq \epsilon \|G\|_{L^2(\hat{Q}_X \times \hat{Q}_Y)}$, where $\hat{Q}_X$ (resp. $\hat{Q}_Y$) denotes a domain slightly larger than Q_X (resp. Q_Y); see Theorem 9 for a precise definition. When $k = \mathcal{O}(\log^{\delta}(1/\epsilon))$ for some small $\delta \in \mathbb{N}$ as $\epsilon \to 0$, then we say that G has rapidly decaying singular values on $Q_X \times Q_Y$.

2.3 Randomized Singular Value Decomposition for Hilbert–Schmidt Operators

Let $D_1, D_2 \subset \mathbb{R}^n$ be two domains, a linear operator $\mathscr{F} : L^2(D_1) \to L^2(D_2)$ is called an HS operator if it has finite HS norm, i.e.,

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

where $\{e_j\}_{j=1}^{\infty}$ is an orthonormal basis of $L^2(D_1)$. HS operators generalize the notion of matrices acting on vectors to infinite dimensions with operators acting on squared-integrable functions. Moreover, the HS norm is the continuous analogue of the Frobenius norm for matrices and $\|\mathscr{F}\|_{\text{HS}} = (\sum_{j=1}^{\infty} \sigma_j^2)^{1/2}$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ denote the singular values of $\mathscr{F}$.

The randomized SVD is one of the most popular algorithms for constructing low-rank approximations of large matrices. Given a matrix A, it uses matrix-vector products with random vectors drawn from a standard Gaussian distribution to find an approximate orthonormal basis Q for the column space of A before computing a low-rank approximation as $QQ^\top A$ (Halko et al., 2011; Martinsson and Tropp, 2020).

A recent generalization of the randomized SVD with random vectors drawn from a general multivariate Gaussian distribution allows its application to learn HS operators using random functions drawn from a Gaussian process (Boullé and Townsend, 2022a,b). The randomized SVD for HS operators uses random functions drawn from a Gaussian process $\mathcal{GP}(0, K)$ with mean $(0, \ldots, 0)$ and continuous symmetric positive definite covariance kernel $K : D_1 \times D_1 \to \mathbb{R}$ to construct a low-rank approximant. The kernel K has positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots > 0$, and there exists an orthonormal basis of $L^2(D_1)$ such that (Hsing and Eubank, 2015, Thm. 4.6.5)

$$K(x, y) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x) \psi_j(y), \quad x, y \in D_1.$$

The trace of the covariance kernel is defined as $\text{Tr}(K) \coloneqq \sum_{j=1}^{\infty} \lambda_j < \infty$ and is finite as K is continuous on $D_1 \times D_1$. A random function $\omega \sim \mathcal{GP}(0, K)$ can be sampled by setting $\omega \sim \sum_{j=1}^{\infty} \sqrt{\lambda_j} c_j \psi_j$, where $c_j \sim \mathcal{N}(0, 1)$ are independent and identically distributed.

Finally, it is convenient to introduce quasimatrices, which extends the notion of tall-skinny matrices to infinite dimensions (Townsend and Trefethen, 2015), to formulate the randomized SVD for HS operators. Let $k \geq 1$ be an integer, a $D_1 \times k$ quasimatrix $\boldsymbol{\Omega} = \begin{bmatrix} \omega_1 & \cdots & \omega_k \end{bmatrix}$ is a matrix with k columns, where each column $1 \leq j \leq k$ is a squared-integrable function $w_j \in L^2(D_1)$. The standard matrix-vector operations generalize naturally to the applications of HS operators to quasimatrices. Then, $\mathscr{F}\boldsymbol{\Omega}$ denotes the quasimatrix obtained by applying $\mathscr{F}$ to each column of $\boldsymbol{\Omega}$ (Boullé and Townsend, 2022a, Sec. 2.1).

We can now state the results of approximating an HS operator with randomized SVD. Let $k \geq 1$ be a target rank, $p \geq 4$ an oversampling parameter, and $\boldsymbol{\Omega}$ be a $D_1 \times (k + p)$ quasimatrix, whose columns are drawn from $\mathcal{GP}(0, K)$. If $\mathbf{Y} = \mathscr{F}\boldsymbol{\Omega}$ and $\mathbf{P}_{\mathbf{Y}}$ is the orthogonal projection onto the vector space spanned by the columns of $\mathbf{Y}$, then for $s, t \geq 1$,

we have (Boullé and Townsend, 2022a, Thm. 1),

$$\|\mathscr{F} - \mathbf{P}_{\mathbf{Y}}\mathscr{F}\|_{\text{HS}} \leq \sqrt{1 + t^2 s^2 \frac{3}{\gamma_k} \frac{k(k+p)}{p+1} \frac{\text{Tr}(K)}{\lambda_1}} \left(\sum_{j=k+1}^{\infty} \sigma_j^2 \right)^{1/2}, \tag{8}$$

with probability of failure bounded by $t^{-p} + [se^{-(s^2-1)/2}]^{k+p}$. Here, $\gamma_k = k/(\lambda_1 \,\text{Tr}(\mathbf{C}^{-1}))$ 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 $1 \leq i, j \leq k$, where v_j is the jth right singular vector of $\mathscr{F}$. The factor $0 < \gamma_k \leq 1$ in Eq. (8) characterizes the quality of the covariance kernel to learn the HS operator $\mathscr{F}$. A refined bound shows that one can enforce prior information on the operator in the covariance kernel to obtain near-best approximation error (Boullé and Townsend, 2022b, Thm. 2).

In the remainder of this manuscript, we apply the randomized SVD for HS operators to learn Green's functions associated with parabolic PDEs on admissible domains $Q_X \times Q_Y$. Green's functions restricted to admissible domains are an example of HS integral operators as the solution operator $\mathscr{F}$ associated with parabolic PDEs can be written as

$$(\mathscr{F} f)(x,t) = \int_{Q_Y} G(x,t,y,s) f(y,s) \,\mathrm{d}y \,\mathrm{d}s, \quad f \in L^2(Q_X),\ (x,t) \in Q_X.$$

Moreover, we can use the relation $\|\mathscr{F}\|_{\text{HS}} = \|G\|_{L^2(Q_X \times Q_Y)}$ to estimate the approximation error between the Green's function and its approximant on $Q_X \times Q_Y$.

3. Low-Rank Approximation of Parabolic Green's Functions

Bebendorf and Hackbusch (2003) show that Green's functions associated with elliptic equations in three dimensions admit a low rank separable approximation on admissible domains. In this section, we extend this result to Green's functions associated with parabolic PDEs so that we obtain low-rank approximants on well-separated domains (see Theorem 9). In particular, approximations in this section are expressed in L^2-norm, and we convert the relations to L^1-norm in the next section.

3.1 Poincaré-type Inequality

We start our derivation by showing a Poincaré-type inequality for the solution of a parabolic equation Eq. (1), which bounds a function by its derivatives and the geometry of its domain. The standard Poincaré's inequality is of the form $\|u - \bar{u}\|_{L^2(D)} \leq C \|\nabla u\|_{L^2(D)}$, where $\bar{u}$ is the average of u in D, and C is some positive constant. In Eq. (10) we find that when D is convex, a closed-form expression for the constant C can be derived.

Lemma 3. *Let D be a bounded convex domain in $\mathbb{R}^n$ and let $u \in W^{1,1}(D) = \{f \in L^1(D) : \partial_x f \in L^1(D)\}$ where ∂_x is taken in the weak sense. Let η be a nonnegative function such that $\int_D \eta \,\mathrm{d}y > 0$ and $0 \leq \eta(y) \leq 1$. Then, for $x \in D$, we have*

$$|u(x) - \bar{u}_\eta| \leq \frac{d^n}{n \int_D \eta \,\mathrm{d}y} \int_D |x - y|^{1-n} \,|\nabla u(y)| \,\mathrm{d}y, \tag{9}$$

where $\bar{u}_\eta = \int_D u(y)\eta(y)\,\mathrm{d}y / \int_D \eta(y)\,\mathrm{d}y$ *and* $d = \operatorname{diam} D$. *In particular, we have*

$$\|u - \bar{u}_\eta\|_{L^2(D)} \leq \frac{\omega_n^{1-\frac{1}{n}} |D|^{\frac{1}{n}}}{\int_D \eta\,\mathrm{d}y} d^n \|\nabla u\|_{L^2(D)}, \tag{10}$$

where ω_n *denotes the volume of the unit ball in* $\mathbb{R}^n$.

Proof From the Fundamental Theorem of Calculus,

$$u(x) - u(y) = -\int_0^{|x-y|} \partial_r u(x + r\omega)\,\mathrm{d}r, \quad \omega = \frac{y-x}{|y-x|}, \quad x, y \in D.$$

By multiplying by $\eta(y)$ on both sides and integrating with respect to y over D, we obtain

$$(u(x) - \bar{u}_\eta) \int_D \eta(y)\,\mathrm{d}y = -\int_D \eta(y) \int_0^{|x-y|} \partial_r u(x + r\omega)\,\mathrm{d}r\,\mathrm{d}y.$$

For $x \in \mathbb{R}^n$, we define the function

$$V(x + r\omega) = \begin{cases} |\partial_r u(x + r\omega)|, & \text{if } x + r\omega \in D, \\ 0, & \text{otherwise.} \end{cases}$$

Then, we have

$$\begin{aligned} |u(x) - \bar{u}_\eta| &\leq \frac{1}{\int_D \eta\,\mathrm{d}y} \int_D \int_0^\infty V(x + r\omega)\,\mathrm{d}r\,\mathrm{d}y \\ &= \frac{1}{\int_D \eta\,\mathrm{d}y} \int_0^\infty \int_{|\omega|=1} \int_0^d V(x + r\omega)\rho^{n-1}\,\mathrm{d}\rho\,\mathrm{d}\omega\,\mathrm{d}r \\ &= \frac{d^n}{n\int_D \eta\,\mathrm{d}y} \int_0^\infty \int_{|\omega|=1} V(x + r\omega)\,\mathrm{d}\omega\,\mathrm{d}r = \frac{d^n}{n\int_D \eta\,\mathrm{d}y} \int_D |x - y|^{1-n} |\nabla u(y)|\,\mathrm{d}y. \end{aligned}$$

Note that we have used the fact that $\eta \leq 1$ to obtain the first inequality. Finally, Eq. (10) is obtained by applying (Gilbarg and Trudinger, 2001, Lem. 7.12) to Eq. (9), with $p = q = 2$ and $\mu = 1/n$. ∎

It is worth noting the differences between the standard Poincaré inequality and the one in Theorem 3 as the average of u is replaced by a "weighted" average of u. This is important for deriving Theorem 4.

The importance of this lemma is that one can bound the L^2-norm of u minus some constant by the L^2 norm of the spatial gradient (not the full gradient) of u when u is a solution of $\mathcal{P}u = 0$. It is complementary to the parabolic Caccioppoli's inequality (see Theorem 5), where the spatial gradient of u is controlled by u itself.

Lemma 4. *Let* $\Omega \subset \mathbb{R}^n$ *be a domain and* $D \subset \mathbb{R}^n$ *be a bounded convex set such that* $\Omega \cap D \neq \emptyset$. *Suppose there is a constant* $\theta \in (0, 1)$ *such that one of the following holds:*

1. $|D \setminus \Omega| \geq \theta |D|$.
2. *There exists a ball* $B \subset \Omega \cap D$ *such that* $|B| \geq \theta |D|$.

Then, for any u satisfying $\mathcal{P}u = 0$ in $Q = (\Omega \cap D) \times I$, where I is an open bounded interval of $\mathbb{R}^+$, and $u = 0$ on $(\partial\Omega \cap D) \times I$, there exists a constant $c \in \mathbb{R}$ such that

$$\|u - c\|_{L^2(Q)} \leq \left(\frac{2^{2n}}{\theta^2} \left(\frac{\omega_n}{|D|} \right)^{2-\frac{2}{n}} d^{2n} + \frac{2^{2n+3}\Lambda^2 \omega_n^{\frac{2}{n}} |I|^2}{\theta^{2+\frac{2}{n}} |D|^{\frac{2}{n}}} \right)^{1/2} \|\nabla u\|_{L^2(Q)}, \tag{11}$$

where d and ω_n are defined in Theorem 3, and Λ is a constant related to uniform parabolicity defined in Eq. (3)*.*

Proof We denote $\tilde{Q} = D \times I$ and extend u by zero on $\tilde{Q} \setminus Q$ so that u is defined on $\tilde{Q}$. The proof is done in two steps assuming one of the conditions.

1. We first consider the case when $|D \setminus \Omega| \geq \theta |D|$. If η is the indicator function of $D \setminus \Omega$, then Theorem 3 yields

$$\int_D |u(x,t)|^2 \,\mathrm{d}x \leq \frac{\omega_n^{2-\frac{2}{n}} |D|^{\frac{2}{n}}}{|D \setminus \Omega|^2} d^{2n} \int_D |\nabla u(x,t)|^2 \,\mathrm{d}x, \quad t \in I.$$

By integrating with respect to $t \in I$ and using $|D \setminus \Omega| \geq \theta |D|$, we find that

$$\int_I \int_D |u(x,t)|^2 \,\mathrm{d}x \,\mathrm{d}t \leq \frac{\omega_n^{2-\frac{2}{n}} |D|^{\frac{2}{n}}}{\theta^2 |D|^2} d^{2n} \int_I \int_D |\nabla u(x,t)|^2 \,\mathrm{d}x \,\mathrm{d}t. \tag{12}$$

Since $u = 0$ in $\tilde{Q} \setminus Q$, (12) implies Eq. (11) with $c = 0$.

2. Next, we consider the case when there exists a ball $B = B(x_0, r) \subset \Omega \cap D$ such that $|B| \geq \theta |D|$. Let η be a smooth function such that

$$0 \leq \eta \leq 1, \quad \operatorname{supp} \eta \subset B, \quad |\nabla \eta| \leq \frac{2}{r}, \quad \text{and} \quad \int_B \eta(x) \,\mathrm{d}x \geq \frac{1}{2^n} |B|. \tag{13}$$

We denote

$$\bar{u}_\eta(t) \coloneqq \frac{1}{\int_D \eta \,\mathrm{d}x} \int_D u(x,t) \eta(x) \,\mathrm{d}x, \quad \hat{u}_\eta \coloneqq \frac{1}{|I|} \int_I \bar{u}_\eta(t) \,\mathrm{d}t = \frac{1}{\int_{\tilde{Q}} \eta \,\mathrm{d}x \,\mathrm{d}t} \int_{\tilde{Q}} u(x,t) \eta(x) \,\mathrm{d}x \,\mathrm{d}t.$$

Then, we have

$$\begin{aligned} \int_{\tilde{Q}} |u(x,t) - \hat{u}_\eta|^2 \,\mathrm{d}x \,\mathrm{d}t &\leq 2 \int_{\tilde{Q}} |u(x,t) - \bar{u}_\eta(t)|^2 + |\bar{u}_\eta(t) - \hat{u}_\eta|^2 \,\mathrm{d}x \,\mathrm{d}t \\ &\leq \frac{2^{2n}}{\theta^2} \left(\frac{\omega_n}{|D|} \right)^{2-\frac{2}{n}} d^{2n} \int_I \int_D |\nabla u(x,t)|^2 \,\mathrm{d}x \,\mathrm{d}t + 2|D| \int_I |\bar{u}_\eta(t) - \hat{u}_\eta|^2 \,\mathrm{d}t, \end{aligned} \tag{14}$$

where the second inequality follows from Eqs. (10) and (13), and the assumption that $|B| \geq \theta |D|$.

We now bound the second term in the right-hand side of Eq. (14). Since $\mathcal{P}u = \partial_t u - \nabla \cdot (A\nabla u) = 0$ in $Q = (\Omega \cap D) \times I$, $u = 0$ on $(\partial\Omega \cap D) \times I$, and $\operatorname{supp}\eta \subset B \subset \Omega \cap D$, we multiply the equation $\mathcal{P}u = 0$ by η and integrate by parts over $\Omega \cap D$ to obtain

$$\int_{\Omega\cap D} \frac{\partial}{\partial t} u(x,t)\eta(x)\,\mathrm{d}x + \int_{\Omega\cap D} A(x,t)\nabla u(x,t)\cdot\nabla\eta(x)\,\mathrm{d}x = 0.$$

Then, integrating over $t \in [t_0, t_1] \subset I$ yields

$$\begin{aligned}\left|\int_{t_0}^{t_1}\left(\frac{d}{dt}\int_{\Omega\cap D} u(x,t)\eta(x)\,\mathrm{d}x\right)\,\mathrm{d}t\right| &= \left|\int_{t_0}^{t_1}\int_{\Omega\cap D} A(x,t)\nabla u(x,t)\cdot\nabla\eta(x)\,\mathrm{d}x\,\mathrm{d}t\right| \\ &\leq \int_{t_0}^{t_1}\int_{\Omega\cap D} |A(x,t)\nabla u(x,t)\cdot\nabla\eta(x)|\,\mathrm{d}x\,\mathrm{d}t \\ &\leq \frac{2\Lambda}{r}\int_Q |\nabla u|\,\mathrm{d}x\,\mathrm{d}t = 2\Lambda\left(\frac{\omega_n}{|B|}\right)^{\frac{1}{n}}\int_{\tilde{Q}} |\nabla u|\,\mathrm{d}x\,\mathrm{d}t,\end{aligned} \tag{15}$$

where the second inequality follows from Eq. (13) and the uniform parabolicity condition. We now use the fact that $\int_B \eta\,\mathrm{d}x \geq \frac{1}{2^n}|B|$ to obtain

$$\left|\int_{t_0}^{t_1}\left(\frac{d}{dt}\int_{\Omega\cap D} u(x,t)\eta(x)\,\mathrm{d}x\right)\,\mathrm{d}t\right| = \left|\int_B \eta\,\mathrm{d}x\int_{t_0}^{t_1}\frac{d}{dt}\bar{u}_\eta(t)\,\mathrm{d}t\right| \geq \frac{|B|}{2^n}|\bar{u}_\eta(t_1) - \bar{u}_\eta(t_0)|,$$

which when combined with Eq. (15), gives the following inequality:

$$\frac{|B|}{2^n}|\bar{u}_\eta(t_1) - \bar{u}_\eta(t_0)| \leq 2\Lambda\left(\frac{\omega_n}{|B|}\right)^{\frac{1}{n}}\int_{\tilde{Q}} |\nabla u|\,\mathrm{d}x\,\mathrm{d}t.$$

Therefore, we have

$$|\bar{u}_\eta(t_1) - \bar{u}_\eta(t_0)| \leq \frac{2^{n+1}\Lambda\omega_n^{\frac{1}{n}}}{|B|^{1+\frac{1}{n}}}\int_{\tilde{Q}} |\nabla u|\,\mathrm{d}x\,\mathrm{d}t \leq \frac{2^{n+1}\Lambda\omega_n^{\frac{1}{n}}}{\theta^{1+\frac{1}{n}}|D|^{1+\frac{1}{n}}}\int_{\tilde{Q}} |\nabla u|\,\mathrm{d}x\,\mathrm{d}t, \tag{16}$$

as $|B| \geq \theta|D|$. Moreover, we have,

$$\bar{u}_\eta(t) - \hat{u}_\eta = \frac{1}{|I|}\int_I \bar{u}_\eta(t) - \bar{u}_\eta(s)\,\mathrm{d}s, \quad t \in I$$

and, using Eq. (16), we deduce that

$$2|D|\int_I |\bar{u}_\eta(t) - \hat{u}_\eta|^2\,\mathrm{d}t \leq 2|\tilde{Q}|\left(\frac{2^{n+1}\Lambda\omega_n^{\frac{1}{n}}}{\theta^{1+\frac{1}{n}}|D|^{1+\frac{1}{n}}}\int_{\tilde{Q}} |\nabla u|\,\mathrm{d}x\,\mathrm{d}t\right)^2.$$

We combine this equation with Eq. (14) to obtain

$$\int_{\tilde{Q}} |u - \bar{u}_\eta|^2\,\mathrm{d}x\,\mathrm{d}t \leq \left(\frac{2^{2n}}{\theta^2}\left(\frac{\omega_n}{|D|}\right)^{2-\frac{2}{n}} d^{2n} + \frac{2^{2n+3}\Lambda^2\omega_n^{\frac{2}{n}}|I|^2}{\theta^{2+\frac{2}{n}}|D|^{\frac{2}{n}}}\right)\int_{\tilde{Q}} |\nabla u|^2\,\mathrm{d}x\,\mathrm{d}t,$$

and choose the constant $c = \bar{u}_\eta$ to conclude the proof of Eq. (11).

■

If the diameter d of the spatial domain D satisfies $d \simeq |I|^{\frac{1}{2}} \simeq |D|^{\frac{1}{n}}$, where $|D| \coloneqq \int_D \mathrm{d}x$ denotes the volume of D, then Eq. (11) can be simplified to

$$\|u - c\|_{L^2(Q)} \lesssim d\|\nabla u\|_{L^2(Q)},$$

where $\lesssim$ denotes an inequality up to the multiplication of the right-hand side by a constant. For example, this situation arises when $D \times I = Q_r^-(X_0) \coloneqq B_r(x_0) \times (t_0 - r^2, t_0]$, for a given $t_0 > 0$. The assumptions of Theorem 4 can be shown to hold in some simple contexts. For example, if Ω satisfies the uniform interior cone condition (Gilbarg and Trudinger, 2001, Chapt. 7.7) and D is such that $|D| \simeq d^n$, such as a cube or a ball, then there exists a constant $\theta \in (0, 1)$ and $\delta > 0$ depending on Ω, such that if $d \leq \delta$, one of the conditions of Theorem 4 holds.

3.2 Cacciopolli's Inequality

Next, we show Cacciopolli's inequality for solutions of parabolic equations (1), which can also be seen as a kind of reverse Poincaré inequality. In particular, it says that the L^2 norm of the spatial gradient (not the full gradient) of u can be bounded above by the L^2 norm of u on a slightly larger domain.

Lemma 5. *Let D be a domain such that $D \cap \Omega \neq \emptyset$, $\Gamma \coloneqq \partial D \cap \overline{\Omega}$, and $K \subset D$ be a domain such that $\delta_0 \coloneqq \mathrm{dist}_{L^2(D)}(K, \Gamma) > 0$. Let $I = (t_0, t_1) \subset \mathbb{R}$ and $I' = (t_0 + \delta_1, t_1)$, for a given $0 < \delta_1 < t_1 - t_0$. Then, for any u satisfying $\mathcal{P}u = 0$ in $Q \coloneqq (D \cap \Omega) \times I$ and $u = 0$ on $(\partial\Omega \cap D) \times I$, we have*

$$\int_{I'} \int_{K \cap \Omega} |\nabla u(x,t)|^2 \,\mathrm{d}x \,\mathrm{d}t \leq \left(\frac{4\Lambda^2}{\lambda^2 \delta_0^2} + \frac{2}{\lambda \delta_1} \right) \|u\|^2_{L^2(Q)}.$$

Proof Let $\eta \in C^1(\mathbb{R}^n)$ such that $0 \leq \eta \leq 1$, $\eta = 1$ on K, $\eta = 0$ in a neighbourhood of Γ, and $|\nabla \eta| \leq 1/\delta_0$. We also consider a similar function $\zeta \in C^1(\mathbb{R})$, defined on the temporal domain, such that $0 \leq \zeta \leq 1$, $\zeta = 1$ on I', $\zeta = 0$ in a neighbourhood of t_0, and $|\zeta'| \leq 1/\delta_1$.

Then, the function $\eta^2\zeta^2 u$, defined on $(D \cap \Omega) \times (t_0, t)$ for $t_0 < t < t_1$, satisfies the equation $\mathcal{P}(\eta^2\zeta^2 u) = 0$. After multiplying the equation by u, integrating by parts over $(D \cap \Omega) \times (t_0, t)$, and using the fact that η vanishes near Γ, which gives $\eta^2\zeta^2 u = 0$ on $\partial(D \cap \Omega)$, we obtain,

$$\int_{t_0}^{t} \int_{D \cap \Omega} \eta^2(x)\zeta^2(t) u(x,t) \frac{\partial}{\partial t} u(x,t) \,\mathrm{d}x \,\mathrm{d}t + \int_{t_0}^{t} \int_{D \cap \Omega} A(x,t) \nabla u \cdot \nabla(\eta^2\zeta^2 u) \,\mathrm{d}x \,\mathrm{d}t = 0. \quad (17)$$

The first term in Eq. (17) can be reformulated as

$$\begin{aligned} \int_{t_0}^{t} \int_{D \cap \Omega} \eta^2(x)\zeta^2(t) u \frac{\partial}{\partial t} u \,\mathrm{d}x \,\mathrm{d}t &= \int_{t_0}^{t} \int_{D \cap \Omega} \frac{\partial}{\partial t} \left(\frac{1}{2} \eta^2(x)\zeta^2(t) u^2 \right) - \eta^2(x)\zeta(t)\zeta'(t) u^2 \,\mathrm{d}x \,\mathrm{d}t. \\ &= \int_{D \cap \Omega} \frac{1}{2} \eta^2(x)\zeta^2(t) u(x,t)^2 \,\mathrm{d}x - \int_{t_0}^{t} \int_{D \cap \Omega} \eta^2 \zeta \zeta' u^2 \,\mathrm{d}x \,\mathrm{d}t, \end{aligned}$$

since ζ vanishes near t_0. Then, the second term of Eq. (17) satisfies

$$\int_{t_0}^{t}\int_{D\cap\Omega} A(x,t)\nabla u\cdot\nabla(\eta^2\zeta^2 u)\,\mathrm{d}x\,\mathrm{d}t = \int_{t_0}^{t}\int_{D\cap\Omega} \eta^2\zeta^2 A\nabla u\cdot\nabla u + 2\eta\zeta^2 uA\nabla u\cdot\nabla\eta\,\mathrm{d}x\,\mathrm{d}t.$$

Combining these two terms gives the following equality:

$$\begin{aligned}\int_{D\cap\Omega}\frac{1}{2}\eta^2\zeta^2(t)u^2\,\mathrm{d}x + \int_{t_0}^{t}\int_{D\cap\Omega}\eta^2\zeta^2 A\nabla u\cdot\nabla u \le{}& \int_{t_0}^{t}\int_{D\cap\Omega} 2\eta\zeta^2|u||A\nabla u\cdot\nabla\eta|\,\mathrm{d}x\,\mathrm{d}t\\ &+\int_{t_0}^{t}\int_{D\cap\Omega}\eta^2\zeta|\zeta'|u^2\,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

We now use the uniform parabolicity of $\mathcal{P}$: $\lambda|\nabla u|^2 \le A\nabla u\cdot\nabla u$ and $|A\nabla u\cdot\nabla\eta| \le \Lambda|\nabla u||\nabla\eta|$ to obtain

$$\begin{aligned}&\int_{D\cap\Omega}\frac{1}{2}\eta^2\zeta^2(t)u^2\,\mathrm{d}x + \lambda\int_{t_0}^{t}\int_{D\cap\Omega}\eta^2\zeta^2|\nabla u|^2\,\mathrm{d}x\,\mathrm{d}t\\ &\qquad\le 2\Lambda\int_{t_0}^{t}\int_{D\cap\Omega}\eta\zeta^2|u||\nabla u||\nabla\eta|\,\mathrm{d}x\,\mathrm{d}t + \int_{t_0}^{t}\int_{D\cap\Omega}\eta^2\zeta|\zeta'|u^2\,\mathrm{d}x\,\mathrm{d}t\\ &\le\epsilon\int_{t_0}^{t}\int_{D\cap\Omega}\eta^2\zeta^2|\nabla u|^2\,\mathrm{d}x\,\mathrm{d}t + \frac{\Lambda^2}{\epsilon}\int_{t_0}^{t}\int_{D\cap\Omega}\zeta^2|\nabla\eta|^2u^2\,\mathrm{d}x\,\mathrm{d}t + \int_{t_0}^{t}\int_{D\cap\Omega}\eta^2\zeta|\zeta'|u^2\,\mathrm{d}x\,\mathrm{d}t,\end{aligned}$$

where the second inequality follows from Young's inequality: $ab \le (a^2+b^2)/2$ with $a = \sqrt{2\epsilon}\eta|\nabla u|$ and $b = \Lambda\sqrt{2/\epsilon}|u||\nabla\eta|$, for any $\epsilon > 0$. Then, choosing $\epsilon = \lambda/2$ and using the properties $|\nabla\eta| \le 1/\delta_0$, $|\zeta'| \le 1/\delta_1$, $0 \le \eta \le 1$, and $0 \le \zeta \le 1$, we find that

$$\frac{1}{2}\int_{D\cap\Omega}\eta^2\zeta^2(t)u^2\,\mathrm{d}x + \frac{\lambda}{2}\int_{t_0}^{t}\int_{D\cap\Omega}\eta^2\zeta^2|\nabla u|^2\,\mathrm{d}x\,\mathrm{d}t \le \left(\frac{2\Lambda^2}{\lambda\delta_0^2}+\frac{1}{\delta_1}\right)\int_{t_0}^{t}\int_{D\cap\Omega}u^2\,\mathrm{d}x\,\mathrm{d}t. \quad (18)$$

The result follows from the properties of η and ζ. ∎

Since $t \in (t_0, t_1)$ is arbitrary in Eq. (18), we also obtain the following inequality:

$$\operatorname*{ess\,sup}_{t\in I'}\int_{K\cap\Omega}|u(x,t)|^2\,\mathrm{d}x \le \left(\frac{4\Lambda^2}{\lambda\delta_0^2}+\frac{2}{\delta_1}\right)\int_I\int_{D\cap\Omega}|u(x,t)|^2\,\mathrm{d}x\,\mathrm{d}t,$$

which can be combined with Theorem 5 to find that

$$\operatorname*{ess\,sup}_{t\in I'}\int_{K\cap\Omega}|u(x,t)|^2\,\mathrm{d}x + \int_{I'}\int_{K\cap\Omega}|\nabla u(x,t)|^2\,\mathrm{d}x\,\mathrm{d}t \le C\int_I\int_{D\cap\Omega}|u(x,t)|^2\,\mathrm{d}x\,\mathrm{d}t, \quad (19)$$

where C is a constant depending on λ, Λ, K, and I'. From the definition of the Banach space V_2, Eq. (19) can be understood as an upper bound of the norm of solution u with respect to this Banach space. Using the bound given by Eq. (19), we can introduce a subspace $\mathscr{X}(D\times I)$ of $L^2(D\times I)$ and prove its closeness.

Lemma 6. *Let D be a domain such that $D\cap\Omega \neq \emptyset$, $\Gamma = \partial D\cap\overline{\Omega}$, $I = (t_0, t_1)$. Define $\mathscr{X}(D\times I)$ to be the subspace of $L^2(D\times I)$ consisting of the functions u satisfying the following conditions.*

1. $u = 0$ *on* $(D \setminus \Omega) \times (t_0, t_1)$.
2. $u \in V_2(Q')$ *for all* $Q' = K \times (t_0 + \tau, t_1)$, *where* $K \subset D$ *with* $\operatorname{dist}(K, \Gamma) > 0$ *and* $0 < \tau < t_1 - t_0$.
3. $\mathcal{P}u = 0$ *in* $(D \cap \Omega) \times (t_0, t_1)$ *in the sense that, for almost all* $t \in (t_0, t_1)$, *the equality*

$$\int_{D\cap\Omega} u(x,t)\eta(x,t)\,\mathrm{d}x - \int_0^t \int_{D\cap\Omega} u\partial_t\eta\,\mathrm{d}x\,\mathrm{d}t + \int_0^t \int_{D\cap\Omega} A\nabla u\cdot\nabla\eta\,\mathrm{d}x\,\mathrm{d}t = 0 \qquad (20)$$

holds for all smooth test function η *vanishing near* $\partial(D\cap\Omega)\times(t_0,t_1)$ *and* $\overline{D\cap\Omega}\times\{t_0\}$.

Then, the space $\mathscr{X}(D \times I)$ *is closed in* $L^2(D \times I)$.

Proof Let $v \in L^2(D \times I)$ and $\{v_k\}_{k\in\mathbb{N}} \subset \mathscr{X}(D \times I)$ be a sequence converging to v, we want to show that $v \in \mathscr{X}(D \times I)$. First, using Eq. (19), we have

$$\|v_k\|_{V_2(Q')} \le C\|v_k\|_{L^2(D\times I)}.$$

Following the Banach–Alaoglu Theorem, there exists a subsequence $\{v_{i_k}\}_{k\in\mathbb{N}}$ of $\{v_k\}_{k\in\mathbb{N}}$ that converges weakly in $V_2(Q')$ to $\hat{v} \in V_2(Q')$. Therefore, for any $w \in L^2(Q')$, we have $\langle v, w\rangle_{L^2(Q')} = \lim_{k\to\infty}\langle v_{i_k}, w\rangle_{L^2(Q')} = \langle \hat{v}, w\rangle_{L^2(Q')}$, which shows that $v = \hat{v} \in V_2(D\times I)$. By the same argument v satisfies Eq. (20). Finally, $v_k = 0$ on $(D \setminus \Omega) \times (t_0, t_1)$ implies that $v = 0$ on $(D \setminus \Omega) \times (t_0, t_1)$, and $v \in \mathscr{X}(D \times I)$. ■

3.3 Separable Approximation

The following two lemmas quantify the dimension of a finite-dimensional subspace W needed to approximate a function in $\mathscr{X}(D \times I)$ up to a prescribed relative error. In this way, given a parabolic equation solution and the desired accuracy, we can determine the rank of the separable approximant of the corresponding Green's function. We begin by defining a finite-dimensional subspace V_k of the space $\mathscr{X}(D \times I)$, and bounding the distance between parabolic equation solutions and V_k.

Lemma 7. *Let* $\Omega \subset \mathbb{R}^n$ *be a domain satisfying the uniform interior cone condition,* $D = \{x \in \mathbb{R}^n : \|x - x_0\|_\infty < d/2\}$ *be a cube in* $\mathbb{R}^n$ *of side length* $d > 0$ *centered at* $x_0 \in \mathbb{R}^n$ *such that* $\Omega \cap D \neq \emptyset$, *and* $I = (t_0, t_0 + \beta d^2)$ *for some constant* $\beta > 0$ *and* $t_0 > 0$. *Then there exists* $\delta_0 > 0$ *depending only on* Ω *such that for any* $k \ge (1 + \lceil d/\delta_0\rceil)^{n+2}$, *there is a subspace* $V_k \subset \mathscr{X}(D \times I)$ *with* $\dim V_k \le k$ *such that*

$$\operatorname{dist}_{L^2(D\times I)}(u, V_k) \le c_{\mathrm{appr}}\, k^{-\frac{1}{n+2}}\, d\, \|\nabla u\|_{L^2(D\times I)}, \quad u \in \mathscr{X}(D \times I) \cap V_2(D \times I), \qquad (21)$$

where $\mathscr{X}(D \times I)$ *is defined by Theorem 6,* $c_{\mathrm{appr}} = 2^{n+2}(\omega_n^{2-2/n}\theta^{-2} + 2\Lambda^2\omega_n^{2/n}\beta^2\theta^{-2-2/n})^{1/2}$, *and* $\theta = \theta(\Omega)$ *is determined by the characteristics of the uniform cone condition.*

Proof Let $\ell \ge 1$, we first subdivide the cube D uniformly into ℓ^n sub-cubes and subdivide the interval (t_0, t_1) into ℓ^2 subintervals to form ℓ^{n+2} cylinders of the form $Q_i = D_i \times I_i$, where

D_i is a cube of side length d/ℓ and I_i is an interval of length $\beta(d/\ell)^2$, for $1 \leq i \leq \ell^{n+2}$. Since Ω satisfies the uniform interior cone condition, there exists $\delta_0 = \delta_0(\Omega) > 0$ and $\theta = \theta(\Omega) > 0$ such that if $d/\ell \leq \delta_0$, then either of the conditions 1 or 2 in Theorem 4 holds for all D_i satisfying $D_i \cap \Omega \neq \emptyset$.

We now choose $\ell \geq \lceil d/\delta_0 \rceil$ so that $d/\ell \leq \delta_0$, and define the space W of piecewise constant functions on the domains Q_i by

$$W \coloneqq \{v \in L^2(D \times I) : v \text{ is constant on } Q_i \text{ for all } 1 \leq i \leq \ell^{n+2}\},$$

then $\dim W = \ell^{n+2}$. Let $1 \leq i \leq \ell^{n+2}$, we first consider the case where $D_i \cap \Omega \neq \emptyset$. We know that D_i is convex, $|D_i| = (d/\ell)^n$, $\operatorname{diam}(D_i) = \sqrt[n]{2}d/\ell$, and $|I_i| = \beta d^2/\ell^2$. According to Theorem 4, there exists a constant $c_i \in \mathbb{R}$ such that

$$\int_{Q_i} |u - c_i|^2 \,\mathrm{d}x\,\mathrm{d}t \leq \left(\frac{2^{2n+2}\omega_n^{2-\frac{2}{n}}}{\theta^2} + \frac{2^{2n+3}\Lambda^2\omega_n^{\frac{2}{n}}\beta^2}{\theta^{2+\frac{2}{n}}} \right) \frac{d^2}{\ell^2} \int_{Q_i} |\nabla u|^2 \,\mathrm{d}x\,\mathrm{d}t. \tag{22}$$

If $D_i \cap \Omega = \emptyset$, then $u \in \mathscr{X}(D \times I)$ implies that $u = 0$ on Q_i, and Eq. (22) holds with $c_i = 0$.

Next, we define a piecewise constant function $\bar{u} \in W$ such that $\bar{u}|_{Q_i} = c_i$ for $1 \leq i \leq \ell^{n+2}$. Summing Eq. (22) over i yields the following inequality

$$\|u - \bar{u}\|_{L^2(D\times I)} \leq \left(\frac{2^{2n+2}\omega_n^{2-\frac{2}{n}}}{\theta^2} + \frac{2^{2n+3}\Lambda^2\omega_n^{\frac{2}{n}}\beta^2}{\theta^{2+\frac{2}{n}}} \right)^{1/2} \frac{d}{\ell} \|\nabla u\|_{L^2(D\times I)}.$$

Let $k \geq (1+\lceil d/\delta_0 \rceil)^{n+2}$ be an integer and choose $\ell = \lfloor k^{\frac{1}{n+2}} \rfloor$ such that $\ell^{n+2} \leq k < (\ell+1)^{n+2}$, $\ell \geq \lceil d/\delta_0 \rceil$, and $\dim W \leq \ell^{n+2} \leq k$. Now, since $1/\ell \leq 2/(\ell+1) \leq 2k^{-\frac{1}{n+2}}$, we have

$$\|u - \bar{u}\|_{L^2(D\times I)} \leq \left(\frac{2^{2n+4}\omega_n^{2-\frac{2}{n}}}{\theta^2} + \frac{2^{2n+5}\Lambda^2\omega_n^{\frac{2}{n}}\beta^2}{\theta^{2+\frac{2}{n}}} \right)^{1/2} k^{-\frac{1}{n+2}} d \|\nabla u\|_{L^2(D\times I)}. \tag{23}$$

Finally, let $P : L^2(D \times I) \to \mathscr{X}(D \times I)$ be the $L^2(D \times I)$-orthogonal projection onto $\mathscr{X}(D \times I)$ and $V_k \coloneqq P(W)$. The statement of the lemma follows from Eq. (23) and

$$\operatorname{dist}_{L^2(D\times I)}(u, V_k) \leq \|u - P(\bar{u})\|_{L^2(D\times I)} = \|P(u - \bar{u})\|_{L^2(D\times I)} \leq \|u - \bar{u}\|_{L^2(D\times I)}.$$

■

From Theorem 7, we can use the constant δ_0 to fix an accuracy and construct a finite-dimensional subspace W of $\mathscr{X}(D \times I)$ such that the $L^2(D \times I)$-distance between solutions to Eq. (1) and W is within the accuracy threshold. In the following lemma, we provide an upper bound on the dimensionality of W.

Lemma 8. *Let $\Omega \subset \mathbb{R}^n$ be a domain satisfying the uniform interior cone condition, $D_1 = \{x \in \mathbb{R}^n : \|x - x_0\|_\infty < d/2\}$, $D = \{x \in \mathbb{R}^n : \|x - x_0\|_\infty < (1/2 + \rho)d\}$, $I_1 = (t_0, t_0 + \beta d^2)$, and $I = (t_0, t_0 + \beta(1 + 2\rho)^2 d^2)$, for some $\beta, \rho > 0$. Assume that $\Omega \cap D_1 \neq \emptyset$ and let δ_0*

and θ be the constant introduced in Theorem 7 characterized by the uniform cone condition. Then, for any $M \geq \exp\{2(\lceil(1+2\rho)d/\delta_0\rceil + 1)\}$, there exists a subspace $W \subset \mathscr{X}(D_1 \times I_1)$ such that

$$\operatorname{dist}_{L^2(D_1\times I_1)}(u, W) \leq \frac{1}{M}\|u\|_{L^2(D\times I)}, \quad \forall u \in \mathscr{X}(D \times I),$$

and

$$\dim(W) \leq c_\rho^{n+2}\lceil \log M\rceil^{n+3} + \lceil \log M\rceil, \quad c_\rho = e(2+\rho^{-1})\kappa_c\, c_{\text{appr}}, \tag{24}$$

where $\kappa_c = \sqrt{4\Lambda^2/\lambda^2 + 1/(2\lambda\beta)}$ and c_{appr} is the constant defined in Theorem 7.

Proof Our proof follows closely the argument for elliptic PDEs (Bebendorf and Hackbusch, 2003, Lem. 2.6). Let $i \in \mathbb{N}_{\geq 1}$ and, for $0 \leq k \leq i$, define

$$D^{(k)} = \{x \in \mathbb{R}^n : \|x - x_0\|_\infty < (1/2 + (1-k/i)\rho)d\},$$
$$I^{(k)} = (t_0, t_0 + \beta(1 + 2(1-k/i)\rho)^2 d^2),$$

such that $D_1 = D^{(i)} \subset D^{(i-1)} \subset \cdots \subset D^{(0)} = D$, and $I_1 = I^{(i)} \subset I^{(i-1)} \subset \cdots \subset I^{(0)} = I$. We also denote $Q^{(k)} = D^{(k)} \times I^{(k)}$ and $\mathscr{X}^{(k)} = \mathscr{X}(Q^{(k)})$. Let $1 \leq j \leq i$. By applying Theorem 5 with the domains $K = D^{(j)}$, $D = D^{(j-1)}$, $I' = I^{(j)}$, and $I = I^{(j-1)}$, we find that

$$\|\nabla v\|_{L^2(Q^{(j)})} \leq \kappa_c \frac{i}{\rho d}\|v\|_{L^2(Q^{(j-1)})}, \quad v \in \mathscr{X}^{(j-1)}. \tag{25}$$

Moreover, Eq. (19) shows that $\mathscr{X}^{(j-1)} \subset \mathscr{X}^{(j)} \cap V_2(Q^{(j)})$. In addition, the choice of $D = D^{(j)}$ and $I = I^{(j)}$ in Theorem 7, shows that there exists a subspace $V_j \subset \mathscr{X}^{(j)}$ such that

$$\operatorname{dist}_{L^2(Q^{(j)})}(v, V_j) \leq c_{\text{appr}}\frac{(1+2\rho)d}{iB}\|\nabla v\|_{L^2(Q^{(j)})}, \quad v \in \mathscr{X}^{(j)} \cap V_2(Q^{(j)}), \tag{26}$$

where B is a constant chosen so that $\lceil(1+2\rho)d/\delta_0\rceil + 1 \leq iB \leq k^{\frac{1}{n+2}}$ and $\dim V_j \leq k$. In particular, we can set $k \coloneqq \lceil (iB)^{n+2}\rceil$ so that $k \geq (1 + \lceil(1+2\rho)d/\delta_0\rceil)^{n+2}$. Combining Eqs. (25) and (26) yields

$$\operatorname{dist}_{L^2(Q^{(j)})}(v, V_j) \leq \frac{1+2\rho}{\rho}\frac{c_{\text{appr}}\,\kappa_c}{B}\|v\|_{L^2(Q^{(j-1)})}, \quad v \in \mathscr{X}^{(j-1)}.$$

We now choose $B \coloneqq B_0 M^{1/i}$ and $B_0 \coloneqq c_{\text{appr}}\kappa_c\frac{1+2\rho}{\rho}$ to obtain the following inequality:

$$\operatorname{dist}_{L^2(Q^{(j)})}(v, V_j) \leq M^{-1/i}\|v\|_{L^2(Q^{(j-1)})}, \quad v \in \mathscr{X}^{(j-1)}. \tag{27}$$

Now let $u \in \mathscr{X}^{(0)}$. We aim to iteratively express u as a sum of functions in smaller subspaces. Initially, we define $v_0 = u$ and use Eq. (27) to decompose v_0 on $Q^{(1)}$ as $v_0|_{Q^{(1)}} = u_1 + v_1$, where $u_1 \in V_1$ and v_1 satisfies $\|v_1\|_{L^2(Q^{(1)})} \leq M^{-1/i}\|v_0\|_{L^2(Q^{(0)})}$. Consequently, we see that $v_1 \in \mathscr{X}^{(1)}$. We can continue this process for $1 \leq j \leq i$, such that $v_{j-1}|_{Q^{(j)}} = u_j + v_j$, where $u_j \in V_j$ and $v_j \in \mathscr{X}^{(j)}$ satisfies $\|v_j\|_{L^2(Q^{(j)})} \leq M^{-1/i}\|v_{j-1}\|_{L^2(Q^{(j-1)})}$. Finally, we define the subspace $W = \operatorname{span}\{V_j|_{D_1\times I_1} : 1 \leq j \leq i\}$ using the restrictions of the V_j to the smallest domain $Q^{(i)} = D^{(i)} \times I^{(i)} = D_1 \times I_1$, which contain $u_j|_{D_1\times I_1} \in V_j|_{D_1\times I_1} \subset W$ for $1 \leq j \leq i$. Therefore, the decomposition of v_0 as $v_0 = v_i + \sum_{j=1}^i u_j$ leads to

$$\operatorname{dist}_{L^2(D_1\times I_1)}(v_0, W) \leq \|v_i\|_{L^2(D_1\times I_1)} \leq (M^{-1/i})^i\|v_0\|_{L^2(Q^{(0)})} = M^{-1}\|u\|_{L^2(D\times I)}.$$

We then choose $i = \lceil \log M \rceil$ and use the definition of W to bound the dimension of W by

$$\dim(W) \leq \sum_{j=1}^{i} \dim(V_j) = i\lceil (iB)^{n+2} \rceil \leq i + B^{n+2} i^{n+3} \leq \lceil \log M \rceil + B_0^{n+2} e^{n+2} \lceil \log M \rceil^{n+3},$$

because $B = B_0 M^{1/i} \leq B_0 e$. The statement of the lemma follows by defining $c_\rho = B_0 e = c_{\text{appr}} \kappa_c e \frac{1+2\rho}{\rho}$. ∎

We are now ready to prove that the Green's function associated with a parabolic PDE has a separable approximation in terms of L^2-norm on well-separated domains.

Theorem 9. *Let $\Omega \subset \mathbb{R}^n$ be a domain satisfying the uniform interior cone condition and $\rho > 0$. Let $D_1, D_2 \subset \mathbb{R}^n$ be two domains such that D_1 is convex and $I_1, I_2 \subset (0, T)$ be two open bounded intervals, such that $Q_X = (D_1 \cap \Omega) \times I_1$ and $Q_Y = (D_2 \cap \Omega) \times I_2$ are admissible, i.e., $\operatorname{dist}(Q_X, Q_Y) \geq \rho \max\{\operatorname{diam} Q_X, \operatorname{diam} Q_Y\}$. Then, for any $\epsilon > 0$ sufficiently small, there exists a separable approximation of the form*

$$G_k(x, t, y, s) = \sum_{i=1}^{k} u_i(x, t) v_i(y, s), \quad (x, t, y, s) \in Q_X \times Q_Y,$$

where $k \leq k_\epsilon = c_{\rho/2}^{n+2} \lceil \log \frac{1}{\epsilon} \rceil^{n+3} + \lceil \log \frac{1}{\epsilon} \rceil$, and c_ρ is defined in (24), such that

$$\|G(\cdot, \cdot, y, s) - G_k(\cdot, \cdot, y, s)\|_{L^2(Q_X)} \leq \epsilon \|G(\cdot, \cdot, y, s)\|_{L^2(\hat{Q}_X)}, \quad (y, s) \in Q_Y, \tag{28}$$

where $\hat{Q}_X \coloneqq \{X \in Q, \operatorname{dist}(X, Q_X) \leq \frac{\rho}{2} \operatorname{diam} Q_X\}$.

Proof Let $\epsilon_0 = e^{-2(\lceil (1+\rho) d/\delta_0 \rceil + 1)}$ with δ_0 defined in Theorem 7, $I_1 = (t_0 - \beta d^2/2, t_0 + \beta d^2/2)$, with $t_0 > 0$ and β defined in Eq. (5), and $d = \max\{\operatorname{diam} Q_X, \operatorname{diam} Q_Y\}$. We also let $D = \{x \in \mathbb{R}^n, \operatorname{dist}(x, D_1 \cap \Omega) \leq \frac{\rho d}{2}\}$ and $I = \{t_0 - \beta d^2 (1+\rho)^2/2, t_0 + \beta d^2 (1+\rho)^2/2\}$. Similarly to Bebendorf and Hackbusch (2003), we observe that because $\operatorname{dist}(\hat{Q}_X, Q_Y) \geq \operatorname{dist}(D \times I, Q_Y) \geq \frac{\rho d}{2} > 0$, the right-hand side $\|G(\cdot, \cdot, y, s)\|_{L^2(\hat{Q}_X)}$ does not contain the singularity of G.

According to Theorem 8, with $M = \epsilon^{-1}$ and ρ replaced with $\rho/2$, we can set $\{u_1, \ldots, u_k\}$ be the basis of the subspace $W \subset \mathscr{X}(D_1 \times I_1)$ with $k = \dim W \leq c_{\rho/2}^{n+2} \lceil \log \frac{1}{\epsilon} \rceil^{n+3} + \lceil \log \frac{1}{\epsilon} \rceil$. For any $(y, s) \in Q_Y$, the function $g_Y := G(\cdot, \cdot, y, s)$ is in $\mathscr{X}(D \times I)$. Moreover, $g_Y = \hat{g}_Y + r_Y$ holds with $\hat{g}_Y \in W$ and $\|r_Y\|_{L^2(Q_X)} \leq \epsilon \|g_Y\|_{L^2(\hat{Q}_X)}$. Then, expressing $\hat{g}_Y$ with the basis, we obtain $\hat{g}_Y = \sum_{i=1}^{k} v_i(y, s) u_i$, with coefficients $v_i(y, s)$ depending on y and s. Since $(y, s) \in Q_Y$, the v_i are functions defined on Q_Y. The function $G_k(x, t, y, s) = \sum_{i=1}^{k} u_i(x, t) v_i(y, s)$ satisfies the estimate (28). ∎

If we integrate Eq. (28) over $(y, s) \in Q_Y$, we obtain the inequality stated in Section 1.3:

$$\|G - G_k\|_{L^2(Q_X \times Q_Y)} \leq \epsilon \|G\|_{L^2(\hat{Q}_X \times Q_Y)}.$$

4. Learning Rate for Green's Functions Associated with Parabolic PDEs

In this section, we combine Theorem 9 and the generalization of the randomized SVD to HS operators (Boullé and Townsend, 2022a,b) to construct a global approximant of Green's functions associated with parabolic PDEs. We suppose that one can generate $N \geq 1$ arbitrary forcing terms $\{f_1, \ldots, f_N\}$ and observe the corresponding solutions $\{u_1, \ldots, u_N\}$ from an unknown parabolic PDE of the form of Eq. (1) and its adjoint, and derive a learning rate by working out the number of training pairs needed to learn the Green's function within a prescribed tolerance. As discussed in Section 1.2, there is an additional difficulty compared with the elliptic case (Boullé and Townsend, 2022a) as Green's functions of parabolic operators are not guaranteed to be squared-integrable in spatial dimensions greater than two. Therefore, we prove the following theorem, which provides rigorous probability bounds for approximating the Green's function in the L^1-norm from a given number of forcing terms and solutions.

Theorem 10. *Let $\Omega \subset \mathbb{R}^n$ be a domain satisfying the uniform interior cone condition, $\mathcal{U} = \Omega \times [0, T]$, $\epsilon > 0$ sufficiently small, and G be the Green's function associated with the parabolic operator $\mathcal{P}$ in Eq. (1). Then, there exists a randomized algorithm that can construct an approximation $\tilde{G}$ of G using $\mathcal{O}(\epsilon^{-\frac{n+2}{2}} \log(1/\epsilon))$ many input-output pairs of (1) and its adjoint such that*

$$\|G - \tilde{G}\|_{L^1(\mathcal{U}\times\mathcal{U})} = \mathcal{O}(\Gamma_\epsilon^{-1/2}\epsilon)\, \|G\|_{L^1(\mathcal{U}\times\mathcal{U})},$$

with probability $\geq 1 - \mathcal{O}(\epsilon^{\log^{n+1}(1/\epsilon)})$. The quantity Γ_ϵ is defined in Eq. (39) and characterizes the quality of the training pairs to learn G.

The proof of the theorem occupies the rest of the section. It exploits the regularity result of Green's functions on admissible domains stated in Theorem 9, and standard Gaussian bounds near the diagonal $\mathcal{D} = \{(x, t, y, s) \in \mathcal{U} \times \mathcal{U},\, (x, t) = (y, s)\}$.

4.1 Hierarchical Partition of the Temporal Domain

We start the proof of Theorem 10 by constructing a hierarchical partition of the domain $\mathcal{U} \times \mathcal{U}$ into admissible and non-admissible domains. We aim to obtain a partition such that the vast majority of the subdomains are admissible, while the remaining non-admissible domains all have small areas. In this way, we can obtain accurate low-rank approximations on admissible domains by combining Theorem 9 with the randomized SVD (see Section 2.3), and safely neglect Green's functions on non-admissible domains.

Without loss of generality, we assume that $\mathcal{U} = \Omega \times I$, where $\Omega \subset D = [0, 1]^n$, $I = [0, 1]$, and $\beta = 1$ in Eq. (5). Otherwise it's straightforward to shift and scale Ω and $[0, T]$ by adjusting β. We construct a partition of $\mathcal{U} \times \mathcal{U}$ such that if $Q_X \times Q_Y$ is an element of the partition, where $Q_X = (\Omega \cap D_x) \times I_X$, then

$$\operatorname{diam}(D_X) = \operatorname{diam}(I_X)^{1/2} = \operatorname{diam}(Q_X) = \operatorname{diam}(Q_Y). \tag{29}$$

This condition is determined by the metric defined in Eq. (5) and guarantees that a large proportion of the domains in the partition are admissible. In this setting, choosing $\rho = 1$ in Eq. (7) is convenient for the definition of admissible sets.

First, we define a hierarchical partition of $D \times I$ for an arbitrary $n_\epsilon \geq 0$ number of partition levels using an octree-type structure. At each level of the partition, the spatial domain is divided into 2^n domains while the temporal domain is divided into 4 subdomains so that Eq. (29) is satisfied. The tree structure of the partition is defined as follows.

- At the level $L = 0$, the domain $I_{1,\dots,1} \coloneqq \underbrace{I_1 \times \ldots \times I_1}_{n \text{ times}} \times I_1 = [0,1]^n \times [0,1]$ is the root of the partitioning tree.
- At a given level $0 \leq L \leq n_\epsilon - 1$, if $I_{j_1,\dots,j_n,j_{n+1}}$ is a node of the tree, then it has 4×2^n children of the form

$$I_{2j_1+k_1,\dots,2j_n+k_n,4j_{n+1}+k_{n+1}}, \quad k_1, \dots k_n \in \{0,1\},\, k_{n+1} \in \{0, \dots 3\}.$$

  Here, if $I_{j_1} = [a,b] \subset [0,1]$, then $I_{2j_1} = [a,(a+b)/2]$ and $I_{2j_1+1} = [(a+b)/2,b]$. The division of the temporal interval $I_{j_{n+1}}$ into four subintervals is performed similarly.

The tree structure of the hierarchical partition in spatial dimension $n = 1$ is displayed in Fig. 2 along with examples of admissible subdomains.

Using the partition of $D \times I$, we can define a tree structure for $(D \times I) \times (D \times I)$ and cluster the tree nodes into admissible and non-admissible sets, respectively denoted by P_{adm} and $P_{\text{non-adm}}$. These two sets also allow us to design a hierarchical partition of the domain $(D \times I) \times (D \times I)$.

- At the level $L = 0$, the root of the tree is given by the domain $I_{1,\dots 1} \times I_{1,\dots 1}$, which belongs to the non-admissible set as it does not satisfy Eq. (7).
- At a given level $0 < L \leq n_\epsilon - 1$, if $I_{j_1,\dots,j_{n+1}} \times I_{\tilde{j}_1,\dots,\tilde{j}_{n+1}}$ is a node of the tree, then it is either in the non-admissible set if all the respective indices are separated by at most one, i.e., $|j_1 - \tilde{j}_1| \leq 1, \dots, |j_{n+1} - \tilde{j}_{n+1}| \leq 1$, or labeled as admissible otherwise. If the node is admissible then it is added to the hierarchical partition. Otherwise, we subdivide it into $(4 \times 2^n)^2$ children using cross-products of the 4×2^n children of $I_{j_1,\dots,j_{n+1}}$ and $I_{\tilde{j}_1,\dots,\tilde{j}_{n+1}}$ in the partition of $D \times [0,1]$.
- At the final level $L = n_\epsilon$, we add both the admissible and non-admissible domains to the partition.

In Fig. 3, we illustrate slices of the spatial and temporal partition of the domain $(D \times I) \times (D \times I)$ when $D = [0,1]$. The regions coloured in green are admissible while the ones coloured in red are non-admissible. In addition, the grey area in Fig. 3(b) shows the domain on which the Green's function is zero. At the final level n_ϵ, all the non-admissible domains have the same diameter and, if $Q_X \times Q_Y$ is non-admissible with $Q_X = D_X \times I_X$, we have $\text{diam}(Q_X) = \text{diam}(D_X) = \text{diam}(I_X)^{1/2} = 2^{-n_\epsilon}$. Therefore, the width of the non-admissible temporal region (see Fig. 3(b)) is given by

$$r_{n_\epsilon} = \sqrt{2} \times 2^{-2n_\epsilon}. \tag{30}$$

We now compute an upper bound on the number of admissible domains in the partition. By construction, the number of non-admissible regions in the hierarchical partition satisfies

$$|P_{\text{non-adm}}| = (3 \times 4^{n_\epsilon} - 2) \times (3 \times 2^{n_\epsilon} - 2)^n. \tag{31}$$

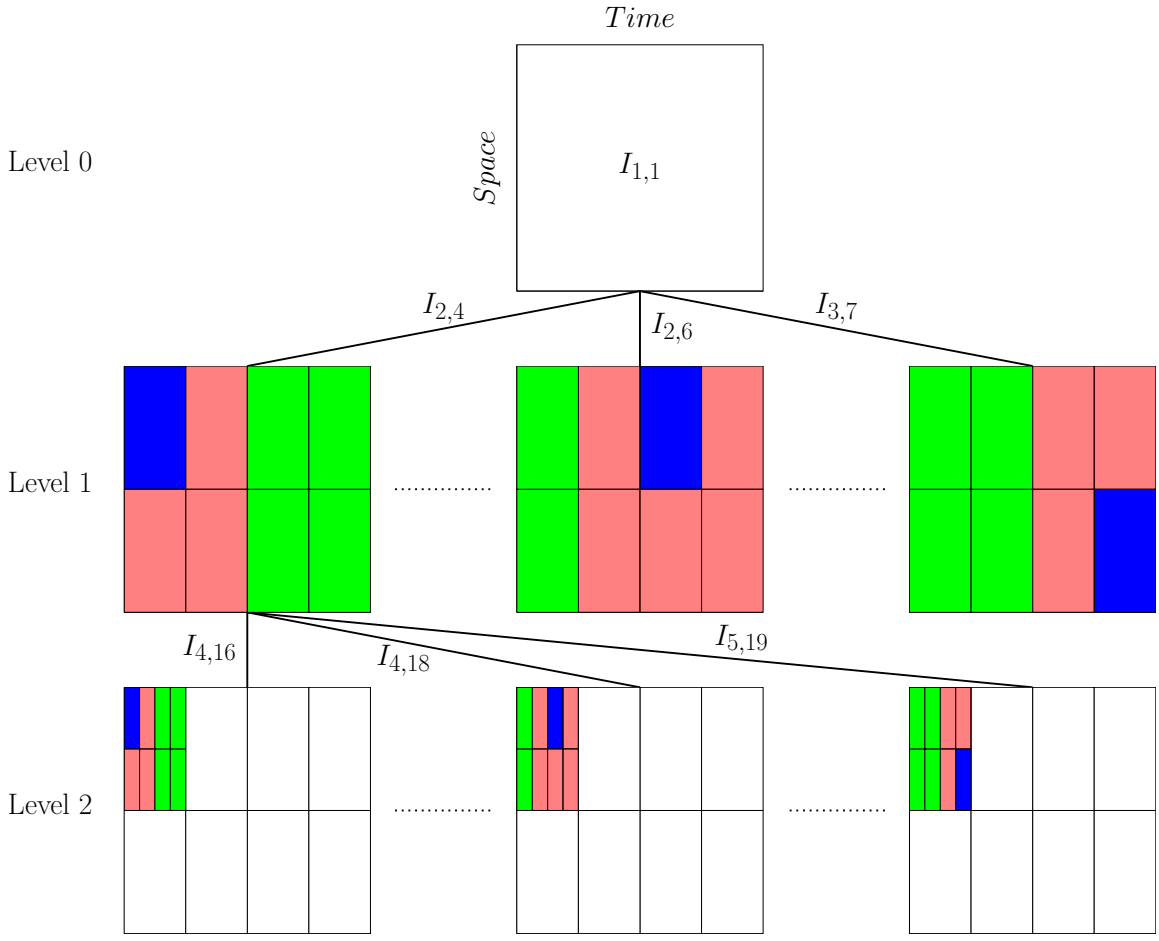

Figure 2: Hierarchical partition of the domain $[0,1] \times [0,1]$, where the spatial domains are divided into 2 at each level, and the temporal domains are divided into 4. At the levels 1 and 2 of the tree, the domains coloured in red (resp. green) are non-admissible (resp. admissible) for the blue domain.

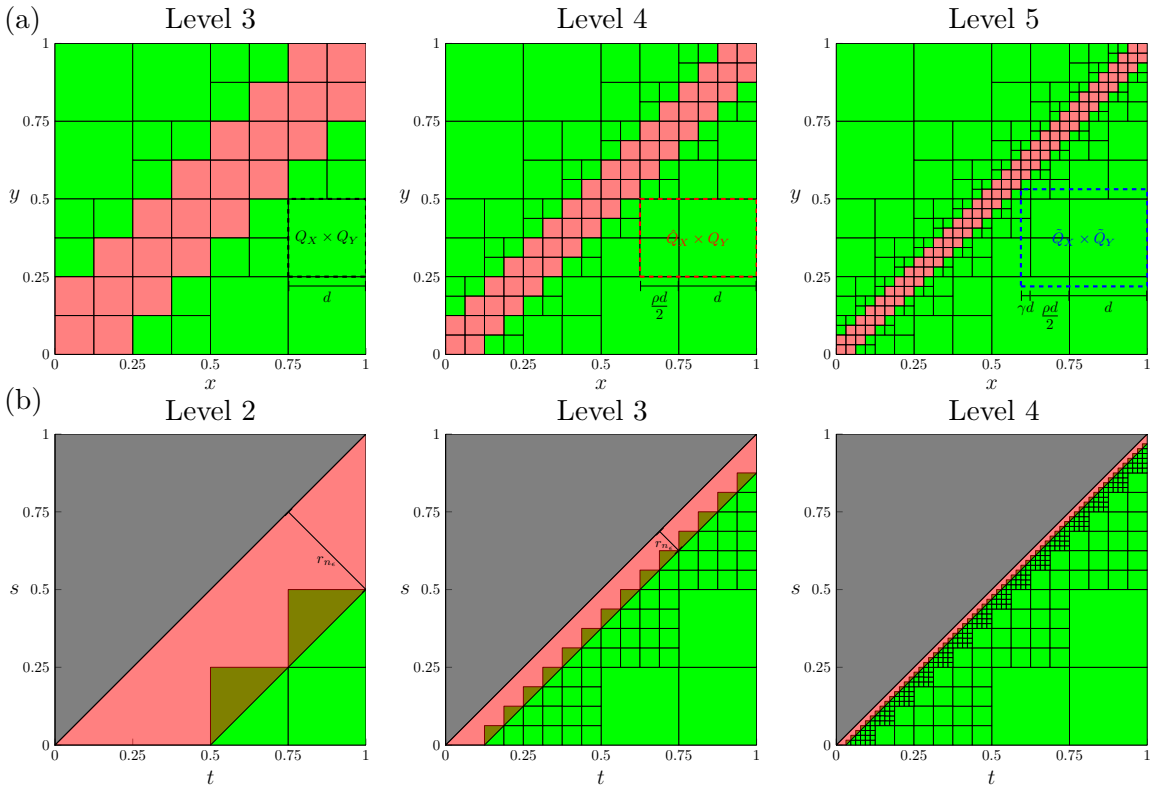

Figure 3: (a) Illustration of level $3, 4$, and 5 of the hierarchical partition of the spatial domain $[0,1] \times [0,1]$. The green blocks are admissible domains with $\rho = 1$ while the pink domains are non-admissible. The left to right panels display one admissible domain $Q_X \times Q_Y$ along with the elongated domains $\hat{Q}_X \times Q_Y$ (dashed red rectangle) and $\tilde{Q}_X \times Q_Y$ (dashed blue rectangle) appearing in the bounds of Section 4.4. (b) Illustration of the hierarchical partition of the temporal domain. At each level the non-admissible blocks are separated into 4^2 domains. For partition level n_ϵ, r_{n_ϵ} is the width of the non-admissible region.

Moreover, the number of new admissible domains, $|P_{\text{adm}}(L)|$, added to the partition at a given level $1 \leq L \leq n_\epsilon$ is given by the number of children of the non-admissible domains at the previous level minus the number of non-admissible sets at the current level, i.e.,

$$|P_{\text{adm}}(L)| = (4 \times 2^n)^2(3 \times 4^{L-1} - 2)(3 \times 2^{L-1} - 2)^n - (3 \times 4^L - 2)(3 \times 2^L - 2)^n, \quad (32)$$

where the number of non-admissible sets is given by Eq. (31). We can then compute the total number of admissible domains, $|P_{\text{adm}}|$ in the partition by summing Eq. (32) over $1 \leq L \leq n_\epsilon$ and obtain

$$\begin{aligned}|P_{\text{adm}}| &= \sum_{L=1}^{n_\epsilon} 2^{2n+4}(3 \times 4^{L-1} - 2)(3 \times 2^{L-1} - 2)^n - (3 \times 4^L - 2)(3 \times 2^L - 2)^n \\ &= 1 - (3 \times 4^{n_\epsilon} - 2)(3 \times 2^{n_\epsilon} - 2)^n + (2^{2n+4} - 1) \sum_{L=0}^{n_\epsilon - 1} (3 \times 4^L - 2)(3 \times 2^L - 2)^n.\end{aligned}$$

We can bound $|P_{\text{adm}}|$ by computing the sum of a geometric series as

$$\sum_{L=0}^{n_\epsilon - 1} (3 \times 4^L - 2) \times (3 \times 2^L - 2)^n \leq 3^{n+1} \sum_{L=0}^{n_\epsilon - 1} (2^{n+2})^L = 3^{n+1} \frac{2^{(n+2)n_\epsilon} - 1}{2^{n+2} - 1} \leq \left(\frac{3}{2}\right)^{n+1} 2^{(n+2)n_\epsilon}.$$

The number of admissible domains is bounded by

$$|P_{\text{adm}}| \leq 2^{2n+4} 3^{n+1} 2^{-(n+1)} 2^{(n+2)n_\epsilon} = 24 \times 6^n 2^{(n+2)n_\epsilon}. \quad (33)$$

We conclude the construction of the hierarchical partition of $\mathcal{U} \times \mathcal{U}$ by intersecting each element of the partition with $(\Omega \times I) \times (\Omega \times I)$. In the following section, we assume that we have already constructed a partition of $\mathcal{U} \times \mathcal{U}$ for a general bounded domain $\Omega \subset \mathbb{R}^n$ and $I = [0, T]$. In fact, the number of admissible domains and size of the non-admissible region (cf. Eqs. (30) and (33)) remain the same up to a constant that depends on the size of the domain $\mathcal{U}$.

4.2 Diagonal Estimate of Green's Functions

This section determines the number of hierarchical partitioning levels needed to neglect the Green's function on the non-admissible regions of the partition of $\mathcal{U} \times \mathcal{U}$ defined in Section 4.1. To start, we use a global Gaussian estimate for Green's functions associated with parabolic PDEs, which guarantees the existence of a positive constant $C > 0$ such that the Green's function is positive and bounded by (Cho et al., 2012, Eq. 4.2)

$$G(x, t, y, s) \leq C \frac{\Theta(t - s)}{(t - s)^{n/2}} \exp\left(-\frac{\kappa |x - y|^2}{t - s}\right), \quad (x, t) \neq (y, s) \in \mathcal{U}, \quad (34)$$

where $\kappa > 0$ is a constant depending on the uniform parabolicity constants (3) and is independent of T. We remark that the Green's function is zero if $t \leq s$, as can be seen in the gray regions of Fig. 3(b). For a given diameter $0 < r_t \leq T$, we define a diagonal subdomain of $\mathcal{U} \times \mathcal{U}$ as

$$\mathcal{D}_{r_t} \coloneqq \{(x, t, y, s) \in \mathcal{U} \times \mathcal{U},\ |t - s| < r_t\}.$$

We use the estimate (34) to bound the Green's function on the domain $\mathcal{D}_{r_t}$ in the L^p-norm, for $p \geq 1$.

Proposition 11. *Let $p \geq 1$, $0 < r_t \leq T$, and assume that $n(p-1) < 2$. Then, there exists a constant $C_{\text{diag}} = C_{\text{diag}}(\Omega, T, p, \kappa) > 0$ such that*

$$\|G\|_{L^p(\mathcal{D}_{r_t})} \leq C_{\text{diag}} r_t^{[1-n(p-1)/2]/p} \|G\|_{L^p(\mathcal{U}\times\mathcal{U})}.$$

Proof Let $y \in \Omega$, $s \in (0,T)$, and denote $I_s = (s, \min(s+r_t, T))$. Integrating Eq. (34), raised to the power p, on the domain $\Omega \times I_s \subset \mathcal{U}$ yields the following inequality,

$$\begin{aligned}\int_{I_s}\int_{\Omega} |G(x,t,y,s)|^p \,\mathrm{d}x\,\mathrm{d}t &\leq \int_{I_s}\int_{\Omega} \frac{(\Theta(t-s))^p C^p}{(t-s)^{pn/2}} \exp\left\{-p\frac{\kappa|x-y|^2}{t-s}\right\} \mathrm{d}x\,\mathrm{d}t \\ &\leq C^p \int_0^{r_t} \tilde{t}^{-pn/2} \int_{\mathbb{R}^n} e^{-p\kappa|x|^2/\tilde{t}} \,\mathrm{d}x\,\mathrm{d}\tilde{t},\end{aligned}$$

where we use the change of variables $\tilde{t} = t - s$. We then make the change of variables $\tilde{x} = x\sqrt{p\kappa/\tilde{t}}$ to obtain

$$\begin{aligned}\|G(\cdot,\cdot,y,s)\|^p_{L^p(\Omega\times I_s)} &\leq C^p \int_0^{r_t} \tilde{t}^{-pn/2} \left(\frac{\tilde{t}}{p}\right)^{n/2} \int_{\mathbb{R}^n} e^{-|\tilde{x}|^2} \,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{t} \leq C^p \left(\frac{\pi}{p}\right)^{n/2} \int_0^{r_t} \tilde{t}^{-n(p-1)/2} \,\mathrm{d}\tilde{t} \\ &\leq \left(\frac{\pi}{p}\right)^{n/2} \frac{C^p}{1-n(p-1)/2} r_t^{1-n(p-1)/2}.\end{aligned}$$

Integrating this expression over $y \in \Omega$ and $s \in (0,T)$ yields

$$\|G\|^p_{L^p(\mathcal{D}_{r_t})} \leq \left(\frac{\pi}{p}\right)^{n/2} \frac{|\Omega| T C^p}{1-n(p-1)/2} r_t^{1-n(p-1)/2}.$$

If $n(p-1) < 2$, then the Green's function is in $L^p(\mathcal{U}\times\mathcal{U})$, and we can introduce a constant $C_{\text{diag}} = C_{\text{diag}}(\Omega, T, p, \kappa) > 0$, such that

$$\|G\|_{L^p(\mathcal{D}_{r_t})} \leq C_{\text{diag}} r_t^{[1-n(p-1)/2]/p} \|G\|_{L^p(\mathcal{U}\times\mathcal{U})},$$

which concludes the proof. ∎

We conclude that the Green's function restricted to the domain $\mathcal{D}_{r_t}$ has a relative small norm if r_t is small. Applying Theorem 11 with the parameter $p = 1$ gives the following L^1 estimate:

$$\|G\|_{L^1(\mathcal{D}_{r_t})} \leq C_{\text{diag}} r_t \|G\|_{L^1(\mathcal{U}\times\mathcal{U})}. \tag{35}$$

Due to the hierarchical partition of the temporal domain, we can easily bound the norm of G on non-admissible sets by using a temporal radius of $r_t = r_{n_\epsilon} = \sqrt{2} \times 2^{-2n_\epsilon}$ up to a constant depending on the size of the domain $\mathcal{U}$, where n_ϵ is the number of hierarchical levels (see. Eq. (30)). Then, since $P_{\text{non-adm}} \subset \mathcal{D}_{r_t}$ as illustrated by Fig. 3, Eq. (35) yields

$$\|G\|_{L^1(P_{\text{non-adm}})} \leq \sqrt{2} C_{\text{diag}} 2^{-2n_\epsilon} \|G\|_{L^1(\mathcal{U}\times\mathcal{U})} \leq \epsilon \|G\|_{L^1(\mathcal{U}\times\mathcal{U})}, \tag{36}$$

where we choose $n_\epsilon \sim (1/2)\log_2(1/\epsilon)$ hierarchical levels such that $\sqrt{2} C_{\text{diag}} 2^{-2n_\epsilon} \leq \epsilon$. This means that we can safely approximate G with the zero approximant on non-admissible domains, and still get an approximation of G within a relative accuracy of ϵ in the L^1 norm.

One might be able to slightly improve Theorem 11 by computing the integral of each non-admissible domain, similarly to the elliptic case (Boullé and Townsend, 2022a). However, the gain is expected to be marginal since the decay of the bound in Eq. (34) is essentially controlled by a well-separation of the temporal variables t and s.

4.3 Approximating Green's Functions on Admissible Domains

The approximation of Green's functions on the well-separated domains of the partition of $\mathcal{U} \times \mathcal{U}$ is achieved using the randomized SVD for HS operators described in Section 2.3. Let $Q_X \times Q_Y \in P_{\text{adm}}$ be an admissible domain and $k = k_\epsilon = c^3_{\rho/2}\lceil \log \frac{1}{\epsilon} \rceil^{n+3} + \lceil \log \frac{1}{\epsilon} \rceil$ be a target rank derived in Theorem 9, we can combine Theorem 9 and the Eckart–Young–Mirsky theorem (Hsing and Eubank, 2015, Thm. 4.4.7), which characterizes the best rank-k approximation error to a HS operator, to bound the singular values of the Green's function restricted to $Q_X \times Q_Y$ by

$$\left(\sum_{j=k_\epsilon+1}^{\infty} \sigma^2_{j,Q_X \times Q_Y} \right)^{1/2} \leq \|G - G_{k_\epsilon}\|_{L^2(Q_X \times Q_Y)} \leq \epsilon \|G\|_{L^2(\hat{Q}_X \times Q_Y)}, \tag{37}$$

where $\sigma_{j,Q_X \times Q_Y}$ are the singular values of G restricted to $Q_X \times Q_Y$. We conclude that the singular values of G decay rapidly to 0 on admissible domains.

With the rapidly decaying singular values, we can follow the arguments in (Boullé and Townsend, 2022a, Sec. 4.1.2) to use the randomized SVD for learning Green's functions on admissible sets. Roughly speaking, we start with a Gaussian process on $\mathcal{U}$ and define a covariance kernel K that restricts onto $Q_Y \times Q_Y$. We then extend the restricted operator by 0 on $\mathcal{U} \times \mathcal{U}$ and apply the randomized SVD. As a result, with a target rank of $k_\epsilon = c^3_{\rho/2}\lceil \log \frac{1}{\epsilon} \rceil^{n+3} + \lceil \log \frac{1}{\epsilon} \rceil$, an oversampling parameter $p = k_\epsilon$, and $t = e$, we combine Theorem 9 and Eq. (37) to obtain an approximant $\tilde{G}_{X \times Y}$ of G on $Q_X \times Q_Y$ such that

$$\|G - \tilde{G}_{X \times Y}\|_{L^2(Q_X \times Q_Y)} \leq \left(1 + se \sqrt{\frac{6k_\epsilon}{\gamma_{k_\epsilon, Q_X \times Q_Y}} \frac{\operatorname{Tr}(K)}{\lambda_1}} \right) \epsilon \|G\|_{L^2(\hat{Q}_X \times Q_Y)}, \tag{38}$$

which holds with probability greater than $1 - e^{-k_\epsilon} - e^{-k_\epsilon(s^2 - 2\log(s) - 1)} \geq 1 - 2e^{-k_\epsilon}$ when $s \geq 3$. The factor $\gamma_{k_\epsilon, Q_X \times Q_Y}$ characterizes the suitability of the covariance kernel for learning G on the domain $Q_X \times Q_Y$. In this way, our algorithm requires $N_{\epsilon, X \times Y} = 2(k_\epsilon + p) = \mathcal{O}\left(\log^{n+3}(1/\epsilon)\right)$ input-output pairs to learn an approximant to G on $Q_X \times Q_Y$.

Remark 12. *To apply the projection operator associated with the randomized approximation on the left of the HS operator in Eq.* (8)*, we need to solve the adjoint equation associated with Eq.* (1)*, which is allowed by Assumption 1.*

4.4 Recovering the Green's Function on the Entire Domain

We can now recover the Green's function G on the entire domain $\mathcal{U} \times \mathcal{U}$ and compute the number of input-output pairs needed to approximate it within accuracy $\epsilon > 0$. With n_ϵ computed in Section 4.2, we can follow the arguments in (Boullé and Townsend, 2022a, Sec. 4.4.1) to quantify the total number of input-output pairs we need to approximate G

using the randomized SVD described in Section 4.3. In particular, we denote the worst $\gamma_{k_\epsilon, Q_X \times Q_Y}$ by

$$\Gamma_\epsilon = \min\{\gamma_{k_\epsilon, Q_X \times Q_Y}, Q_X \times Q_Y \in P_{\text{adm}}\}, \tag{39}$$

so that we need

$$N_\epsilon = \mathcal{O}(|P_{\text{adm}}| \log^{n+3}(1/\epsilon)) = \mathcal{O}(\epsilon^{-\frac{n+2}{2}} \log^{n+3}(1/\epsilon))$$

input-output pairs to capture the Green's function on admissible domains with $n_\epsilon \sim (1/2)\log_2(1/\epsilon)$ hierarchical levels (see Section 4.2), and the number of admissible domains given by Eq. (33).

We now provide an explicit bound for the approximation $\tilde{G}$ if we use zero approximant on non-admissible sets and learn with $N_\epsilon = \mathcal{O}(\epsilon^{-\frac{n+2}{2}} \log^{n+3}(1/\epsilon))$ many input-output pairs on admissible domains. First, we separate the norm error into error on admissible sets and that on non-admissible sets as

$$\begin{aligned} \|G - \tilde{G}\|_{L^1(\mathcal{U}\times\mathcal{U})} &\leq \|G - \tilde{G}\|_{L^1(P_{\text{non-adm}})} + \|G - \tilde{G}\|_{L^1(P_{\text{adm}})} \\ &\leq \epsilon \|G - \tilde{G}\|_{L^1(\mathcal{U}\times\mathcal{U})} + \sum_{Q_X \times Q_Y \in P_{\text{adm}}} \|G - \tilde{G}\|_{L^1(Q_X \times Q_Y)}, \end{aligned} \tag{40}$$

where the second inequality comes from Eq. (36). Let $Q_X \times Q_Y \in P_{\text{adm}}$, we focus on bounding the error on this subdomain with Eq. (38). Using Hölder's inequality, we have

$$\|G(\cdot,\cdot,y,s) - G_k(\cdot,\cdot,y,s)\|_{L^1(Q_X)} \leq |Q_X|^{1/2} \|G(\cdot,\cdot,y,s) - G_k(\cdot,\cdot,y,s)\|_{L^2(Q_X)}, \quad (y,s) \in Q_Y,$$

which implies that

$$\|G - \tilde{G}\|_{L^1(Q_X \times Q_Y)} \leq |Q_X|^{1/2} |Q_Y|^{1/2} \|G - \tilde{G}\|_{L^2(Q_X \times Q_Y)}. \tag{41}$$

We then apply Eq. (38) to estimate the term $\|G - \tilde{G}\|_{L^2(Q_X \times Q_Y)}$ in Eq. (41) and bound the resulting right-hand side term $\|G\|_{L^2(\hat{Q}_X \times Q_Y)}$ by an L^1-estimate to complete the bound of the approximation error on admissible domains. Let $\gamma > 0$ be an arbitrary constant and define $\tilde{Q}_X \coloneqq \{X \in \mathcal{U}, \operatorname{dist}(X, \hat{Q}_X) \leq \frac{\gamma}{2} \operatorname{diam} \hat{Q}_X\}$. We first remark that for $(y,s) \in Q_Y$, $G(\cdot,\cdot,y,s)$ satisfies $\mathcal{P}G(\cdot,\cdot,y,s) = 0$ in $(\tilde{D} \cap \Omega) \times \tilde{I}$ and vanishes on $(\tilde{D} \cap \partial\Omega) \times \tilde{I}$, where $\tilde{Q}_X = (\tilde{D} \cap \Omega) \times \tilde{I}$. Therefore, by Moser's local maximum estimate (Lieberman, 1996, Thm. 6.30), we have

$$\|G(\cdot,\cdot,y,s)\|_{L^\infty(\hat{Q}_X)} \leq C_1 |\tilde{Q}_X|^{-1} \|G(\cdot,\cdot,y,s)\|_{L^1(\tilde{Q}_X)}, \tag{42}$$

where $C_1 = C_1(n, \lambda, \Lambda, \gamma) > 0$. Eq. (42) implies that for all $(y,s) \in Q_Y$, we have

$$\int_{\hat{Q}_X} |G(x,t,y,s)|^2 \,\mathrm{d}x \,\mathrm{d}t \leq C_1^2 |\tilde{Q}_X|^{-2} |\hat{Q}_X| \left(\int_{\tilde{Q}_X} |G(x,t,y,s)| \,\mathrm{d}x \,\mathrm{d}t \right)^2. \tag{43}$$

By integrating over $(y,s) \in Q_Y$ and using an integral version of Minkowski's inequality, we obtain

$$\begin{aligned} \int_{Q_Y} \int_{\hat{Q}_X} |G(x,t,y,s)|^2 \,\mathrm{d}x \,\mathrm{d}t \,\mathrm{d}y \,\mathrm{d}s &\leq C_1^2 |\tilde{Q}_X|^{-2} |\hat{Q}_X| \int_{Q_Y} \left(\int_{\tilde{Q}_X} |G(x,t,y,s)| \,\mathrm{d}x \,\mathrm{d}t \right)^2 \mathrm{d}y \,\mathrm{d}s \\ &\leq C_1^2 |\tilde{Q}_X|^{-2} |\hat{Q}_X| \left\{ \int_{\tilde{Q}_X} \left(\int_{Q_Y} |G(x,t,y,s)|^2 \,\mathrm{d}y \,\mathrm{d}s \right)^{1/2} \mathrm{d}x \,\mathrm{d}t \right\}^2. \end{aligned} \tag{44}$$

On the other hand, for $(x,t) \in \tilde{Q}_X$, $G(x,t,\cdot,\cdot)$ satisfies $\mathcal{P}^* G(x,t,\cdot,\cdot) = 0$ in Q_Y, where $\mathcal{P}^*$ is the adjoint operator of $\mathcal{P}$ (Cho et al., 2008). Similarly to Eqs. (42) and (43), we have

$$\int_{Q_Y} |G(x,t,y,s)|^2 \,\mathrm{d}y\,\mathrm{d}s \leq C_2^2 |\tilde{Q}_Y|^{-2} |Q_Y| \left(\int_{\tilde{Q}_Y} |G(x,t,y,s)| \,\mathrm{d}y\,\mathrm{d}s \right)^2, \tag{45}$$

where $\tilde{Q}_Y \coloneqq \{Y \in \mathcal{U},\, \mathrm{dist}(Y, Q_Y) \leq \frac{\gamma}{2} \operatorname{diam} Q_Y\}$ and $C_2 > 0$ is a constant. Combining (44) and (45) yields

$$\int_{Q_Y} \int_{\hat{Q}_X} |G(x,t,y,s)|^2 \,\mathrm{d}x\,\mathrm{d}t\,\mathrm{d}y\,\mathrm{d}s \leq C_1^2 C_2^2 \frac{|\hat{Q}_X||Q_Y|}{|\tilde{Q}_X|^2 |\tilde{Q}_Y|^2} \left\{ \int_{\tilde{Q}_X} \int_{\tilde{Q}_Y} |G(x,t,y,s)| \,\mathrm{d}y\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t \right\}^2,$$

or equivalently,

$$\|G\|_{L^2(\hat{Q}_X \times Q_Y)} \leq C_1 C_2 |\tilde{Q}_X|^{-1} |\hat{Q}_X|^{1/2} |\tilde{Q}_Y|^{-1} |Q_Y|^{1/2} \|G\|_{L^1(\tilde{Q}_X \times \tilde{Q}_Y)}. \tag{46}$$

Finally, the multiplication of Eq. (46) by the term $|Q_X|^{1/2}|Q_Y|^{1/2}$ from Eq. (41) yields

$$|Q_X|^{1/2} |Q_Y|^{1/2} \|G\|_{L^2(\hat{Q}_X \times Q_Y)} \leq C \|G\|_{L^1(\tilde{Q}_X \times \tilde{Q}_Y)},$$

where $C > 0$ is a constant, because $Q_X \subset \hat{Q}_X \subset \tilde{Q}_X$ and $Q_Y \subset \tilde{Q}_Y$.

We then choose $\gamma = 1/16$ so that the domain $\tilde{Q}_X \times \tilde{Q}_Y$ is included in a finite number, C_n, of neighbors in the hierarchical partition of the domain including itself (see Fig. 3(a)). Combining Eqs. (38) and (40) and the L^1 argument described in the previous paragraph, we conclude that

$$\begin{aligned} \|G - \tilde{G}\|_{L^1(\mathcal{U}\times\mathcal{U})} &\leq \epsilon \|G\|_{L^1(\mathcal{U}\times\mathcal{U})} + \sum_{Q_X \times Q_Y \in P_{\mathrm{adm}}} \|G - \tilde{G}\|_{L^1(Q_X \times Q_Y)} \\ &\leq \epsilon \|G\|_{L^1(\mathcal{U}\times\mathcal{U})} + \sum_{Q_X \times Q_Y \in P_{\mathrm{adm}}} s C_{\mathrm{svd}} k_\epsilon^{1/2} \Gamma_\epsilon^{-1/2} \epsilon \|G\|_{L^1(\tilde{Q}_X \times \tilde{Q}_Y)} \\ &\leq s(2C_n + 1) C_{\mathrm{svd}} k_\epsilon^{1/2} \Gamma_\epsilon^{-1/2} \epsilon \|G\|_{L^1(\mathcal{U}\times\mathcal{U})}, \end{aligned} \tag{47}$$

where $C_{\mathrm{svd}} > 0$ is a constant.

Finally, we choose $s = 3$ to obtain a probability of failure of the randomized SVD less than $2e^{-k_\epsilon}$ on each admissible domain (cf. Section 4.3). Following Eq. (47), as $\epsilon \to 0$, the global approximation error between G and the constructed approximant $\tilde{G}$ on $\mathcal{U}\times\mathcal{U}$ satisfies

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

with probability $\geq (1 - 2e^{-k_\epsilon})^{24^{n+1} \times 2^{(n+2)n_\epsilon}} = 1 - \mathcal{O}(\epsilon^{\log^{n+2}(1/\epsilon) - \frac{n+2}{2}})$. This indicates that $\tilde{G}$ is a good approximation of G with high probability. We then make the change of variable $\tilde{\epsilon} \coloneqq \epsilon \log^{(n+3)/2}(1/\epsilon)$ to obtain the bound

$$\|G - \tilde{G}\|_{L^1(\mathcal{U}\times\mathcal{U})} = \mathcal{O}(\Gamma_{\tilde{\epsilon}}^{-1/2} \tilde{\epsilon}) \, \|G\|_{L^1(\mathcal{U}\times\mathcal{U})},$$

with probability $\geq 1 - \mathcal{O}(\tilde{\epsilon}^{\log^{n+1}(1/\tilde{\epsilon})})$ for ϵ small enough. Note that the factor $\Gamma_{\tilde{\epsilon}}$ is changed according to its implicit definition given by Eq. (39). As a result, the number of input-output pairs is given by

$$N_{\tilde{\epsilon}} = \mathcal{O}(\tilde{\epsilon}^{-\frac{n+2}{2}} \log^{(n+3)(1-\frac{n+2}{4})}(1/\tilde{\epsilon})) = \mathcal{O}(\tilde{\epsilon}^{-\frac{n+2}{2}} \log(1/\tilde{\epsilon})),$$

and we can drop the tilde symbol to conclude the proof of Theorem 10.

5. Summary and Discussion

We derive a rigorous learning rate for parabolic operators, by giving an upper bound on the number of training data needed to learn Green's functions within a prescribed relative accuracy. Our analysis relies on an extension of a result from Bebendorf and Hackbusch (2003) to show that Green's functions of parabolic operators admit low-rank properties on well-separated domains, i.e., away from the singularity near the diagonal. A similar low rank property is derived for elliptic operators in dimension three. This result may motivate the development of novel algorithms that use the hierarchical structures of parabolic operators to discretize time-dependent equations. One interesting outcome of this work is that the analysis and the resulting approximation error bounds are obtained using the L^1-norm since Green's functions of parabolic operators in spatial dimension greater than one are usually not square-integrable. This fact may partially explain the challenges met by the current deep learning techniques that attempt to learn the solution operators of time-dependent mathematical models using a quadratic loss function. Hence, Krishnapriyan et al. (2021) and Wang et al. (2022) recently observed and analysed mode failure issues in existing physics-informed neural network architectures. The development of PDE learning techniques based on the L^1 loss and NN architectures exploiting singularities of the underlying model, such as rational NNs (Boullé et al., 2020), is of significant interest for the field to overcome the challenges resulting from learning PDEs with short-lived transient dynamics. Finally, we note that the analysis performed in this paper is applicable to obtain a learning rate for elliptic PDEs in any spatial dimension in L^1-norm. This generalizes the previous results from Boullé and Townsend (2022a), and concludes the study of elliptic and parabolic PDEs. However, Green's functions associated with hyperbolic PDEs do not admit a similar low-rank structure on well-separated domains due to singularity near characteristics lines. Thus, the theoretical analysis remains a future challenge.

Acknowledgments

N.B. was supported by the EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling through grant EP/L015803/1 in collaboration with Simula Research Laboratory. S.K. was supported by National Research Foundation of Korea grants NRF-2019R1A2C2002724 and NRF-2022R1A2C1003322. T.S. and A.T. were supported by the National Science Foundation grants DMS-1818757, DMS-1952757, and DMS-2045646.

References

M. Alber et al. Integrating machine learning and multiscale modeling–perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *NPJ Digit. Med.*, 2(1):1–11, 2019.

J. Ballani and D. Kressner. Matrices with hierarchical low-rank structures. In *Exploiting hidden structure in matrix computations: algorithms and applications*, pages 161–209. Springer, 2016.

M. Bebendorf. *Hierarchical Matrices: A Means to Efficiently Solve Elliptic Boundary Value Problems*. Springer, 1st edition, 2008.

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(1):1–28, 2003.

M. Bebendorf and S. Rjasanow. Adaptive low-rank approximation of collocation matrices. *Computing*, 70(1):1–24, 2003.

S. Börm, L. Grasedyck, and W. Hackbusch. Introduction to hierarchical matrices with applications. *Eng. Anal. Bound. Elem.*, 27(5):405–422, 2003.

N. Boullé and A. Townsend. Learning elliptic partial differential equations with randomized linear algebra. *Found. Comput. Math.*, pages 1–31, 2022a.

N. Boullé and A. Townsend. A generalization of the randomized singular value decomposition. In *International Conference on Learning Representations*, 2022b.

N. Boullé, Y. Nakatsukasa, and A. Townsend. Rational neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 14243–14253, 2020.

N. Boullé, C. J. Earls, and A. Townsend. Data-driven discovery of Green's functions with human-understandable deep learning. *Sci. Rep.*, 12(1):4824, 2022.

S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA*, 113 (15):3932–3937, 2016.

C.-T. Chen and G. X. Gu. Learning hidden elasticity with deep neural networks. *Proc. Natl. Acad. Sci. USA*, 118(31), 2021.

T. Chen and H. Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Trans. Neur. Netw.*, 6(4):911–917, 1995.

S. Cho, H. Dong, and S. Kim. On the Green's matrices of strongly parabolic systems of second order. *Indiana Univ. Math. J.*, 57(4):1633–1677, 2008.

S. Cho, H. Dong, and S. Kim. Global estimates for Green's matrix of second order parabolic systems with application to elliptic systems in two dimensional domains. *Potential Anal.*, 36(2):339–372, 2012.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.*, 2(4):303–314, 1989.

M. V. de Hoop, N. B. Kovachki, N. H. Nelsen, and A. M. Stuart. Convergence rates for learning linear operators from noisy data. *arXiv preprint arXiv:2108.12515*, 2021.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

L. C. Evans. *Partial Differential Equations*. American Mathematical Society, 2010.

J. Feliu-Faba, Y. Fan, and L. Ying. Meta-learning pseudo-differential operators with deep neural networks. *J. Comput. Phys.*, 408:109309, 2020.

D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer, 2nd edition, 2001.

N. Gillis and S. A. Vavasis. On the complexity of robust PCA and ℓ_1-norm low-rank matrix approximation. *Math. Oper. Res.*, 43(4):1072–1084, 2018.

C. R. Gin, D. E. Shea, S. L. Brunton, and J. N. Kutz. DeepGreen: deep learning of Green's functions for nonlinear boundary value problems. *Sci. Rep.*, 11(1):1–14, 2021.

L. Greengard and P. Lin. Spectral approximation of the free-space heat kernel. *Appl. Comput. Harmon. Anal.*, 9(1):83–97, 2000.

L. Greengard and J. Strain. A fast algorithm for the evaluation of heat potentials. *Commun. Pure Appl. Math.*, 43(8):949–963, 1990.

G. Gupta, X. Xiao, and P. Bogdan. Multiwavelet-based Operator Learning for Differential Equations. In *Advances in Neural Information Processing Systems*, volume 34, pages 24048–24062, 2021.

W. Hackbusch. A sparse matrix arithmetic based on $\mathcal{H}$-matrices. I. Introduction to $\mathcal{H}$-matrices. *Computing*, 62(2):89–108, 1999.

W. Hackbusch. *Hierarchical Matrices: Algorithms and Analysis*. Springer, 2015.

W. Hackbusch and B. N. Khoromskij. A sparse $\mathcal{H}$-matrix arithmetic. II. Application to multi-dimensional problems. *Computing*, 64(1):21–47, 2000.

W. Hackbusch, B. N. Khoromskij, and R. Kriemann. Hierarchical matrices based on a weak admissibility criterion. *Computing*, 73(3):207–243, 2004.

N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53 (2):217–288, 2011.

T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, 2015.

S. Jiang, L. Greengard, and S. Wang. Efficient sum-of-exponentials approximations for the heat kernel and their applications. *Adv. Comput. Math.*, 41(3):529–551, 2015.

F. John. Numerical solution of the equation of heat conduction for preceding times. *Ann. Mat. Pura Appl.*, 40(1):129–142, 1955.

G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nat. Rev. Phys.*, 3(6):422–440, 2021.

D. Kochkov, J. A. Smith, A. Alieva, Q. Wang, M. P. Brenner, and S. Hoyer. Machine learning–accelerated computational fluid dynamics. *Proc. Natl. Acad. Sci. USA*, 118(21), 2021.

N. Kovachki, S. Lanthaler, and S. Mishra. On universal approximation and error bounds for Fourier Neural Operators. *J. Mach. Learn. Res.*, 22:1–76, 2021a.

N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar. Neural operator: Learning maps between function spaces. *arXiv preprint arXiv:2108.08481*, 2021b.

A. Krishnapriyan, A. Gholami, S. Zhe, R. Kirby, and M. W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 26548–26560, 2021.

J. N. Kutz. Deep learning in fluid dynamics. *J. Fluid Mech.*, 814:1–4, 2017.

S. Lanthaler, S. Mishra, and G. E. Karniadakis. Error estimates for DeepONets: A deep learning framework in infinite dimensions. *Trans. Math. Appl.*, 6(1), 2022.

J.-R. Li and L. Greengard. On the numerical solution of the heat equation I: Fast solvers in free space. *J. Comput. Phys.*, 226(2):1891–1901, 2007.

J.-R. Li and L. Greengard. High order accurate methods for the evaluation of layer heat potentials. *SIAM J. Sci. Comput.*, 31(5):3847–3860, 2009.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Neural operator: Graph kernel network for partial differential equations. arXiv preprint arXiv:2003.03485, 2020a.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, A. Stuart, K. Bhattacharya, and A. Anandkumar. Multipole graph neural operator for parametric partial differential equations. In *Advances in Neural Information Processing Systems*, volume 33, pages 6755–6766, 2020b.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

G. M. Lieberman. *Second order parabolic differential equations*. World Scientific, 1996.

L. Lin, J. Lu, and L. Ying. Fast construction of hierarchical matrix representation from matrix–vector multiplication. *J. Comput. Phys.*, 230(10):4071–4087, 2011.

L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nat. Mach. Intell.*, 3(3):218–229, 2021.

P.-G. Martinsson and J. A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numer.*, 29:403–572, 2020.

W. L. Miranker. A well posed problem for the backward heat equation. *Proc. Amer. Math. Soc.*, 12(2):243–247, 1961.

E. Qian, B. Kramer, B. Peherstorfer, and K. Willcox. Lift & learn: Physics-informed machine learning for large-scale nonlinear dynamical systems. *Phys. D*, 406:132401, 2020.

E. Qian, I.-G. Farcas, and K. Willcox. Reduced operator inference for nonlinear partial differential equations. arXiv preprint arXiv:2102.00083, 2021.

M. Raissi, A. Yazdani, and G. E. Karniadakis. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 367(6481):1026–1030, 2020.

S. Rasp, M. S. Pritchard, and P. Gentine. Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci. USA*, 115(39):9684–9689, 2018.

S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz. Data-driven discovery of partial differential equations. *Sci. Adv.*, 3(4), 2017.

F. Schäfer and H. Owhadi. Sparse recovery of elliptic solvers from matrix-vector products. *arXiv preprint arXiv:2110.05351*, 2021.

Z. Song, D. P. Woodruff, and P. Zhong. Low rank approximation with entrywise ℓ_1-norm error. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 688–701, 2017.

A. Townsend and L. N. Trefethen. Continuous analogues of matrix factorizations. *Proc. Roy. Soc. A*, 471, 2015.

S.-M. Udrescu and M. Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Sci. Adv.*, 6(16), 2020.

S.-M. Udrescu, A. Tan, J. Feng, O. Neto, T. Wu, and M. Tegmark. AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. In *Advances in Neural Information Processing Systems*, volume 33, pages 4860–4871, 2020.

S. Wang, H. Wang, and P. Perdikaris. Learning the solution operator of parametric partial differential equations with physics-informed DeepONets. *Sci. Adv.*, 7(40):eabi8605, 2021.

S. Wang, X. Yu, and P. Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *J. Comput. Phys.*, 449:110768, 2022.

L. Zanna and T. Bolton. Data-driven equation discovery of ocean mesoscale closures. *Geophys. Res. Lett.*, 47(17), 2020.

S. Zhang and G. Lin. Robust data-driven discovery of governing physical laws with error bars. *Proc. R. Soc. A*, 474, 2018.