Convergence and error control of consistent PINNs for elliptic PDEs

Andrea Bonito, Ronald DeVore, Guergana Petrova, and Jonathan W. Siegel

June 14, 2024

Abstract

We provide an a priori analysis of a certain class of numerical methods, commonly referred to as *collocation methods*, for solving elliptic boundary value problems. They begin with information in the form of point values of the right side f of such equations and point values of the boundary function g and utilize only this information to numerically approximate the solution u of the Partial Differential Equation (PDE). For such a method to provide an approximation to u with guaranteed error bounds, additional assumptions on f and g, called *model class assumptions*, are needed. We determine the best error (in the energy norm) of approximating u, in terms of the number of point samples m, under all Besov class model assumptions for the right hand side f and boundary g.

We then turn to the study of numerical procedures and asks whether a proposed numerical procedure (nearly) achieves the optimal recovery error. We analyze numerical methods which generate the numerical approximation to u by minimizing a specified data driven loss function over a set Σ which is either a finite dimensional linear space, or more generally, a finite dimensional manifold. We show that the success of such a procedure depends critically on choosing a correct data driven loss function that is consistent with the PDE and provides sharp error control. Based on this analysis a loss function $\mathcal{L}^*$ is proposed.

We also address the recent methods of Physics Informed Neural Networks (PINNs). Minimization of the new loss $\mathcal{L}^*$ over neural network spaces Σ is referred to as consistent PINNs (CPINNs). We prove that CPINNs provides an optimal recovery of the solution u, provided that the optimization problem can be numerically executed and Σ has sufficient approximation capabilities. Finally, numerical examples illustrating the benefits of the CPINNs are given.

1 Introduction

This paper is concerned with numerical methods for solving elliptic partial differential equations (PDEs). Our primary goal is to provide a rigorous analysis of rates of convergence for collocation methods, including PINNs (Physics Informed Neural Networks), for solving such differential equations.

Let Ω be a bounded domain in the Euclidean space $\mathbb{R}^d$, $d \geq 2$, and $\overline{\Omega}$ be its closure in $\mathbb{R}^d$. Consider the elliptic Dirichlet problem

$$\begin{aligned} -\nabla \cdot (a(\mathbf{x})\nabla u(\mathbf{x})) &= f(\mathbf{x}), \quad \mathbf{x} \in \Omega, \\ u(\mathbf{z}) &= g(\mathbf{z}), \quad \mathbf{z} \in \partial\Omega, \end{aligned} \tag{1.1}$$

where $\partial\Omega$ is the boundary of Ω. In order to prove the existence and uniqueness of a solution to (1.1), one needs to impose conditions on f, g, a, and Ω. A standard set of assumptions that guarantee the existence and uniqueness of a solution (1.1) is the following:

A1: Ω is a Lipschitz (graph) domain, i.e., Ω is an open connected set in $\mathbb{R}^d$ with a Lipschitz boundary in the sense described in [57];

A2: $f \in H^{-1}(\Omega)$;

A3: $g \in H^{1/2}(\partial\Omega)$;

A4: the diffusion coefficient a satisfies the coercivity condition

$$0 < r \leq a(\mathbf{x}) \leq M, \quad \mathbf{x} \in \Omega, \tag{1.2}$$

for some constants $r, M > 0$. Here the spaces $H^s(\Omega)$, $s \in \mathbb{R}$, are the Sobolev spaces of order s in $L_2(\Omega)$.

Under the assumptions **A1-A4**, the Lax-Milgram theorem (see e.g. [25]) implies that (1.1) has a unique solution $u \in H^1(\Omega)$ which satisfies

$$c\{\|f\|_{H^{-1}(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)}\} \leq \|u\|_{H^1(\Omega)} \leq C\{\|f\|_{H^{-1}(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)}\}, \tag{1.3}$$

where the constants c, C depend on Ω, r and M [1]. Therefore, if we consider the theoretical loss function

$$\mathcal{L}_T(v) := \|f + \Delta v\|_{H^{-1}(\Omega)} + \|g - v\|_{H^{1/2}(\partial\Omega)}, \tag{1.4}$$

then minimizing this loss over the whole of $H^1(\Omega)$ has u as its unique solution.

In order to make our presentation as clear as possible, we want to avoid the technicalities of handling general domains and general diffusion coefficients a. Accordingly, we restrict our presentation to the case $\Omega = (0,1)^d$ and $a \equiv 1$. This will allow us to concentrate on the novelties of PINNs and alternative collocation methods. The treatment of more general settings for the diffusion a and domains Ω is left to future works. We also want to point out that assumptions **A2-A3** could be replaced with other smoothness assumptions on f and g, respectively, which will result in a theory similar to the one described in this paper.

1.1 Numerical methods for solving (1.1)

We are interested in numerical methods for solving (1.1). The most prominent of these are the Finite Element Methods (FEMs) and the adaptive variations of these (AFEMs). Over the last decades, a rigorous theory has been developed for FEMs and AFEMs which guarantees, a priori, the convergence of these numerical methods and even provides bounds on their rate of convergence under additional assumptions on f and g. These additional assumptions stipulate extra smoothness of f and g than those assumed in **A2-A3**.

A numerical method assumes that f, g, and Ω are known and suggests a procedure for solving (1.1) based on that information. The numerical procedure generates a function $\hat{u}$ which is an approximation to u. The efficiency of the numerical method is then evaluated in the following sense. One chooses a norm $\|\cdot\|_X$ in which to measure performance and then seeks to quantify how fast the error $\|u - \hat{u}\|_X$ between the true solution u and the output $\hat{u}$ of the numerical procedure tends to zero in terms of the complexity of the numerical procedure. The classical norm $\|\cdot\|_X$ used to measure error is the energy norm which corresponds to choosing $X = H^1(\Omega)$. Other function norms sometimes used to measure error correspond to $X = L_p(\Omega)$, $1 \leq p \leq \infty$. We restrict our analysis of convergence and rates of convergence to the case $X = H^1(\Omega)$ in going further.

An important question is how one can compare the performance of different numerical methods for solving (1.1) in order to give a fair competition between all possible numerical methods. This is typically done by assigning some form of complexity to the numerical method. In the case of FEMs and AFEMs, this complexity can be measured in terms of the number n of machine operations used to compute $\hat{u}_n = \hat{u}$ and there are theorems that give a priori bounds for the error $\|u - \hat{u}_n\|_{H^1(\Omega)}$ under additional assumptions on the smoothness of f and g (see e.g. [5]). Some alternatives to measuring the complexity n in terms of machine operations are the following. If the numerical approximation $\hat{u}$ to u is an element from a linear n-dimensional space V_n, then the associated complexity is typically assigned to be the dimension n. This may not be directly converted to an appropriate number of machine operations because numerical stability becomes an issue. If the approximation $\hat{u}$ comes from a nonlinear manifold $\mathcal{M}_n$, then one can use the number of parameters n determining the manifold as a complexity measure. However, in the latter case, one has to impose extra conditions on the manifold (or the numerical procedure) to prevent the use of space

[1] Throughout this paper, we use the notation c, C, c_1, C_1, etc, to denote constants. Unless they are absolute constants, we indicate the quantities on which they depend when they occur. The value of these constants can change at each occurrence. We also use the notation $A \asymp B$ to denote that $c_1 A \leq B \leq c_2 A$ with $c_1, c_2 > 0$.

filling manifolds and thus avoid unstable numerical methods, see [9]. In the absence of such restrictions, the numerical method may require an inordinate amount of computational resources to achieve a desired accuracy even in the case when the dimension n is small.

Another issue to consider is in what sense f, g, Ω are known/given to us. One setting is to assume that we can query (ask questions about) f, g, Ω, and receive the answer to such queries. In this setting, one has to decide which type of questions are allowed and whether these queries are answered free of charge or should the cost of asking/answering such queries be included in the numerical cost of the algorithm. For example, in FEMs, the data used are certain linear functionals (inner products with the basis functions of the Galerkin space) which are then utilized in the FEM approximation.

In this paper, we are interested in collocation methods as described in the next section. Accordingly, the linear functionals we consider are point evaluations of f and g at points from $\overline{\Omega}$. In order for these to make sense, we assume that f and g are continuous functions, i.e., $f \in C(\overline{\Omega})$, $g \in C(\partial\Omega)$. The accuracy of how well $\hat{u}$ approximates u will depend on two components. The first is the number m of point evaluations and their positioning. The optimal accuracy that any numerical method can achieve from these m pieces of information under the model class assumption on the function is called the *optimal recovery rate*. A central question in this paper is to determine this optimal recovery rate under various model class assumptions, provided the m data sites are optimally positioned. This is the subject of §3.

A second item in collocation methods is how one numerically uses the m pieces of information provided to create a numerical approximation to u. If Σ_n is the underlying space used to create $\hat{u}$ then the accuracy of the numerical recovery will depend on n and the choice of Σ_n. In other words, $\hat{u}$ can be viewed as $\hat{u} = \hat{u}_{n,m}$. If one fixes m, then one can study how fast the accuracy of approximation $\|u - \hat{u}_{n,m}\|$ tends to the optimal recovery rate as $n \to \infty$. This issue is discussed in §8.

Concerning the sense in which we know Ω, one usually considers polyhedral domains whose vertices and edges are given to us. As already noted, for simplicity, we assume that $\Omega = (0,1)^d$. However, with some additional technicality, we could equally well start with polyhedral domain. The standard way of handling more general domains Ω is to first find a polyhedral domain $\widehat{\Omega}$ that approximates Ω, solve the corresponding PDE on $\widehat{\Omega}$, and then analyze the effect of the approximation of Ω by $\widehat{\Omega}$.

1.2 Collocation methods and PINNs

Recently, there has been significant interest in using neural networks (NNs) as a nonlinear manifold to generate the approximation $\hat{u}$ to the solution u of (1.1). Let Σ_n denote the set of outputs of a neural network with n parameters and fixed architecture. Once Σ_n is chosen, the numerical procedure queries f and g and then uses this information to numerically construct a function $\hat{u} \in \Sigma_n$ which serves as an approximation to u in a specified norm $\|\cdot\|_X$. The queries of f and g are taken as point evaluations at specified points from $\overline{\Omega}$, thereby providing the values

$$\mathbf{f} = (f_1, \dots, f_{\widetilde{m}}), \quad f_i := f(\mathbf{x}_i), \quad i = 1, \dots, \widetilde{m}; \quad \mathbf{g} = (g_1, \dots, g_{\overline{m}}), \qquad g_i := g(\mathbf{z}_i), \quad i = 1, \dots, \overline{m}, \tag{1.5}$$

where $\mathbf{x}_i \in \overline{\Omega}$, $i = 1, \dots, \widetilde{m}$, and $\mathbf{z}_i \in \partial\Omega$, $i = 1, \dots, \overline{m}$, are the query sites. We refer to these points as *data sites* and denote them by

$$\mathcal{X} := \{\mathbf{x}_1, \dots, \mathbf{x}_{\widetilde{m}}\}, \quad \mathcal{Z} := \{\mathbf{z}_1, \dots, \mathbf{z}_{\overline{m}}\}.$$

The performance of such a numerical method will depend on the numbers $\widetilde{m}, \overline{m}$, which we refer to as the *query budget*, and also on the positioning of these points.

In PINNs, the numerical procedure to find $\hat{u}$ is to search over Σ_n and find a $\hat{u} \in \Sigma_n$ which 'fits the data'. The most frequently used procedure (there are many variants) is to take $\hat{u}$ as one of the elements of the set

$$\hat{u} \in \operatorname*{argmin}_{S \in \Sigma_n} \mathcal{L}_0(S), \tag{1.6}$$

where

$$\mathcal{L}_0(S) := \frac{1}{\widetilde{m}} \sum_{i=1}^{\widetilde{m}} [\Delta S(\mathbf{x}_i) + f(\mathbf{x}_i)]^2 + \frac{\lambda}{\overline{m}} \sum_{i=1}^{\overline{m}} [S(\mathbf{z}_i) - g(\mathbf{z}_i)]^2, \tag{1.7}$$

with λ being a tuning parameter. For simplicity, throughout the paper we will set $\lambda = 1$ and consider the loss

$$\mathcal{L}(S) := \left[\frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}}[\Delta S(\mathbf{x}_i) + f(\mathbf{x}_i)]^2\right]^{1/2} + \left[\frac{1}{\overline{m}}\sum_{i=1}^{\overline{m}}[S(\mathbf{z}_i) - g(\mathbf{z}_i)]^2\right]^{1/2} \tag{1.8}$$

in our analysis. Note that this is equivalent to the square root of the typical PINNs loss with $\lambda = 1$. The appropriateness of the loss $\mathcal{L}_0$ and how $\widetilde{m}, \overline{m}, \lambda$ and the data sites $(\mathcal{X}, \mathcal{Z})$ should be chosen are major issues and one of the focal points of this paper.

The PINNs numerical procedures fall into the broad class of *collocation methods* for solving PDEs, i.e. numerical methods that use only values of f and g at specified data sites $\mathcal{X}$ and $\mathcal{Z}$. Such methods were studied in the last century (see e.g. [1, 34, 31]) but became less popular with the advent of FEMs. The novelty of PINNs is to use neural networks instead of polynomials or splines to build the approximation $\hat{u}$ from the given data.

Some assumptions are necessary for collocation methods to make sense. Firstly, to ensure that point values of f and g make sense, we assume that f and g are continuous function on $\overline{\Omega}$ and $\partial\Omega$, respectively. Secondly, to have any hope of proving a priori guarantees on convergence or rates of convergence as $\widetilde{m}, \overline{m}, n \to \infty$, we need at a minimum that f comes from a compact subset $\mathcal{F}$ of $C(\overline{\Omega})$ and g comes from a compact subset $\mathcal{G}$ of $C(\partial\Omega)$. This in turn guarantees that u is in a compact subset $\mathcal{U}$ of $H^1(\Omega)$.

1.3 Overview of this paper

The idea of using neural networks to solve PDEs goes back to the last century [20, 39], and has recently become extremely popular with the introduction of physics-informed neural networks (PINNs) [50]. Despite the increasing usage of PINNs for numerically solving a wide range of PDEs (see for example, [36, 6, 10, 6, 43] and the references therein), a satisfactory analysis of the convergence and performance of these methods has not been given. Some partial progress towards this goal has recently been made, see for instance [24, 45, 53, 54, 65].

For example, in [53], convergence is proved in the $C(\overline{\Omega})$-norm for a modified physics-informed neural network for elliptic PDEs under the assumption that the right hand side f and boundary data g are Hölder continuous. Under the additional assumption that the outputs of the neural network satisfy the boundary conditions exactly, convergence is obtained in $H^1(\Omega)$. In this situation, the boundary values are *not enforced* via the loss function and must instead be implemented through the neural network architecture. We remark that although methods for exactly enforcing boundary conditions with neural networks have been proposed in [39, 40, 3], this approach appears unable to rigorously handle arbitrary boundary conditions on general domains.

Further, in [54] and [65], an abstract framework for analyzing methods based upon residual minimization, such as PINNs, is presented. This framework has been used in a variety of follow up works to analyze PINNs (see [2, 58, 22] for instance). However, the current analysis is, to the best of our knowledge, not able to obtain convergence rates with respect to the number of collocation points when a fixed (as opposed to random) set of collocation points are used.

An analysis taking into account the rate of convergence with respect to the number of collocation points has been obtained in [45]. It applies to the heat equation (instead of elliptic PDEs) under the assumption that the initial data lies in the Sobolev space $H^s(\Omega)$ for sufficiently large s to ensure that point values are well-defined.

However, so far, none of the existing literature treats the problem of obtaining convergence rates in terms of both the number of collocation points and the network size for elliptic PDEs under general Besov regularity assumptions on the right hand side f and boundary value g. A proper a priori analysis that addresses this problem would answer the following questions:

Q1: Given the model class assumptions on f and g, which query sites should be used and how large would we have to take $\widetilde{m}$ and $\overline{m}$ to be able to reach a target accuracy ε for the error $\|u - \hat{u}\|_{H^1(\Omega)}$?

Q2: Given a budget of m queries and given model class assumption on f and g, what is the smallest error that can be achieved in the recovery of u? This is called the *optimal recovery* accuracy.

Q3: If we use a NN space Σ_n with n parameters to build the approximation $\hat{u}$ to u, then how large do we need to choose n to achieve near optimal accuracy?

Q4: If we use minimization of a data driven loss function to generate the numerical approximation $\hat{u}$, then what form should the loss function take so that small values of this loss function guaranees small values of the solution approximation error, up to the optimal recovery rate ?

The goal of the present paper is to answer the above questions for all Besov space model classes and thereby establish such an a priori analysis of collocation methods and in particular PINNs. In contrast to previous works, we also consider the error in the $H^1(\Omega)$-norm, and treat the situation where the boundary values are enforced through the loss function. As already noted, in order to simplify our presentation, we will only consider the case of Laplace's equation (1.1) with $a \equiv 1$ on Ω and $\Omega = (0,1)^d$. We leave the problem of handling more general settings to future work. We remark that an analysis of other methodologies for solving (1.1) using neural networks, such as the deep Ritz method [60] and finite neuron method [61], can be found in [42, 41, 56, 47, 23, 21].

As we have already mentioned, any a priori convergence analysis of collocation methods requires assumptions on the functions f and g in the form that $f \in \mathcal{F}$, $g \in \mathcal{G}$, where $\mathcal{F}$ and $\mathcal{G}$ are compact subsets of the space of continuous functions on the domains $\overline{\Omega}$ and $\partial\Omega$, respectively. Such compact sets are typically described by smoothness conditions. In this paper, we measure smoothness by membership in Besov classes. For this reason, we begin in §2 with the definition of Besov spaces and the properties we will need in the sequel. That section will also discuss how well functions in Besov classes can be approximated (in various norms) by piecewise polynomials. All of the results in that section are known and are therefore presented in a summary form. However, for completeness of the paper, and for the specific presentation of these results, we give proofs of the results we need in the appendix.

We use membership in Besov spaces $B^s_q(L_p(\Omega))$ as the smoothness we impose on f and g. Once such assumptions are placed on f and g, we answer questions **Q1-Q2** in §3. Questions of this type are commonly referred to as *optimal recovery* (OR) of a function u from data. Since our data is given by point samples, this is also known as *optimal sampling*. Our general setting is not usually addressed in the OR literature because our data is not point evaluations of the target function u, but rather of f and g. Nevertheless, we consider the results of §3 to be for the most part known in the sense that we are simply piecing together various known results and techniques such as those in [12, 13, 38, 49]. The paper [59], which studies optimal sampling with respect the negative Sobolev norm $H^{-1}(\Omega)$ on Lipschitz domains Ω, comes closest to presenting the optimal recovery results in the form we need. However, note that the cases $0 < p, q < 1$ for all d, or $p = 1$, $0 < q \leq \infty$, $d = 2$ are not covered in [59]. Notably, we observe that the optimal sampling rates for Besov spaces in $H^{-1}(\Omega)$ coincide with the optimal sampling rates in a certain space $L_\gamma(\Omega)$, determined by the Sobolev embedding (see Theorem 3.1 (iv)).

The results of §3 do not give a numerical method for finding $\hat{u}$ from the given data that provides the error bound $\|u - \hat{u}\|_X \leq \varepsilon$. They simply establish the theoretical benchmark for the optimal accuracy of any numerical procedure for the recovery of u based on the point samples of f and g. This theoretical analysis does not involve NNs or any other proposed collocation method. NNs and PINNs only enter the picture when one wants to transform the theoretical analysis into a numerical procedure that utilizes NNs to find an approximant to u.

We next discuss in §4 the use of classical numerical methods such as Finite Element Methods (FEMs) for optimal recovery. In §5, we turn to the question of using optimization to solve the PDE. We observe that minimizing the theoretical loss $\mathcal{L}_T$ over a space Σ_n will approximate well u, provided Σ_n has sufficient approximation power. However, this is not a numerical algorithm per se because it does not provide a numerical recipe for evaluating $\mathcal{L}_T(S)$ when $S \in \Sigma_n$.

The remainder of this paper discusses possible numerical methods to achieve the optimal recovery rate, i.e., to solve the PDE to the highest accuracy possible given only the information $(\mathbf{f}, \mathbf{g})$ (see (1.5)). An important consequence of the analysis in §5 is that the widely used loss function $\mathcal{L}_0$ (or $\mathcal{L}$) may not be an

appropriate discretization of the theoretical loss $\mathcal{L}_T$. We show that the correct loss function $\mathcal{L}^* : H^1(\Omega) \to \mathbb{R}$ to use in collocation methods in our setting is

$$\mathcal{L}^*(v) := \left[\frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} |f(\mathbf{x}_i) + \Delta v(\mathbf{x}_i)|^\gamma\right]^{1/\gamma} + \left[\frac{1}{\overline{m}^2}\sum_{\substack{i,j=1\\ i\neq j}}^{\overline{m}} \frac{|[g-v](\mathbf{z}_i) - [g-v](\mathbf{z}_j)|^2}{|\mathbf{z}_i - \mathbf{z}_j|^2}\right]^{1/2} + \left[\frac{1}{\overline{m}}\sum_{j=1}^{\overline{m}} |g(\mathbf{z}_j) - v(\mathbf{z}_j)|^2\right]^{1/2}, \tag{1.9}$$

where γ is the smallest number (if this is possible) so that $L_\gamma(\Omega)$ embeds in $H^{-1}(\Omega)$. For example, $\gamma = \frac{2d}{d+2}$ in the case $d \geq 3$. The case $d = 2$ is more complicated and we will discuss it later in the paper.

The loss $\mathcal{L}^*$ and its properties are given in §5, §6 and §7. A large component of its development centers on how to discretize L_p (quasi-)norms and the $H^{1/2}$ norm. Our analysis on this subject may be useful in the development of other methods for solving PDEs numerically.

The first term in the loss $\mathcal{L}^*$ are shown to bound $\|f + \Delta v\|_{H^{-1}(\Omega)}$ and the second and third terms are shown to bound $\|g - v\|_{H^{1/2}(\partial\Omega)}$. These in turn give a bound for $\|u - v\|_{H^1(\Omega)}$ via (1.3). We go on to show how near minimizers of the suggested loss $\mathcal{L}^*$ give an a priori bound of the approximation error, provided that the approximation method Σ_n has suitable approximation properties. The a priori guarantee provides a near optimal recovery whenever Σ_n provides sufficiently good approximation of the elements in the solution model class $\mathcal{U}$. Notice that this new loss function replaces the first discrete ℓ_2 term in (1.7) by a discrete ℓ_γ loss for a proper choice of γ and the second discrete ℓ_2 loss (used to match boundary values) by a weighted ℓ_2 loss.

We next develop in §8 the properties needed of a set Σ_n so that minimizing the loss $\mathcal{L}^*$ over Σ_n results in a near optimal recovery of u. It turns out that a certain restricted approximation property, guaranteed to hold when Σ_n is sufficiently good at approximating the elements of solution model class $\mathcal{U}$, is sufficient. In particular, we show that this is the case for suitable NN spaces Σ_n, provided n is large enough.

The remainder of this paper focuses on using minimization over neural network spaces as the numerical method as is the case in PINNs. As noted above, the loss $\mathcal{L}^*$ is the correct loss to use in such a minimization since it is consistent with recovering u in the $H^1(\Omega)$. When the loss $\mathcal{L}^*$ is minimized over a neural network space, we call the numerical procedure consistent PINNs (CPINNs).

The later sections of the paper analyze the theoretical gains in using CPINNs over PINNS. We give a priori bounds on how well the solution CPINNs approximates u and, in particular, we provide a sufficient condition on the size n of the neural network to guarantee optimal recovery. These bounds are established under the assumption that the optimization problem is solved (or approximately solved) when optimizing over the NN space Σ_n. As we know, proving that a particular numerical method of optimization (such as gradient decent) converges to a minimizer (or near minimizer) of the loss is a serious issue in optimization, which we do not address here. However, we do prove that for any function $v \in H^1(\Omega)$, the quantity $\mathcal{L}^*(v)$ always provides an upper bound of the true error $\|u - v\|_{H^1(\Omega)}$ up to the optimal recovery rate for the solution model class $\mathcal{U}$. This means that $\mathcal{L}^*(v)$ can be used as an a posteriori error estimator for any proposed numerical approximation v to u. This error estimate can be used to check whether the output of a PINNs optimization achieves the desired accuracy ε. In other words, although we do not guarantee a priori that PINNs will achieve the accuracy ε (because of lack of performance analysis of optimization procedures like gradient descent or stochastic gradient descent), we do give an a numerically implementable a posteriori bound on performance which may serve to guarantee optimality.

In summary, a theorem which guarantees that PINNs provide an approximation to u in the $\|\cdot\|_{H^1(\Omega)}$ norm with a prescribed accuracy ε requires not only smoothness assumptions on f and g as described above, assumptions on the spacing and number of elements in the data sites $\mathcal{X}$ and $\mathcal{Z}$, and assumptions on the size of n, i.e., the complexity of Σ_n, but also an efficient numerical method that properly chooses a good approximation $\hat{u}$ to u in the norm $\|\cdot\|_{H^1(\Omega)}$ through minimizing the loss $\mathcal{L}^*$ over Σ_n.

Acknowledgments

We are grateful to Peter Binev, Albert Cohen, Wolfgang Dahmen, and Jinchao Xu for many insightful conversations about the material in this paper.

This research was supported by the NSF Grants DMS-2409807 (AB), DMS-2424305 (JWS), CCF-2205004 (JWS), and DMS 2134077 (RD and GP), and the MURI ONR Grant N00014-20-1-278 (RD, GP, and JWS).

2 Besov spaces

We start this section by recalling the definition of Besov spaces and their properties. We confine this paper to the case of functions defined on the domain $\Omega = (0,1)^d$, $d \geq 2$. For the range $s > 0$, and $0 < p, q \leq \infty$, the Besov space $B^s_q(L_p(\Omega))$ is a space of functions with smoothness of order $s > 0$ in $L_p(\Omega)$. Here q is a secondary index that gives a fine gradation of these spaces. The material in this section is taken for the most part from the papers [18, 16, 17] and the reader will have to refer to those papers for some of the definitions and proofs. Let us also mention that the univariate case is covered in the book [15].

If r is a positive integer and $0 < p \leq \infty$ and $f \in L_p(\Omega)$, we define the modulus of smoothness $\omega_r(f,\cdot)_p$ of f by

$$\omega_r(f,t)_p := \omega_r(f,t,\Omega)_p := \sup_{|\mathbf{h}|\leq t} \|\Delta^r_{\mathbf{h}}(f,\cdot)\|_{L_p(\Omega_{r\mathbf{h}})}, \quad t > 0, \tag{2.1}$$

where

$$\Delta^r_{\mathbf{h}}(f,\cdot) := (-1)^r \sum_{k=0}^{r} (-1)^k \binom{r}{k} f(\cdot + k\mathbf{h}), \tag{2.2}$$

is the r-th difference of f for $\mathbf{h} \in \mathbb{R}^d$ and $\Omega_{\mathbf{h}} := \{\mathbf{x} \in \Omega : [\mathbf{x}, \mathbf{x}+\mathbf{h}] \subset \Omega\}$. Here $[\mathbf{x}, \mathbf{x}+\mathbf{h}]$, $\mathbf{x}, \mathbf{h} \in \mathbb{R}^d$, denotes the line segment in $\mathbb{R}^d$ between $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$, and $|\mathbf{h}|$ denotes the Euclidean norm of $\mathbf{h}$. If $s > 0$ and $0 < p \leq \infty$, then $B^s_q(L_p(\Omega))$ is defined as the set of all functions in $L_p(\Omega)$ for which

$$|f|_{B^s_q(L_p(\Omega))} := \left[\int_0^1 [t^{-s}\omega_r(f,t)_p]^q \frac{dt}{t} \right]^{1/q} < \infty, \quad 0 < q < \infty, \tag{2.3}$$

where r can be taken as any integer strictly bigger than s. When $q = \infty$ we replace the integral by a supremum in the definition. This is a (quasi-)semi-norm and we obtain the (quasi-)norm for $B^s_q(L_p(\Omega))$ by adding $\|f\|_{L_p(\Omega)}$ to it. An equivalent (quasi-)semi-norm is given by

$$|f|_{B^s_q(L_p(\Omega))} \asymp \left[\sum_{k=0}^{\infty} [2^{ks}\omega_r(f,2^{-k})_p]^q \right]^{1/q}, \quad 0 < q < \infty, \tag{2.4}$$

with equivalency constants independent of f. This is proved by discretizing the integral in (2.3) and using the monotonicity of ω_r as a function of t. When $q = \infty$, (2.4) uses the ℓ_∞ norm in place of the ℓ_q norm, i.e.,

$$|f|_{B^s_\infty(L_p(\Omega))} \asymp \sup_{k\geq 0} 2^{ks}\omega_r(f,2^{-k})_p. \tag{2.5}$$

While different choices of r result in different (quasi-)semi-norms, the corresponding Besov (quasi-)norms are equivalent, provided $r > s$. To fix matters, we define the Besov norm with the value of $r = r(s)$ as the smallest integer strictly larger than $\max\{s, 1\}$. It follows that $r(s)$ is always larger than or equal to 2. This choice is for notational convenience in the material that follows.

Let us make some remarks on these spaces. Consider the role of q in this definition. If $q_2 < q_1$, then $B^s_{q_2}(L_p(\Omega)) \subset B^s_{q_1}(L_p(\Omega))$, i.e., these spaces get smaller as q gets smaller. Thus, all $B^s_q(L_p(\Omega))$, $q > 0$, are contained in $B^s_\infty(L_p(\Omega))$ once s and p are fixed. The effect of q in the definition of the Besov spaces is subtle.

In this paper, the space $B^s_\infty(L_p(\Omega))$ is especially important, and accordingly, we use the abbreviated notation

$$B^s_p := B^s_p(\Omega) := B^s_\infty(L_p(\Omega)) \tag{2.6}$$

in going forward. It follows that a function $f \in B^s_p(\Omega)$ if and only if

$$\omega_r(f,t)_p \le Mt^s, \quad t > 0, \tag{2.7}$$

and the smallest M for which (2.7) is valid is the (quasi-)semi-norm $|f|_{B^s_p}$. This space is commonly referred to as (generalized) Hölder smoothness of order s in L_p. It is important to note that when s is a positive integer, we take $r > s$ in its definition. Therefore the Besov space B^s_p is not a Lipschitz space when s is a positive integer. In view of (2.5), we have that a function f is in B^s_p if and only if

$$\omega_r(f, 2^{-k})_{L_p(\Omega)} \le M' 2^{-ks}, \quad k = 0, 1, \ldots, \tag{2.8}$$

and the smallest M' for which this is true is an equivalent (quasi-)semi-norm.

2.1 The Sobolev spaces $H^s(\Omega)$

Since the results of this paper heavily use the Sobolev spaces $H^s(\Omega)$, we give here a short review of these spaces and list some of their properties. The classical Sobolev spaces $H^s(\Omega)$, $s > 0$, are the Besov spaces

$$H^s(\Omega) = B^s_2(L_2(\Omega)) \subset B^s_\infty(L_2(\Omega)) =: B^s_2(\Omega).$$

Two of the $H^s(\Omega)$ spaces play an important role in what follows, namely $H^{-1}(\Omega)$ and $H^{1/2}(\partial\Omega)$. Let us recall the definition of $H^{-1}(\Omega)$. The space $H^1_0(\Omega)$ is the collection of functions in $H^1(\Omega)$ which vanish on the boundary of Ω, i.e., it is the closure of smooth functions in $H^1(\Omega)$ which are identically zero on $\partial\Omega$. The space $H^{-1}(\Omega)$ is by definition the dual space of $H^1_0(\Omega)$ and is equipped with the norm

$$\|f\|_{H^{-1}(\Omega)} := \sup_{\|v\|_{H^1_0(\Omega)}=1} \langle f, v\rangle_{H^{-1}\times H^1_0}. \tag{2.9}$$

There are two equivalent ways to describe the space $H^{1/2}(\partial\Omega)$ appearing in **A3**. The first is as the Besov space $B^{1/2}_2(L_2(\partial\Omega))$. Here, one needs the concept of Besov spaces on manifolds. In our case, the boundary of Ω is quite simple since it is the union of the faces of Ω. This definition gives the norm in $H^{1/2}(\partial\Omega)$ through the modulus of smoothness. Using the averaged modulus of smoothness, see the Appendix, one finds that the semi-norm of $g \in H^{1/2}(\partial\Omega)$, is

$$|g|^2_{H^{1/2}(\partial\Omega)} := \int_{\partial\Omega}\int_{\partial\Omega} \frac{|g(\mathbf{z}) - g(\mathbf{z}')|^2}{|\mathbf{z} - \mathbf{z}'|^d}\, d\mathbf{z}\, d\mathbf{z}', \tag{2.10}$$

which is commonly referred to as the *intrinsic semi-norm* on this space. We obtain the norm on this space by adding $\|g\|_{L_2(\partial\Omega)}$ to this semi-norm.

The second way to describe $H^{1/2}(\partial\Omega)$ and its norm is through the trace operator $Tr = Tr_{\partial\Omega}$. Recall that if g is a continuous function on $\overline{\Omega}$ then the trace $Tr(g)$ is simply the restriction of g to $\partial\Omega$. While the trace operator has an extension to certain functions that are not continuous, in our applications that appear later the function g will always be continuous. Using the trace operator one defines $H^{1/2}(\partial\Omega)$ as the trace of functions in $H^1(\Omega)$ and its *trace norm* is

$$\|g\|_{H^{1/2}(\partial\Omega)} := \inf\{\|v\|_{H^1(\Omega)} : \ Tr(v) = g\}. \tag{2.11}$$

It is well known that the trace norm and the intrinsic norms for $H^{1/2}(\Omega)$ are equivalent for Lipschitz domains and in particular for our case of $\Omega = (0,1)^d$ (see [27]). We use both of these norms in what follows while always making clear which norm is being employed.

2.2 Piecewise polynomial approximation and interpolation

Recall that one can characterize membership in Besov spaces by piecewise polynomial approximation. We describe and prove such characterizations in the appendix. An important additional fact is that the piecewise polynomials in such characterizations can be described by *interpolation*. This allows us to generate near best piecewise polynomial approximations to f and g by using only the data (f_i), respectively (g_i).

Let us begin with the cube $\Omega = (0,1)^d$ and discuss polynomial interpolation on $\overline{\Omega}$ which we will later rescale to any dyadic cube. We fix $r \in \mathbb{N}$, $r > 1$ and let

$$G_r := \left\{ \left(\frac{j_1}{r-1}, \dots, \frac{j_d}{r-1} \right),\ j_i \in \{0, 1, \dots, r-1\} \right\} \subset [0,1]^d,$$

be the tensor product grid of r^d equally spaced points in $\overline{\Omega}$. Consider the simplicial (Kuhn-Tucker) decomposition of $\overline{\Omega}$ into simplices T,

$$\overline{\Omega} = \bigcup \overline{T}.$$

Given any one of these simplices T, the number n_r of grid points in this simplex $\overline{T}$ is the same as the dimension of $\mathcal{P}_r := \mathcal{P}_k^d$, where $\mathcal{P}_r^d$ denotes the linear space of algebraic polynomials of order r (total degree $r-1$), namely,

$$\mathcal{P}_r^d := \left\{ \sum_{|\mathbf{k}|_1 < r} a_{\mathbf{k}} \mathbf{x}^{\mathbf{k}},\ a_{\mathbf{k}} \in \mathbb{R} \right\}, \quad \text{where} \quad \mathbf{x}^{\mathbf{k}} := x_1^{k_1} \cdots x_d^{k_d}, \quad \mathbf{k} := (k_1, \dots, k_d), \quad k_j \geq 0, \quad |\mathbf{k}|_1 := \sum_{j=1}^d k_j.$$

Let us consider the standard closed simplex

$$\overline{T}_0 := \{\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d :\ x_1 + \ldots + x_d \leq 1,\ x_i \geq 0,\ i = 1, \dots, d\}.$$

It is well known that polynomial interpolation using elements of $\mathcal{P}_r$ at the grid points in $G_r \cap \overline{T_0}$ is well posed (see e.g. Section 4 in [51] or Chapter 2 in [7]. Let $\phi_j := \phi_{j,T_0}$, $j = 1, \dots, n_r$, be the Lagrange polynomial basis for $\mathcal{P}_r$ corresponding to the points in $G_r \cap \overline{T}_0$. Then, the operator

$$L_{T_0}(f) := \sum_{\mathbf{x}_j \in G_r \cap T_0} f(\mathbf{x}_j) \phi_j, \tag{2.12}$$

is a bounded projector from $C(\overline{\Omega})$ onto $\mathcal{P}_r$, i.e.,

$$\|L_{T_0}(f)\|_{C(\overline{T_0})} \leq \Lambda_r \|f\|_{C(\overline{T_0})}, \quad \text{where} \quad \Lambda_r := \| \sum_{j=1}^{n_r} |\phi_{j,T_0}(\mathbf{x})| \|_{C(\overline{T_0})}, \tag{2.13}$$

is the Lebesgue constant which depends only on r, d, and T_0.

Now, fix $k \geq 0$ and consider the dyadic cubes $I \in \mathcal{D}_k$. If we rescale to any such dyadic cube I and any simplex T in the simplicial decomposition of I and denote by $\phi_{j,T}$ the rescaled polynomial $\phi_{j,T} := \phi_{j,T_0} \circ F_T^{-1}$ where $F_T : \overline{T}_0 \to \overline{T}$ is an affine map from the reference simplex $\overline{T}_0$ to $\overline{T}$, we obtain the projector L_T onto $\mathcal{P}_r$ which satisfies

$$\|L_T(f)\|_{C(\overline{T})} \leq \Lambda_r(T) \|f\|_{C(\overline{T})}, \quad \text{where} \quad \Lambda_r(T) := \| \sum_{j=1}^{n_r} |\phi_{j,T}(\mathbf{x})| \|_{C(\overline{T})}, \quad T \subset I \in \mathcal{D}_k,\ k \geq 0. \tag{2.14}$$

Notice that $\Lambda_r = \Lambda(T)$. Moreover, the interpolating polynomial $L_T(f)$ provides a good approximation to f on T. For example, for any polynomial $P \in \mathcal{P}_r$, we have

$$\|f - L_T(f)\|_{C(\overline{T})} \leq \|f - P\|_{C(\overline{T})} + \|L_T(f - P)\|_{C(\overline{T})} \leq (1 + \Lambda_r) \|f - P\|_{C(\overline{T})}. \tag{2.15}$$

Let us denote by $S_k^*(f)$ the piecewise polynomial interpolant

$$S_k^*(f) := S_{k,r}^*(f) := \sum_{T\in\mathcal{T}_k} L_T(f)\chi_T, \tag{2.16}$$

where $\mathcal{T}_k$ is the collection of all simplicies arising from the (Kuhn-Tucker) decomposition of the dyadic cubes in $\mathcal{D}_k(\overline{\Omega})$ and χ_T is the characteristic function of T. The function $S_k^*(f)$ is continuous on $\overline{\Omega}$ since the number points from G_r that are on a shared face of any two simplicies from $\mathcal{T}_k$ is the same as the dimension of the polynomial space $\mathcal{P}_r^{d-1}$.

The following theorem (proved in the appendix) describes the appproximation accuracy of S_k^*.

Theorem 2.1. *Let $s > 0$, $0 < p \le \infty$, be fixed with $s > d/p$. Let $r = r(s)$ be the smallest integer strictly larger than* $\max(s, 1)$ *and let S_k^* be the interpolation operator* (2.16)*. Then, for any $f \in B_p^s(\Omega) = B_\infty^s(L_p(\Omega))$ and any $\tau \ge p$, we have*

$$\|f - S_k^*(f)\|_{L_\tau(\Omega)} \le C|f|_{B_p^s(\Omega)} 2^{-k(s-d/p+d/\tau)}, \tag{2.17}$$

with C independent of f and k.

We will also need similar results where the approximation takes place in the $H^1(\Omega) = B_2^1(L_2(\Omega))$ norm.

Theorem 2.2. *Let $d \ge 2$. If $f \in B_q^s(L_p(\Omega))$ with $s > d/p$ and $0 < p \le 2$ and $S_k^*(f)$ is as defined in* (2.16)*, then*

$$\|f - S_k^*(f)\|_{H^1(\Omega)} \le C|f|_{B_q^s(L_p(\Omega))} 2^{-k(s-1-d/p+d/2)}, \quad k \ge 0. \tag{2.18}$$

There are many results concerning the embeddings of Besov spaces into other Besov spaces or L_p spaces. We are particularly interested in those when a Besov space is embedded into $C(\Omega)$. For this, we can use the following theorem proved in the appendix

Theorem 2.3. *For every $f \in B_p^s(\Omega)$, $s > d/p$, $0 < p \le \infty$, there is a continuous function $\widetilde{f}$ such that $f = \widetilde{f}$ a.e. In fact, $\widetilde{f} \in Lip\ \alpha$, with $\alpha := s - d/p$, i.e.,*

$$\omega_r(\widetilde{f}, t)_{C(\overline{\Omega})} \le C|f|_{B_p^s(\Omega)} t^{s-d/p}, \quad t > 0. \tag{2.19}$$

3 Optimal recovery

In this section, we study the question of whether the information $(\mathbf{f}, \mathbf{g})$ at the **fixed** data sites $\mathcal{X} := \{\mathbf{x}_1, \dots, \mathbf{x}_{\widetilde{m}}\}$ and $\mathcal{Z} := \{\mathbf{z}_1, \dots, \mathbf{z}_{\overline{m}}\}$ is sufficient to determine u to a prescribed accuracy ε in the $H^1(\Omega)$ norm. Questions of this type are well studied and referred to as optimal recovery (OR). The answer to our specific question depends on the numbers $\widetilde{m}, \overline{m}$, the positions of the points in $\mathcal{X}$ and $\mathcal{Z}$, and the assumptions we make on f and g. As we have mentioned, we assume that f and g are continuous so that point evaluation makes sense. It follows that the unknown function u is continuous as well. If we wish to ensure an error bound on how well we can recover u in the $H^1(\Omega)$ norm, then the conditions we impose on f and g also have to guarantee that u lies in a compact subset of $H^1(\Omega)$. We require that f, g are in certain Besov spaces that compactly embed into the appropriate space of continuous functions, i.e., f lies in a Besov space that compactly embeds into $C(\Omega)$ and g is in a Besov space that compactly embeds into $C(\partial\Omega)$.

Before presenting the results on optimal recovery of this section, let us remark that the results given below are (in essence) all known. For example, the optimal recovery rates for the Besov model classes of this section have all been given in the papers [12, 13, 59] save for one small discrepancy on optimal recovery in $H^{-1}(\Omega)$ (explained below) and for the fact that they do not directly consider recovery on manifolds like $\partial\Omega$. The main distinction between the results in [59] and ours given below is in how the optimal recovery is obtained/proven.The optimal recovery method in [59] uses linear combinations of certain 'bump functions', whereas we use continuous piecewise polynomial Lagrange interpolation. One can argue that the fact that optimal recovery can be obtained by interpolation is also known since it is a prominent method of approximation in Finite Element Methods. However, the latter community usually considers Sobolev

classes, rather than the more general Besov classes, and also typically restricts $p \geq 1$. In summary, we could not find a direct reference which proves the results given below based on continuous piecewise polynomial interpolation. In any case, those not familiar with optimal recovery may find the results below useful and the OR community can just do a very light reading of this section. The results stated in this section are all proved in detail in the appendix.

We assume that f, g belong to the unit ball of a smoothness space that compactly embeds into the space of continuous functions on the relevant domain. The function f is defined on the domain $\Omega \subset \mathbb{R}^d$ with continuous extension on $\overline{\Omega}$. The classical assumptions on function smoothness for a domain Ω is membership in a Besov space. Accordingly, we shall assume that

$$f \in \mathcal{F} := U(\mathcal{B}), \quad \mathcal{B} = B^s_q(L_p(\Omega)), \quad 0 < q, p \leq \infty, \quad s > d/p, \tag{3.1}$$

where $U(\mathcal{B})$ is the unit ball of $\mathcal{B}$ and the restriction on s ensures that $\mathcal{F}$ compactly embeds into the space of continuous functions $C(\overline{\Omega})$.

The function g is defined on the manifold $\partial\Omega$. There are two ways to describe smoothness conditions on such functions g. The first is through the trace operator $Tr := Tr_{\partial\Omega}$. The second is to place smoothness conditions directly on g as a function from $C(\partial\Omega)$. The latter is usually referred to as an intrinsic definition. We choose in this paper to define the set $\mathcal{G}$ via the trace operator. Namely, we define the model class $\mathcal{G}$ via the trace of functions from the unit ball of a Besov class $B^{\overline{s}}_{\overline{q}}(L_{\overline{p}}(\Omega))$ with $\overline{s} > d/\overline{p}$, to ensure continuity. We also place the restrictions $0 < \bar{p}, \bar{q} \leq 2$ in order to simplify the presentation that follows. In other words, our model class assumptions on the boundary function g take the form

$$g \in \mathcal{G} := Tr(U(\overline{\mathcal{B}})), \quad \text{where } \overline{\mathcal{B}} := B^{\overline{s}}_{\overline{q}}(L_{\overline{p}}(\Omega)), \quad \text{with } \bar{s} > d/\bar{p}, \quad 0 < \bar{p}, \bar{q} \leq 2. \tag{3.2}$$

Then $\mathcal{G}$ is a compact subset of $H^{1/2}(\partial\Omega)$. Note that in general the parameters s, p, q used in defining $\mathcal{F}$ are different from $\overline{s}, \overline{p}, \overline{q}$ used in defining $\mathcal{G}$. These assumptions imply that the function u we want to numerically approximate is an element of the set

$$\mathcal{U} := \{\widetilde{u} \in C(\overline{\Omega}) : \ \widetilde{u} \text{ satisfies (1.1) with } f \in \mathcal{F}, \ g \in \mathcal{G}\}. \tag{3.3}$$

Given data $\mathbf{f} := (f_1, \ldots, f_{\widetilde{m}})$, the totality of information we have about f is that it belongs to the set

$$\mathcal{F}_{\text{data}} := \mathcal{F}_{\text{data}(f)} := \{\widetilde{f} \in \mathcal{F} : \ \widetilde{f}_i = f_i, \ i = 1, \ldots, \widetilde{m}\}.$$

Similarly, if the data $\mathbf{g} := (g_1, \ldots, g_{\overline{m}})$ is coming from a $g \in \mathcal{G}$, then, what we know about g is that in lies in the set

$$\mathcal{G}_{\text{data}} := \mathcal{G}_{\text{data}(g)} := \{\widetilde{g} \in \mathcal{G} : \ \widetilde{g}_i = g_i, \ i = 1, \ldots, \overline{m}\}. \tag{3.4}$$

Notice that the $g_i = u(\mathbf{z}_i)$, $i = 1, \ldots, \overline{m}$. Finally, the totality of information we have about the sought after u is that it is in the set

$$\mathcal{U}_{\text{data}} := \mathcal{U}_{\text{data}(u)} := \{\widetilde{u} \in \mathcal{U}: \ -\Delta\widetilde{u}(\mathbf{x}_i) = f_i, \ i = 1, \ldots, \widetilde{m}, \quad \widetilde{u}(\mathbf{z}_i) = g_i, \quad i = 1, \ldots, \overline{m}\}.$$

or equivalently

$$\mathcal{U}_{\text{data}} := \mathcal{U}_{\text{data}(u)} := \{\widetilde{u} : \widetilde{u} \text{ is a solution to (1.1) with } \ f \in \mathcal{F}_{\text{data}}, \quad g \in \mathcal{G}_{\text{data}}\}.$$

We are interested in knowing to what extent the information that $u \in \mathcal{U}_{\text{data}}$ identifies u.

There is a simple theoretical answer to such questions. Given a compact set K of a Banach space X, we let $B(K) := B(K)_X$ denote a ball with the smallest radius in X which contains K, called a *Chebyshev ball* of K in X. If we wish to provide an element from X that simultaneously approximates all elements in K (in the norm of X), then the center of this ball is the best we can do and its radius

$$R(K)_X := \text{rad}(B(K))_X \tag{3.5}$$

is the *optimal error* we can obtain. It is called *the error of optimal recovery* of K.

If $\mathcal{U}_{\rm data}$ is a compact subset of X, let $B = B(\mathcal{U}_{\rm data})_X$ be a Chebyshev ball of this set. Since all we know about u is that it is in $\mathcal{U}_{\rm data}$, any element from $B(\mathcal{U}_{\rm data})_X$ will give a near best approximation to u in $\|\cdot\|_X$. The radius of B

$$R(\mathcal{U}_{{\rm data}(u)})_X := {\rm rad}(B(\mathcal{U}_{{\rm data}(u)}))_X \tag{3.6}$$

is then the optimal recovery error. We are interested in bounds for $R(\mathcal{U}_{\rm data})_X$, where $\|\cdot\|_X$ is the norm in which we wish to measure accuracy. We introduce similar notation for the recovery of f and g.

The recovery rate (3.6) will not only depend on the data sites $\mathcal{X}$, and $\mathcal{Z}$, but also on the values $(\mathbf{f}, \mathbf{g})$ assigned at these data sites. In order to obtain uniform estimates, we fix the data sites $(\mathcal{X}, \mathcal{Z})$ and introduce

$$R^*(\mathcal{U}, \mathcal{X}, \mathcal{Z})_X := \sup_{u\in\mathcal{U}} R(\mathcal{U}_{{\rm data}(u)})_X, \tag{3.7}$$

where the data values come from any $u \in \mathcal{U}$. This is measuring the worst possible performance over the class $\mathcal{U}$ and is called the *uniform optimal recovery rate* at the fixed data sites $(\mathcal{X}, \mathcal{Z})$. If we prescribe a budget for the number $m = |\mathcal{X}| + |\mathcal{Z}|$ of data sites, we can ask for the optimal error we can achieve under such a budget restriction. Accordingly, we define

$$R^*_m(\mathcal{U})_X := \inf_{\mathcal{X}\subset\overline{\Omega},\, \mathcal{Z}\subset\partial\Omega,\, |\mathcal{X}|+|\mathcal{Z}|=m} R^*(\mathcal{U}, \mathcal{X}, \mathcal{Z})_X, \quad m \ge 2.$$

Similarly,

$$R^*(\mathcal{F}, \mathcal{X})_X := \sup_{f\in\mathcal{F}} R(\mathcal{F}_{{\rm data}(f)})_X \quad \text{and} \quad R^*_{\widetilde{m}}(\mathcal{F})_X := \inf_{\mathcal{X}\subset\overline{\Omega}, |\mathcal{X}|=\widetilde{m}} R^*(\mathcal{F}, \mathcal{X})_X, \quad \widetilde{m} \ge 1,$$

where the latter is the minimum error we can achieve over all the possible choices of $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_{\widetilde{m}}\}$. We can similarly ask the question of how well can we can recover g from the data $\mathbf{g}$, and therefore how well we can recover u? Each of these questions has an associated uniform optimal recovery error which is dependent on which norm $\|\cdot\|_X$ we use to measure the error of recovery.

3.1 Optimal recovery of f

Let $\mathcal{F} := U(\mathcal{B})$ where $\mathcal{B} = B^s_q(L_p(\Omega))$ with $0 \le q, p \le \infty$ and $s > \frac{d}{p}$. We define the tensor product grid $G_{k,r}$ of $\overline{\Omega} = [0,1]^d$ for $r \ge 2$,

$$G_{k,r} := \{\mathbf{x}_1, \dots, \mathbf{x}_{\widetilde{m}}\} \subset \overline{\Omega}, \quad \widetilde{m} = (r2^k)^d, \tag{3.8}$$

with spacing $h = 2^{-k}(r-1)^{-1}$ (the role of r will be made clear below) and prove that, see Theorem 3.1,

$$R^*(\mathcal{F}, G_{k,r})_X \lesssim \widetilde{m}^{-\alpha_X}.$$

Moreover, we show that for any choice of data cites $\mathcal{X} \subset \overline{\Omega}$ with $|\mathcal{X}| = \widetilde{m}$, we have

$$R^*(\mathcal{F}, \mathcal{X})_X \gtrsim \widetilde{m}^{-\alpha_X},$$

and therefore $R^*_{\widetilde{m}}(\mathcal{F})_X \asymp \widetilde{m}^{-\alpha_X}$. The proof of this theorem is presented in the Appendix.

Theorem 3.1. *Let $\Omega = (0,1)^d$ and $\mathcal{F} := U(\mathcal{B})$, where $\mathcal{B}$ is the Besov space $B^s_q(L_p(\Omega))$ with $s > d/p$, and $0 < p, q \le \infty$. Then the following holds for the optimal recovery rate $R^*_{\widetilde{m}}(\mathcal{F})_X$ in the norm of X:*

(i) *If $X = C(\Omega)$, then*

$$R^*_{\widetilde{m}}(\mathcal{F})_X \asymp \widetilde{m}^{-\alpha_C},\ \widetilde{m} \ge 1, \quad \text{where} \quad \alpha_C := \frac{s}{d} - \frac{1}{p}, \tag{3.9}$$

with constants of equivalence independent of $\widetilde{m}$.

(ii) *If* $X = L_\tau(\Omega)$, $\tau > 0$, *then*

$$R^*_{\widetilde{m}}(\mathcal{F})_X \asymp \widetilde{m}^{-\alpha_\tau}, \ \widetilde{m} \geq 1, \quad \text{where} \quad \alpha_\tau := \frac{s}{d} - \left[\frac{1}{p} - \frac{1}{\tau}\right]_+, \tag{3.10}$$

with constants of equivalence independent of $\widetilde{m}$.

(iii) *If* $X = H^1(\Omega)$, $0 < p \leq 2$, *then*

$$R^*_{\widetilde{m}}(\mathcal{F})_X \asymp \widetilde{m}^{-\alpha_H}, \ \widetilde{m} \geq 1, \quad \text{where} \quad \alpha_H := \frac{s-1}{d} - \left[\frac{1}{p} - \frac{1}{2}\right], \tag{3.11}$$

with constants of equivalence independent of $\widetilde{m}$.

(iv) *If* $X = H^{-1}(\Omega)$, *then*

- *if* $d \geq 3$ *and* $0 < p \leq \infty$, *or* $d = 2$ *and* $1 < p \leq \infty$, *we have*

$$R^*_{\widetilde{m}}(\mathcal{F})_X \asymp \widetilde{m}^{-\alpha_{-1}}, \ \widetilde{m} \geq 1 \quad \text{where} \quad \alpha_{-1} := \frac{s}{d} - \left[\frac{1}{p} - \frac{1}{\delta}\right]_+ \text{ and } \frac{1}{\delta} := \frac{1}{2} + \frac{1}{d}, \tag{3.12}$$

  with constants of equivalence independent of $\widetilde{m}$.

- *if* $d = 2$ *and* $0 < p \leq 1$, *we have*

$$R^*_{\widetilde{m}}(\mathcal{F})_X \lesssim \log(\widetilde{m})\widetilde{m}^{-\alpha_{-1}}, \qquad R^*_{\widetilde{m}}(\mathcal{F})_X \gtrsim \widetilde{m}^{-\alpha_{-1}}, \qquad \widetilde{m} \geq 1, \tag{3.13}$$

  with α_{-1} *and* δ *as in* (3.12).

Moreover, when $\widetilde{m} = (r2^k)^d$, *the upper bounds in each of these estimates are obtained when we take the data sites* $\mathcal{X} = G_{k,r}$, *provided* $r > \max(s, 1)$. *In this case, the approximant to* f *is given by the continuous piecewise polynomial*

$$S^*_k(f) = \sum_{T \in \mathcal{T}_k} L_T(f)\chi_T,$$

described in (2.16), *where* $\mathcal{T}_k = \mathcal{T}_k(\Omega) = \bigcup T$ *is the collection of all simplicies arising from the (Kuhn-Tucker) conforming decomposition of the dyadic cubes in* $\mathcal{D}_k(\Omega)$, *the* χ_T*'s are the characteristic functions of* T, *and the* $L_T(f)$*'s are the polynomials of order* r *(*$L_T(f) \in \mathcal{P}^d_r$*), gotten by interpolating* f *at the data points* $T \cap G_{k,r}$.

We remark that in the case $d = 2$ and $0 < p \leq 1$, the logarithm appears because of the failure of the Sobolev embedding to provide $L_1(\Omega) \subset H^{-1}(\Omega)$. However, it is actually true that the Hardy space $\mathcal{H}^1(\Omega)$ embeds into $H^{-1}(\Omega)$. This could potentially be used to improve our later analysis, but for the sake of simplicity we do not pursue this here.

We make some further observations that will help explain the above theorem, especially what is happening for recovery in $H^{-1}(\Omega)$. Notice that under the model class assumption $f \in \mathcal{F} = U(B^s_q(L_p(\Omega)))$ with $s > d/p$, the OR for this class is the same whether we measure error in $H^{-1}(\Omega)$ or in $L_\delta(\Omega)$ when $d \geq 3$ or $d = 2$, $1 < p \leq \infty$. We also have that

$$\|f\|_{H^{-1}(\Omega)} \leq \begin{cases} C(d)\|f\|_{L_\delta(\Omega)}, & d \geq 3, \\ \\ \frac{C}{\tau - 1}\|f\|_{L_\tau(\Omega)}, & d = 2, \quad \tau > 1, \end{cases} \tag{3.14}$$

with C an absolute constant, as discussed in Lemma 12.4 of the appendix . These two facts explain the form of the loss function $\mathcal{L}^*$ which we discuss in detail later.

3.2 Optimal recovery of g

In this section, we study the optimal recovery rate of functions $g \in \mathcal{G}$ in the norm of $H^{1/2}(\partial\Omega)$. Let $Tr = Tr_{\partial\Omega}$ be the trace operator onto $\partial\Omega$ and let $\overline{\mathcal{B}} := B^{\bar{s}}_{\bar{q}}(L_{\bar{p}}(\Omega))$, where we fix $\bar{s}, \bar{p}, \bar{q}$ and assume that $\overline{s} > d/\overline{p}$ so that we are sure that the functions in $\overline{\mathcal{B}}$ are continuous on $\overline{\Omega}$. We consider the case $0 < \bar{p} \leq 2$, which together with the restriction $d \geq 2$ implies that $\bar{s} - d(1/\bar{p} - 1/2) > 1$, and thus the unit ball of $\overline{\mathcal{B}}$ compactly embeds into $H^1(\Omega)$. We define the model class

$$\mathcal{G} := \{g : \ g = Tr(v) : \ \|v\|_{\overline{\mathcal{B}}} \leq 1\}, \tag{3.15}$$

and use the trace norm definition of $\|\cdot\|_{H^{1/2}(\partial\Omega)}$ throughout this section.

Recall the tensor product grid $G_{k,r}$ for $\overline{\Omega} = [0,1]^d$, see (3.8). We let

$$\overline{G}_{k,r} := \{\mathbf{z}_i : \ i = 1, \ldots, \overline{m}\} = G_{k,r} \cap \partial\Omega \tag{3.16}$$

denote the set of those grid points of $G_{k,r}$ that are on the boundary $\partial\Omega$. We will use the points $\mathcal{Z} = \overline{G}_{k,r}$ to recover g. Note that the number $\bar{m}$ of data sites in $\overline{G}_{k,r}$ is

$$\overline{m} = \#(\overline{G}_{k,r}) = 2d[r2^k]^{d-1} \asymp 2^{k(d-1)}, \tag{3.17}$$

with constants of equivalency depending only on r and d.

Given $g \in \mathcal{G}$, we define the continuous piecewise polynomial

$$\overline{S}_k(g) := \sum_{i=1}^{\overline{m}} g(\mathbf{z}_i)\overline{\phi}_i, \tag{3.18}$$

where each $\overline{\phi}_i$ is the trace of the Lagrange element ϕ_i centered at $\mathbf{z}_i \in \overline{G}_{k,r}$. Note that if $v \in \overline{\mathcal{B}}$ is any function whose trace on $\partial\Omega$ is g, and $S^*_k(v)$ is its continuous piecewise polynomial Lagrange interpolant at the grid points $G_{k,r}$, see (2.16), then we have

$$Tr(S^*_k(v)) = \overline{S}_k(g).$$

This follows from the fact that each Lagrange interpolant ϕ_i centered at $\mathbf{x}_i \in G_{k,r} \setminus \overline{G}_{k,r}$ vanishes at the faces of the simplex from the corresponding Kuhn-Tucker decomposition that contains $\mathbf{x}_i$. We also want to point out that $\overline{S}_k(g)$ does not depend on v.

The proof of the following result is provided in the appendix.

Theorem 3.2. *Let $\overline{\mathcal{B}} = B^{\overline{s}}_{\overline{q}}(L_{\bar{p}}(\Omega))$ with $\bar{s} > d/\bar{p}$, $0 < \bar{p} \leq 2$, $0 < q \leq \infty$, and let $\mathcal{G} = \{Tr(v) : \|v\|_{\overline{\mathcal{B}}} \leq 1\}$. Then the optimal recovery rates of the model classe $\mathcal{G}$ in $H^{1/2}(\partial\Omega)$ is*

$$R^*_{\overline{m}}(\mathcal{G})_{H^{1/2}(\partial\Omega)} \asymp \overline{m}^{-\beta}, \quad \overline{m} \geq 1, \quad \textit{where} \quad \beta := \frac{\overline{s}-1}{d-1} - \frac{d}{d-1}\left[\frac{1}{\bar{p}} - \frac{1}{2}\right], \tag{3.19}$$

with constants of equivalency independent of $\overline{m}$.

Moreover, when $\overline{m} = \#(\overline{G}_{k,r})$, the upper bound in (3.19) is achieved when we take the data sites $\mathcal{Z} = \overline{G}_{k,r}$, provided $r > \max(s, 1)$ and the approximant to g is given by the function $\overline{S}_k(g)$.

To understand the exponent β in (3.19) better, let v be a function in $U(B^{\bar{s}}_q(L_{\bar{p}}(\Omega)))$ whose trace is g. Note that β can be rewritten as

$$\beta = \frac{\overline{s} - 1 - (\frac{d}{\bar{p}} - \frac{d}{2})}{d-1},$$

where the numerator is the excess regularity of v in $H^1(\Omega)$ and thereby corresponds to the excess regularity of g in $H^{1/2}(\partial\Omega)$, and the denominator is $d-1$ since $\partial\Omega$ has dimension $d-1$.

3.3 Optimal recovery of u

We turn now to determining the optimal $H^1(\Omega)$ recovery rate for the functions u in the model class $\mathcal{U}$ from data

$$-\Delta u(\mathbf{x}_i) = f_i, \quad \mathbf{x}_i \in \mathcal{X},\ i = 1, \dots, \widetilde{m}; \quad u(\mathbf{z}_i) = g_i, \quad \mathbf{z}_i \in \mathcal{Z},\ i = 1, \dots \overline{m}, \tag{3.20}$$

under a given budget $m = \widetilde{m} + \overline{m}$ of data observations. Recall that $\mathcal{F} := U(B^s_q(L_p(\Omega)))$, where $s > d/p$ and $0 < q, p \le \infty$, $\mathcal{G} := Tr(U(B^{\overline{s}}_{\overline{q}}(L_{\overline{p}}(\Omega)))$, where $\overline{s} > d/\overline{p}$ and $0 < \overline{p} \le 2$, and

$$\mathcal{U} := \{u \in C(\overline{\Omega}) \ :\ u \text{ satisfies (1.1) with } f \in \mathcal{F}, g \in \mathcal{G}\}.$$

The following theorem gives the optimal recovery rate for $\mathcal{U}$.

Theorem 3.3. *Let $\Omega = (0,1)^d$ and let $\mathcal{F}$, $\mathcal{G}$, $\mathcal{U}$ be as above. Then the following holds for the optimal recovery rate $R^*_m(\mathcal{U})_{H^1(\Omega)}$:*

- *if $d \ge 3$, or $d = 2$ and $p > 1$, we have*

$$R^*_m(\mathcal{U})_{H^1(\Omega)} \asymp m^{-\min(\alpha,\beta)}, \quad m \ge 2, \tag{3.21}$$

- *if $d = 2$ and $0 < p \le 1$, we have*

$$m^{-\min(\alpha,\beta)} \lesssim R^*_m(\mathcal{U})_{H^1(\Omega)} \lesssim \log(m) m^{-\alpha} + m^{-\beta}, \quad m \ge 2, \tag{3.22}$$

where $\alpha := \alpha_{-1}$ is given by (3.12)*, β is given by* (3.19)*, and all constants of equivalence are independent of m.*

Proof. Let us fix the data cites $\mathcal{X}, \mathcal{Z}$ with $|\mathcal{X}| = \widetilde{m}$, $|\mathcal{Z}| = \overline{m}$, $m = \overline{m} + \widetilde{m}$, and let $u_1, u_2 \in \mathcal{U}_{\rm data}$ with corresponding $f_1, f_2 \in \mathcal{F}_{\rm data}$ and $g_1, g_2 \in \mathcal{G}_{\rm data}$. It follows from (1.3) that

$$\|u_1 - u_2\|_{H^1(\Omega)} \asymp \|f_1 - f_2\|_{H^{-1}(\Omega)} + \|g_1 - g_2\|_{H^{1/2}(\Omega)}.$$

It follows that

$$R^*(\mathcal{U}, \mathcal{X}, \mathcal{Z})_{H^1(\Omega)} \asymp R^*(\mathcal{F}, \mathcal{X})_{H^{-1}(\Omega)} + R^*(\mathcal{G}, \mathcal{Z})_{H^{1/2}(\Omega)}. \tag{3.23}$$

Clearly, if $\overline{m} = \widetilde{m} \asymp m/2$ and using the rates from Theorem 3.1 and Theorem 3.2, we have that

$$R^*_m(\mathcal{U})_{H^1(\Omega)} \lesssim \begin{cases} m^{-\alpha} + m^{-\beta} \lesssim m^{-\min(\alpha,\beta)}, & d \ge 3, \text{ or } d = 2,\ p > 1, \\ \log(m) m^{-\alpha} + m^{-\beta}, & d = 2,\ 0 < p \le 1. \end{cases}$$

On the other hand, using the same theorems and (3.23), we obtain

$$\begin{aligned} m^{-\min(\alpha,\beta)} \lesssim \inf_{m=\overline{m}+\widetilde{m}} (\widetilde{m}^{-\alpha} + \overline{m}^{-\beta}) &\lesssim \inf_{m=\overline{m}+\widetilde{m}} \left(R^*_{\widetilde{m}}(\mathcal{F})_{H^{-1}(\Omega)} + R^*_{\overline{m}}(\mathcal{G})_{H^{1/2}(\Omega)}\right) \\ &\lesssim \inf_{m=|\mathcal{X}|+|\mathcal{Z}|} R^*(\mathcal{F}, \mathcal{X})_{H^{-1}(\Omega)} + R^*(\mathcal{G}, \mathcal{Z})_{H^{1/2}(\Omega)} \\ &\lesssim \inf_{m=|\mathcal{X}|+|\mathcal{Z}|} R^*(\mathcal{U}, \mathcal{X}, \mathcal{Z})_{H^1(\Omega)} = R^*_m(\mathcal{U})_{H^1(\Omega)}, \end{aligned}$$

and the proof is completed. ∎

3.4 Final observations on optimal recovery

This section gave the optimal recovery rates for the model classes $\mathcal{F}, \mathcal{G}, \mathcal{U}$ from $\widetilde{m}, \overline{m}, m$, data observations, respectively. While we began by considering general model class assumptions $\mathcal{F}$ and $\mathcal{G}$ for f and g to be the unit balls of Besov spaces $B^s_q(L_p)$, we found that the optimal recovery rate is the same for many of these model classes. This means that some of the model classes are superflous in that they are contained in a larger model class with the same recovery rate. It makes sense only to use these largest model classes in going forward.

Consider, for example, the model classes $\mathcal{F} = U(\mathcal{B})$ with $\mathcal{B} = B^s_q(L_p(\Omega))$ for f. We found that in the case $d \geq 3$ and $0 < p \leq \infty$, or $d = 2$ and $1 < p \leq \infty$, the optimal recovery rate of $\mathcal{F}$ in $H^{-1}(\Omega)$ is the same as its optimal recovery rate in $L_\delta(\Omega)$ with $\delta = \frac{2d}{d+2}$. Moreover, the model class $\mathcal{F}$ has an optimal recovery rate $m^{-\alpha/d}$ in $H^{-1}(\Omega)$ if and only if it is contained in $U(B^\alpha_\infty(L_\gamma(\Omega)))$ and, in addition, $\mathcal{F}$ embeds into $C(\Omega)$.

Largest Model Classes for f:

In the case $d \geq 3$, $0 < p \leq \infty$ or $d = 2$, $1 < p \leq \infty$ the class

$$\mathcal{F} := U(\mathcal{B}), \quad \mathcal{B} = B^s_\infty(L_p(\Omega)), \text{ with } p \geq \delta \text{ and } s > \frac{d}{p} \tag{3.24}$$

has the optimal recovery rate

$$R^*_m(\mathcal{F})_{H^{-1}(\Omega)} \asymp m^{-s/d}, \quad m \geq 1, \tag{3.25}$$

and any Besov model class that gives the optimal recovery rate $O(m^{-s/d})$ is contained in one of these model classes. In each of these cases, the number s represents the smoothness of $f \in \mathcal{F}$ in $L_\delta(\Omega)$. These largest model classes are pictured in Figure 3.1.

In the case $d = 2$, $0 < p \leq 1$, the largest class

$$\mathcal{F} := U(\mathcal{B}), \quad \mathcal{B} = B^s_\infty(L_1(\Omega)), \text{ with } s > 2, \tag{3.26}$$

has recovery rate

$$m^{-s/2} \lesssim R^*_m(\mathcal{F})_{H^{-1}(\Omega)} \lesssim (1 + \log(m))m^{-s/2}, \quad m \geq 1. \tag{3.27}$$

Largest Model Classes for g: We have considered the model classes $\mathcal{G} = Tr(B^{\bar{s}}_\infty(L_{\bar{p}}(\Omega))$ with $0 < \bar{p} \leq 2$ and $\bar{s} > \max\{\frac{d}{\bar{p}}, 1\}$. Thus, all of the model classes which give a given rate $O(\overline{m}^{-\alpha})$ will all be contained in one model class

$$\mathcal{G} := Tr(U(\overline{\mathcal{B}})), \quad \overline{\mathcal{B}} = B^{\bar{s}}_\infty(L_2(\Omega)), \quad \text{with } \bar{s} > d/2. \tag{3.28}$$

This class has optimal recovery rate in $H^{1/2}(\partial\Omega)$

$$R^*_m(\mathcal{G})_{H^{1/2}(\partial\Omega)} \asymp m^{-(\bar{s}-1)/(d-1)}, \quad m \geq 1. \tag{3.29}$$

Moreover, any Besov model class that gives this recovery rate is contained in one of these largest model classes. For each of these largest model classes $\bar{s} - 1$ is the excess smoothness of v in $H^1(\Omega)$ and also the excess smoothness of $g = Tr(v)$ in $H^{1/2}(\Omega)$.

Remark 3.4. *Because of the above remarks, in going further, we always take $\mathcal{F}$ as in* (3.24) *and $\mathcal{G}$ as in* (3.28).

Model Classes for u: If we use one of the model classes from (3.24) for $\mathcal{F}$ and one of the model classes from (3.28) for $\mathcal{G}$, we obtain a model class $\mathcal{U}$ for u. The optimal recovery rate for this model class is

$$R^*_m(\mathcal{U})_{H^1(\Omega)} \asymp m^{-\min\{s/d, (\bar{s}-1)/(d-1)\}}, \quad m \geq 2,$$

which could be obtained by assigning $\widetilde{m} \asymp m/2$ points in $\overline{\Omega}$ and $\overline{m} \asymp m/2$ points on the boundary $\partial\Omega$.

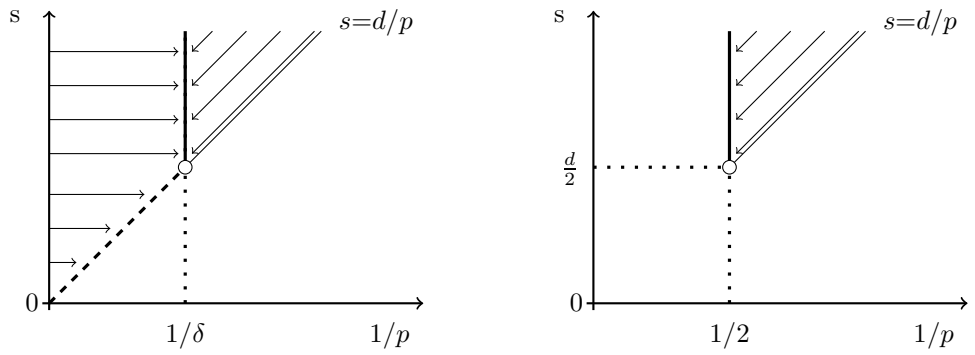

Figure 3.1: The spaces $B^s_\infty(L_p(\Omega))$ are represented by the point $(1/p, s)$ in the first quadrant. The arrows indicate embedding with a slope of d for the oblique lines. In particular, the spaces above the line $s = d/p$ embed into $C(\overline{\Omega})$. (Left) The largest model classes for f are the unit balls of spaces just above the thick and dashed segment or on the on the thick vertical half-line (excluding the point $(1/\delta, d/\delta)$). (Right) The largest model classes for g are the traces of the unit balls of spaces on the thick vertical half-line (excluding the point $(1/2, d/2)$).

4 Numerical algorithms based on linear approximation

The above analysis of optimal recovery tells us to what extent the information $(\mathbf{f}, \mathbf{g})$ determines u. That is, the optimal recovery analysis tells us that if two functions $u, \widetilde{u}$ from our model class $\mathcal{U}$ satisfy the same data, then they are close in the $H^1(\Omega)$ norm, that is

$$\|u - \widetilde{u}\|_{H^1(\Omega)} \lesssim R^*_m(\mathcal{U})_{H^1(\Omega)}.$$

However, it does not give a numerical algorithm that takes the data and creates an approximation $\widehat{u}$ to u with accuracy of the optimal recovery rate. In this section, we discuss two numerical methods which would accomplish the latter task. The methods discussed in this section are not in the form of PINNs. The latter will be discussed in §9.

We assume in this section that the data sites form a tensor product grid G of $\overline{\Omega}$. We know such data sites G provide a near optimal recovery. We have shown that if we use the data $\mathbf{f}$ at G, then we can numerically construct a piecewise polynomial interpolant $\widehat{f}$ to f at the data sites which is a near optimal (in terms of the budget $\widetilde{m}$) approximation to f in the $H^{-1}(\Omega)$ norm. Similarly, we can numerically construct a $\widehat{g}$ from the data $\mathbf{g}$ which is a near optimal recovery of g in the $H^{1/2}(\partial\Omega)$ norm.

We define $\widetilde{u}$ as the solution to (1.1) with right hand side $\widehat{f}$ and boundary value $\widehat{g}$. We know that $\widetilde{u}$ is a near optimal recovery of u in the $H^1(\Omega)$ norm. Then, in view of (1.3), we have

$$\|u - \widetilde{u}\|_{H^1(\Omega)} \lesssim R^*_m(\mathcal{U})_{H^1(\Omega)}. \tag{4.1}$$

Note that the constructions of $\widehat{f}$ and $\widehat{g}$ are numerical and we can control their complexity. In order to do that for $\widetilde{u}$, we use an existing numerical method for solving (1.1). This will, of course, incur a numerical error depending on the algorithm we choose. We mention two natural possibilities.

4.1 Using FEMs or AFEMs

Let $\widehat{u}_n$ be the approximation to $\widetilde{u}$ obtained from the Galerkin projection onto a standard finite element space V_n of dimension n. That is, $\widehat{u}_n$ is the Galerkin solution to (1.1) with right side $\widehat{f}$ and boundary value $\widehat{g}$. Then, we have

$$\|u - \widehat{u}_n\|_{H^1(\Omega)} \lesssim R^*_m(\mathcal{U})_{H^1(\Omega)} + \varepsilon_n(\mathcal{U})_{H^1(\Omega)}, \tag{4.2}$$

where

$$\varepsilon_n(\mathcal{U})_{H^1(\Omega)} := \operatorname{dist}(\mathcal{U}, V_n)_{H^1(\Omega)}, \quad n \geq 1. \tag{4.3}$$

The term ε_n will tend to zero as $n \to \infty$ and we would obtain the optimal recovery error on the right side of (4.2) if n is chosen suitably large. The actual decay rate of ε_n depends on the regularity of the model class $\mathcal{U}$ in $H^1(\Omega)$. Therefore, the issue becomes what do our model class assumptions $f \in \mathcal{F}$, $g \in \mathcal{G}$, say about the regularity of u in the $H^1(\Omega)$ norm. If we use linear spaces such as standard finite element spaces, then we would want to know the regularity of u in the scale of $H^t(\Omega)$ spaces, $t > 1$. There are several theorems that obtain such regularity results (see e.g. [11, 35, 46] and the references therein) and thereby obtain concrete bounds on the decay rate ε_n. We do not elaborate on this further but refer the interested reader to the papers cited above.

If we use nonlinear numerical methods such as AFEMs from for solving (1.1) with right side $\widehat{f}$, and boundary value $\widehat{g}$, then the decay rate for $\varepsilon_n(\mathcal{U})$ can be improved (see [5] for a summary of such results).

4.2 Reduced models

If one is faced with solving several PDEs with the same data sites $(\mathcal{X}, \mathcal{Z})$ but with different data $(\mathbf{f}, \mathbf{g})$, one can build a numerical algorithm as follows. Suppose that $\mathcal{X} = G_{k,r}$, $|G_{k,r}| = \widetilde{m} = [r2^k]^d$, see (3.8) and $\mathcal{Z} = \overline{G}_{k,r}$, $|\overline{G}_{k,r}| = \overline{m} = 2d[r2^k]^{d-1}$, see (3.16) $\overline{\Omega}$ with spacing $h = r^{-1}2^{-k}$.

We know that the simplicial interpolation operator S_k^* defined in (2.16) provides a near optimal recovery for $\mathcal{F}$. Namely,

$$\|f - S_k^*(f)\|_{H^{-1}(\Omega)} \lesssim R_{\widetilde{m}}^*(\mathcal{F})_{H^{-1}(\Omega)}, \quad \widetilde{m} = [r2^k]^d. \tag{4.4}$$

We write

$$S_k^*(f) = \sum_{i=1}^{\widetilde{m}} f(\mathbf{x}_i)\phi_i, \tag{4.5}$$

where the ϕ_i are the local Lagrange functions with $\phi_i(\mathbf{x}_i) = 1$ and $\phi_i(\mathbf{x}_j) = 0$ when $j \neq i$, supported on the simplex containing the point $\mathbf{x}_i$. Let ϕ_i^* be the solution to (1.1) with right side ϕ_i and zero boundary conditions.

Similarly, the interpolant $\overline{S}_k(g)$ to $g \in \mathcal{G}$ used for optimal recovery, see (3.18), can be written as

$$\overline{S}_k(g) = \sum_{i=1}^{\overline{m}} g(\mathbf{z}_i)\psi_i, \quad \overline{m} = 2d[r2^k]^{d-1}. \tag{4.6}$$

Let ψ_i^* be the solution to (1.1) with $g = \psi_i$ and zero right side.

Given any data $(\mathbf{f}, \mathbf{g})$, the function

$$\widetilde{u} := \sum_{i=1}^{\widetilde{m}} f(\mathbf{x}_i)\phi_i^* + \sum_{i=1}^{\overline{m}} g(\mathbf{z}_i)\psi_i^* \tag{4.7}$$

is a near optimal recovery of the solution u to (1.1) with right side f and boundary values g. Note that $\widetilde{u}$ is the solution to (1.1) with right side $S_k^*(f)$ and boundary value $\overline{S}_k(g)$.

Given a budget m, one computes offline to a sufficiently high accuracy suitable approximations $\widehat{\phi}_i$ and $\widehat{\psi}_i$ in $H^1(\Omega)$ to the ϕ_i^* and ψ_i^*, respectively. Then the function

$$\widehat{u} := \sum_{i=1}^{\widetilde{m}} f(\mathbf{x}_i)\widehat{\phi}_i + \sum_{i=1}^{\overline{m}} g(\mathbf{z}_i)\widehat{\psi}_i \tag{4.8}$$

is a near optimal recovery for the solution u to (1.1) with $f \in \mathcal{F}$, $g \in \mathcal{G}$, where $f|_{\mathcal{X}} = \mathbf{f}$ and $g|_{\mathcal{Z}} = \mathbf{g}$.

Note also that the near optimality can be guaranteed for all model class assumptions provided s and $\overline{s}$ are all less than a fixed number s_0.

5 Numerical methods based on optimization

Another approach to finding a near optimal approximation to u from the given data is through optimization, such as the one used in PINNS. To analyze this approach, we begin this section by first considering a theoretical optimization algorithm which we shall later turn into a numerical procedure. We begin by introducing the functional $\mathcal{L}_T : H^1(\Omega) \to \mathbb{R}$, defined as

$$\mathcal{L}_T(v) := \|f + \Delta v\|_{H^{-1}(\Omega)} + \|v - g\|_{H^{1/2}(\partial\Omega)}, \tag{5.1}$$

for fixed $f \in \mathcal{F} \subset H^{-1}(\Omega)$ and $g \in \mathcal{G} \subset H^{1/2}(\partial\Omega)$. Clearly, the unique solution $u \in H^1(\Omega)$ to (1.1) with right side f and boundary condition g is the unique minimizer of this functional since $\mathcal{L}_T(u) = 0$. In other words, we have

$$u = \operatorname*{argmin}_{v \in H^1(\Omega)} \mathcal{L}_T(v). \tag{5.2}$$

Note also that because of (1.3), we have for every $v \in H^1(\Omega)$

$$\|u - v\|_{H^1(\Omega)} \asymp \|f + \Delta v\|_{H^{-1}(\Omega)} + \|g - v\|_{H^{1/2}(\partial\Omega)} = \mathcal{L}_T(v), \tag{5.3}$$

with absolute constants in the equivalence. In other words, the loss $\mathcal{L}_T(v)$ for any v tells us how close v is to the true solution u.

The minimization in (5.2) is over all $v \in H^1(\Omega)$. To make this minimization more amenable to numerical implementation, we take this minimization over a set $\Sigma = \Sigma_n$, where Σ is either a linear space of dimension n or a nonlinear manifold of order n (i.e., depending on n parameters), and study any function

$$S_\Sigma \in \{\operatorname*{argmin}_{S \in \Sigma} \mathcal{L}_T(S)\}. \tag{5.4}$$

We want to understand how close such an S_Σ is to u with error measured in the norm of $H^1(\Omega)$. Therefore, we need to require that $\Sigma \subset H^1(\Omega)$.

Let us fix our model class $\mathcal{U}$ as given in (3.3). If we want S_Σ to be a good approximation to u, then we need Σ to be good at approximating the elements of $\mathcal{U}$. We define the error

$$E(\mathcal{U}, \Sigma) := \sup_{v \in \mathcal{U}} \inf_{S \in \Sigma} \|v - S\|_{H^1(\Omega)}.$$

If we return to our specific $u \in \mathcal{U}$ determined by f and g, then for any S_Σ from (5.4) and any $S \in \Sigma$, we have from (5.3) that

$$\|u - S_\Sigma\|_{H^1(\Omega)} \lesssim \mathcal{L}_T(S_\Sigma) \lesssim \mathcal{L}_T(S) \lesssim \|u - S\|_{H^1(\Omega)}, \tag{5.5}$$

where the last inequality is a consequence of the left inequality in (1.3). Since (5.5) holds for all $S \in \Sigma$, we have that

$$\|u - S_\Sigma\|_{H^1(\Omega)} \lesssim \inf_{S \in \Sigma} \|u - S\|_{H^1(\Omega)} \lesssim E(\mathcal{U}, \Sigma). \tag{5.6}$$

The discussion that we have just given shows that if we solve the continuous minimization problem (5.2) over Σ instead of all of $H^1(\Omega)$, then the solution will get as close to u as the efficiency of Σ in approximating $\mathcal{U}$, i.e., $E(\mathcal{U}, \Sigma)$. This error can be made as small as we wish by taking finer spaces for Σ, i.e., letting n increase. The error we incur is then proportional to the error in $E(\mathcal{U}, \Sigma)$ which tell us that the best candidates for Σ are those that approximate our model class $\mathcal{U}$ well.

Several issues arise that prevent the direct implementation of the above loss in our setting. The first of these is that we do not know neither f nor g. We only know these functions through the given data. The remedy for this is to introduce a surrogate for the $H^{-1}(\Omega)$ and $H^{1/2}(\partial\Omega)$ norms that use only the data information $(\mathbf{f}, \mathbf{g})$ we have about f and g. The next section addresses this issue.

6 Discretizing norms

The optimization procedure of the previous section is not a numerical algorithm since it does not incorporate numerical methods for estimating the norms appearing in the loss $\mathcal{L}_T$, namely $\|\cdot\|_{H^{-1}(\Omega)}$ and $\|\cdot\|_{H^{1/2}(\partial\Omega)}$. In this section, we address this issue by introducing discrete norms which involve only the values of functions at the data sites. Replacing the norms in $\mathcal{L}$ by these discrete norms leads to a loss which can be numerically computed. The equivalence between the norms in $\mathcal{L}_T$ and their discrete counterparts holds (modulo corresponding optimal recovery rates) for functions whose Laplacians are uniformly bounded in $\mathcal{B}$ and whose traces are uniformly bounded in $\overline{\mathcal{B}}$.

6.1 A discrete L_τ norm

In this section, we study the discretization of the L_τ norm for any $1 \le \tau \le \infty$. Let $\mathcal{F} = U(B^s_q(L_p(\Omega)))$, $s > d/p$, be a model class assumption on f. We consider the uniform grid $G_{k,r} \subset \overline{\Omega}$, see (3.8), consisting of m points, $m = |G_{k,r}| = [r2^k]^d$, $k \ge 0$ and $r > \max\{s, 1\}$. For any continuous function f and any $1 \le \tau < \infty$, we define

$$\|f\|^*_{L_\tau} := \left[\frac{1}{m}\sum_{j=1}^{m} |f(\mathbf{x}_j)|^\tau\right]^{1/\tau}, \quad \text{where} \quad \mathbf{x}_j \in G_{k,r}. \tag{6.1}$$

When $\tau = \infty$ we make the usual modification to define the $\|f\|^*_{L_\infty}$ norm.

The following lemma shows that the discrete L_τ norm is close to the true L_τ norm for functions in the model class $\mathcal{F}$.

Lemma 6.1. *Let $\mathcal{B} = B^s_q(L_p(\Omega))$, $\Omega = (0,1)^d$, be a Besov space with $s > d/p$. If $f \in \mathcal{B}$, then for any $1 \le \tau \le \infty$, we have*

$$\|f\|_{L_\tau(\Omega)} \lesssim \|f\|^*_{L_\tau} + \|f\|_{\mathcal{B}} m^{-\alpha_\tau}, \quad \text{and} \quad \|f\|^*_{L_\tau} \lesssim \|f\|_{L_\tau(\Omega)} + \|f\|_{\mathcal{B}} m^{-\alpha_\tau},$$

with $\alpha_\tau := \frac{s}{d} - \left(\frac{1}{p} - \frac{1}{\tau}\right)_+$. Recall that $m^{-\alpha_\tau}$ is the uniform rate of optimal recovery of the model class $\mathcal{F}$ in $L_\tau(\Omega)$. The constant in $\lesssim$ depends only on r and d.

Proof. We prove the lemma for $1 \le \tau < \infty$. A simple modification of the proof handles the case $\tau = \infty$. We use the simplicial interpolation operator S^*_k given in (2.16). Let $\mathcal{T}_k := \mathcal{T}_k(\Omega)$ be the set of simplices that make up the simplicial decomposition of the cubes in $\mathcal{D}_k(\overline{\Omega})$. Then, from (2.16) we have

$$\|S^*_k(f)\|^\tau_{L_\tau(\Omega)} = \sum_{T \in \mathcal{T}_k} \|L_T(f)\|^\tau_{L_\tau(T)}, \tag{6.2}$$

where $L_T(f)$ is the polynomial in $\mathcal{P}^d_r$ that interpolates the data $\{f(\mathbf{x}_i) : \mathbf{x}_i \in \overline{T}\}$. By equivalence of norms on $\mathcal{P}^d_r$, we know that

$$\|L_T(f)\|_{L_\tau(T)} \asymp \Big[\frac{1}{m}\sum_{\mathbf{x}_i \in \overline{T}} |f(\mathbf{x}_i)|^\tau\Big]^{1/\tau}, \tag{6.3}$$

with constants of equivalency depending only on d, r. Summing over $T \in \mathcal{T}_k$ the quantities in (6.3) gives

$$\|S^*_k(f)\|_{L_\tau(\Omega)} \asymp \Big[\frac{1}{m}\sum_{i=1}^{m} |f(\mathbf{x}_i)|^\tau\Big]^{1/\tau} = \|f\|^*_{L_\tau(\Omega)}. \tag{6.4}$$

For $f \in \mathcal{B} = B^s_q(L_p(\Omega)$, the piecewise polynomial interpolant $S^*_k(f)$ provides a near optimal recovery of f from the data $\mathbf{f}$, that is

$$\|f - S^*_k(f)\|_{L_\tau(\Omega)} \le C\|f\|_{\mathcal{B}} m^{-\alpha_\tau}. \tag{6.5}$$

It follows then from (6.4) that

$$\big|\|f\|_{L_\tau(\Omega)} - \|S^*_k(f)\|_{L_\tau(\Omega)}\big| \le C\|f\|_{\mathcal{B}} m^{-\alpha_\tau}, \tag{6.6}$$

and the proof is completed. ∎

6.2 Discrete $H^{1/2}(\partial\Omega)$ norms

In this section, we introduce discrete $H^{1/2}(\partial\Omega)$ norms and discuss their accuracy in computing the true $H^{1/2}(\partial\Omega)$ norm for functions g in our model class $\mathcal{G}$. The discrete norm we introduce will only uses the values of g at the data sites and can therefore be computed from the given data. We fix any $k \geq 1$ and let $\mathcal{Z} := \overline{G}_{k,r} = \{\mathbf{z}_j\}_{j=1}^{\bar{m}} = G_{k,r} \cap \partial\Omega$, see (3.16), be the grid points on the boundary $\partial\Omega$. The number of grid points in $\mathcal{Z}$ is $\overline{m} = 2d[r2^k]^{d-1}$.

We fix a model class $\mathcal{G} = Tr(U(\overline{\mathcal{B}}))$ of the form (3.28). The optimal recovery rate for this class is

$$R_m^*(\mathcal{G})_{H^{1/2}(\partial\Omega)} \asymp m^{-\beta}, \quad m \geq 1, \quad \beta = \frac{\bar{s}-1}{d-1}. \tag{6.7}$$

This also serves to fix r which we recall is the smallest integer satisfying $r > \max(\bar{s}, 1)$. For any continuous function g which is the trace of a $v \in \overline{B}$, we use the trace norm

$$\|g\|_{Tr(\overline{\mathcal{B}})} := \inf_{Tr(v)=g} \|v\|_{\overline{B}}. \tag{6.8}$$

Recall that the intrinsic semi-norm for $H^{1/2}(\partial\Omega)$ is given by

$$|g|^2_{H^{1/2}(\partial\Omega)} := \int\limits_{\partial\Omega\times\partial\Omega} \frac{|g(\mathbf{x}) - g(\mathbf{y})|^2}{|\mathbf{x}-\mathbf{y}|^d}\, d\mathbf{x}d\mathbf{y}. \tag{6.9}$$

We obtain the intrinsic $H^{1/2}(\partial\Omega)$ norm by adding $\|g\|_{L_2(\partial\Omega)}$ to this semi-norm. It is known that this definition is equivalent to the trace norm definition (6.8) for this domain Ω (see [27]). We shall use (6.9) through out this section.

Let $\overline{S}_k$ be as defined in (3.18). We have proven in Theorem 3.2 that

$$\|g - \overline{S}_k(g)\|_{H^{1/2}(\partial\Omega)} \leq C\|g\|_{Tr(\overline{\mathcal{B}})}\bar{m}^{-\beta}, \tag{6.10}$$

with C independent of $\overline{m}$ and g. This gives the comparison

$$\big|\|g\|_{H^{1/2}(\partial\Omega)} - \|\overline{S}_k(g)\|_{H^{1/2}(\partial\Omega)}\big| \leq \|g - \overline{S}_k(g)\|_{H^{1/2}(\partial\Omega)} \leq C\|g\|_{Tr(\overline{\mathcal{B}})}\bar{m}^{-\beta}.$$

We concentrate therefore on finding a discrete $H^{1/2}(\partial\Omega)$ norm for the functions $\overline{S}_k(g)$, $g \in \mathcal{G}$. Note that $\overline{S}_k(g)$ is completely determined by the values of g at the data sites $\overline{G}_{k,r}$.

Given $k \geq 0$, let $\mathcal{T}_k := \mathcal{T}_k(\partial\Omega)$ denote the collection of all $d-1$ dimensional simplicies given by the Kuhn-Tucker partition of the dyadic cubes $I \in \mathcal{D}_k(\partial\Omega)$ of the boundary $\partial\Omega$. For each dyadic cube $I \in \mathcal{D}_k(\partial\Omega)$, there are $(d-1)!$ simplices contained in I

Let us further introduce the finite dimensional linear space $V^r(\mathcal{T}_k)$ of functions that are continuous on $\partial\Omega$ and piecewise polynomial (subordinate to $\mathcal{T}_k$) of order r on $\partial\Omega$. The linear operator $\overline{S}_k$ is a projector onto $V^r(\mathcal{T}_k)$.

If g is in $C(\partial\Omega)$, we define

$$|g|^*_{H^{1/2}(\partial\Omega)} := \left[\frac{1}{\overline{m}^2}\sum_{i\neq j}\frac{|g(\mathbf{z}_i) - g(\mathbf{z}_j)|^2}{|\mathbf{z}_i - \mathbf{z}_j|^d}\right]^{1/2} \tag{6.11}$$

and prove that the the two semi-norms $|\cdot|^*_{H^{1/2}(\partial\Omega)}$ and $|\cdot|_{H^{1/2}(\partial\Omega)}$ are equivalent on $V^r(\mathcal{T}_k)$.

We begin with the following lemma.

Lemma 6.2. *For each pair $T, T' \in \mathcal{T}_k$ and each $S \in V^r(\mathcal{T}_k)$, we have*

$$\left[\frac{1}{\bar{m}^2}\sum_{\substack{i\neq j\\ \mathbf{z}_i\in\overline{T},\ \mathbf{z}_j\in\overline{T}'}}\frac{|S(\mathbf{z}_i) - S(\mathbf{z}_j)|^2}{|\mathbf{z}_i - \mathbf{z}_j|^d}\right]^{1/2} \asymp \left[\int\limits_{T\times T'}\frac{|S(\mathbf{z}) - S(\mathbf{z}')|^2}{|\mathbf{z}-\mathbf{z}'|^d}d\mathbf{z}\, d\mathbf{z}'\right]^{1/2}, \tag{6.12}$$

where the constants in $\asymp$ depend only on r and d.

Proof. We fix d, r, and a pair $T, T' \in \mathcal{T}_k := \mathcal{T}_k(\partial\Omega)$. Let us first consider the case when $k \leq 4$. Let $\mathcal{N}(S) := \mathcal{N}_{T,T'}(S)$ denote the expression in (6.12) involving the sum, and $\mathcal{N}'(S)$ the expression involving the integral. Both are semi-norms on the finite dimensional linear space $X = X(T, T')$ of functions from $V^r(\mathcal{T}_k)$ restricted to $T \cup T'$. A function $S \in X$ satisfies $\mathcal{N}(S) = 0$ if and only if S is constant on T, T' and, additionally, S is constant on $T \cup T'$ when T and T' touch. A similar statement holds for the semi-norm $\mathcal{N}'$. It follows that for any pair T, T', we have

$$c\mathcal{N}(S) \leq \mathcal{N}'(S) \leq C\mathcal{N}(S), \quad S \in X(T, T'), \tag{6.13}$$

where the constants c, C depend on T, T', r and d, but are independent of S. Since there are only a finite number (depending on d) of pairs of simplicies in $\mathcal{T}_k(\partial\Omega)$, $k \leq 4$, there are constants c_1, C_1 which depend only on r, d for which (6.13) is valid for all pairs of simplices from $\mathcal{T}_k(\partial\Omega)$, $k \leq 4$. Then (6.13) holds for all such pairs with $c = c_1$, $C = C_1$, and we have completed the proof in this case.

We now consider the case $k > 4$. Given a pair T, T', we let $c(T, T')$ be the largest number such that the lower inequality in (6.13) holds uniformly for all $S \in X(T, T')$. Similarly $C(T, T')$ is the smallest constant so that the upper inequality of (6.13) holds uniformly on $X(T, T')$. The above argument using equivalence of semi-norms shows that there always exist positive constants $c(T, T'), C(T, T')$. We are left to show that there are constants $0 < c^* < C^* < \infty$, depending only on d and r, for which $c(T, T') \geq c^*$ and $C(T, T') \leq C^*$ holds for all pairs T, T' such that $T, T' \in \mathcal{T}_k$ and $k > 4$.

We consider two cases.

Case 1: $T \cap T' \neq \emptyset$**.**

In this case there is a pair $T_0, T_0' \in \mathcal{T}_4$, and a linear mapping consisting of a translation and a dilation with factor 2^{-k+4} that rigidly transforms T_0, T_0' to the pair T, T'. If we use this linear transformation to change variables in (6.2) we obtain the validity of (6.2) for this case with the same constants c_1, C_1.

Case 2: $T \cap T' = \emptyset$**.** First note that from the L_2 discretization of the previous section, we have that there are constants c_2, C_2 depending only on r, d such that for each $S \in V^r(\mathcal{T}_k(\partial\Omega))$

$$c_2\Big[\int\limits_{T\times T'} |S(\mathbf{z}) - S(\mathbf{z}')|^2 d\mathbf{z}\, d\mathbf{z}'\Big]^{1/2} \leq \Big[\frac{1}{\overline{m}^2} \sum_{\substack{\mathbf{z}\neq\mathbf{z}'\in\mathcal{Z}\\ \mathbf{z}\in T,\ \mathbf{z}'\in T'}} |S(\mathbf{z}) - S(\mathbf{z}')|^2\Big]^{1/2} \leq C_2\Big[\int\limits_{T\times T'} |S(\mathbf{z}) - S(\mathbf{z}')|^2 d\mathbf{z}\, d\mathbf{z}'\Big]^{1/2}. \tag{6.14}$$

Now, on $T \times T'$, the expressions $|\mathbf{z} - \mathbf{z}'|$ that appear in (6.12) are all comparable with absolute constants. Therefore, we obtain (6.12) holds in this case as well with constants depending only on r, d. □

We can now prove that the discrete semi-norm and the true semi-norm of functions in $\mathcal{G}$ are comparable.

Theorem 6.3. *For any $g \in \mathcal{G} = Tr(U(\overline{B}))$, where $\overline{B} = B^{\bar{s}}_\infty(L_2(\Omega))$ with $\bar{s} > d/2$, and any $\overline{m} \geq 1$, we have*

$$|g|_{H^{1/2}(\partial\Omega)} \lesssim |g|^*_{H^{1/2}(\partial\Omega)} + \|g\|_{Tr(\overline{\mathcal{B}})}\overline{m}^{-\beta}, \quad \textit{and} \quad |g|^*_{H^{1/2}(\partial\Omega)} \lesssim |g|_{H^{1/2}(\partial\Omega)} + \|g\|_{Tr(\overline{\mathcal{B}})}\overline{m}^{-\beta}, \tag{6.15}$$

with

$$\beta = \frac{\bar{s} - 1}{d - 1}, \tag{6.16}$$

and constants in $\lesssim$ depending only on r, d. Note that the optimal recovery rate of $\mathcal{G}$ in the $H^{1/2}(\partial\Omega)$ is $\bar{m}^{-\beta}$.

Proof. For any $S \in V^r(\mathcal{T}_k)$, we have

$$|S|_{H^{1/2}(\partial\Omega)} = \left[\sum_{T,T'\in\mathcal{T}_k} \int\limits_T \int\limits_{T'} \frac{|S(\mathbf{z}) - S(\mathbf{z}')|^2}{|\mathbf{z} - \mathbf{z}'|^d}\, d\mathbf{z}\, d\mathbf{z}'\right]^{1/2}$$

and

$$|S|^*_{H^{1/2}(\partial\Omega)} = \left[\sum_{\mathbf{z}\neq\mathbf{z}'\in\overline{G}_{k,r}} \frac{1}{\overline{m}^2}\frac{|S(\mathbf{z})-S(\mathbf{z}')|^2}{|\mathbf{z}-\mathbf{z}'|^d}\right]^{1/2} \asymp \left[\sum_{T,T'\in\mathcal{T}_k}\sum_{\substack{\mathbf{z}\neq\mathbf{z}'\in\overline{G}_{k,r}\\ \mathbf{z}\in T,\ \mathbf{z}'\in T'}} \frac{1}{\overline{m}^2}\frac{|S(\mathbf{z})-S(\mathbf{z}')|^2}{|\mathbf{z}-\mathbf{z}'|^d}\right]^{1/2}.$$

Here, in the second expression, we only have equivalence since a given data site may be used for more than one simplex. Note, however that the constant in the second equivalence depends only on d and r.

In view of Lemma 6.2, we know that there exist constants $0 < c < C < \infty$ that depend only on d and r such that for each $S \in V^r(\mathcal{T}_k)$, we have

$$c|S|_{H^{1/2}(\partial\Omega)} \le |S|^*_{H^{1/2}(\partial\Omega)} \le C|S|_{H^{1/2}(\partial\Omega)}. \tag{6.17}$$

In particular this holds for $S = \overline{S}_k(g)$ whenever $g \in \mathcal{G}$. If we combine (6.17) with (6.10) and observe that $|g|^*_{H^{1/2}(\partial\Omega)} = |\overline{S}_k(g)|^*_{H^{1/2}(\partial\Omega)}$ we obtain the theorem. □

In order to define a discrete norm for $H^{1/2}(\partial\Omega)$, we will also need a discrete $L_2(\partial\Omega)$ norm for the functions in our model class $\mathcal{G}$. Using the same $\overline{m}$ data sites $\mathcal{Z} := \overline{G}_{k,r}$, we define

$$\|g\|^*_{L_2(\partial\Omega)} := \left[\frac{1}{\overline{m}}\sum_{j=1}^{\overline{m}} |g(\mathbf{z}_j)|^2\right]^{1/2}, \quad g \in \mathcal{G}, \quad \mathbf{z}_j \in \overline{G}_{k,r}. \tag{6.18}$$

Arguing as we have done in §6.1, we have for $\alpha = \frac{\bar{s}-1/2}{d-1} > \beta$ and any $g \in Tr(\overline{\mathcal{B}})$

$$\|g\|_{L_2(\partial\Omega)} \lesssim \|g\|^*_{L_2(\partial\Omega)} + \|g\|_{Tr(\overline{\mathcal{B}})}\overline{m}^{-\alpha}, \quad \text{and} \quad \|g\|^*_{L_2(\partial\Omega)} \lesssim \|g\|_{L_2(\partial\Omega)} + \|g\|_{Tr(\overline{\mathcal{B}})}\overline{m}^{-\alpha}. \tag{6.19}$$

We now define for any continuous function g the discrete $H^{1/2}(\partial\Omega)$ norm

$$\|g\|^*_{H^{1/2}(\partial\Omega)} := \|g\|^*_{L_2(\partial\Omega)} + |g|^*_{H^{1/2}(\partial\Omega)}. \tag{6.20}$$

Theorem 6.4. *For any $g \in \mathcal{G} = Tr(U(\overline{B}))$, where $\overline{B} = B^{\bar{s}}_\infty(L_2(\Omega)$ with $\bar{s} > d/2$, and any $\overline{m} \ge 1$, we have*

$$\|g\|_{H^{1/2}(\partial\Omega)} \lesssim \|g\|^*_{H^{1/2}(\partial\Omega)} + \|g\|_{Tr(\overline{\mathcal{B}})}\overline{m}^{-\beta}, \quad \textit{and} \quad \|g\|^*_{H^{1/2}(\partial\Omega)} \lesssim \|g\|_{H^{1/2}(\partial\Omega)} + \|g\|_{Tr(\overline{\mathcal{B}})}\overline{m}^{-\beta}, \tag{6.21}$$

where β is given in (6.16)*, and the constants in $\lesssim$ depend only on r and d.*

Proof. The theorem follows from the comparisons (6.15), (6.19), and the fact that $\alpha > \beta$. □

7 A discrete loss function with error control

In this section, we introduce a discrete loss function $\mathcal{L}^*$ which is a surrogate for the theoretical loss function $\mathcal{L}_T$ of (5.1). The advantage of $\mathcal{L}^*$ is that it can be computed directly from the data $(\mathbf{f}, \mathbf{g})$. Let k, r be fixed and let $\mathcal{X} = G_{k,r} \subset \overline{\Omega}$ and $\mathcal{Z} = \overline{G}_{\bar{k},r}$ be the data sites of $\partial\Omega$. Note that $\bar{k}$ could be different from k. This fixes $\widetilde{m} = \#(G_{k,r})$ and $\overline{m} := \#(\overline{G}_{\bar{k},r})$ for the remainder of this section.

Remark 7.1. *In this and the following two sections, the results and the proofs of the results are simplest in the case $d \ge 3$. When $d = 2$, the results and proofs are clouded by the appearance of log factors. For this reason, we state the results in correct form (including the log factors) for all cases $d \ge 2$ but give the exposition and proofs only for $d \ge 3$. The proofs for the case $d = 2$ are given in the Appendix, §12.7.*

Let $\|\cdot\|^*_{L_\gamma}$ be the discrete norm, see (6.3), introduced in §6.1 for the following choices of

$$\gamma = \begin{cases} \frac{2d}{d+2}, & d \geq 3, \\ \\ 1 + [\log(\widetilde{m})]^{-1}, & d = 2, \end{cases}$$

and let $\|\cdot\|^*_{H^{1/2}}$ be the discrete norm, see (6.11), for $H^{1/2}(\partial\Omega)$ introduced in §6.2. Given the data vectors for $(\mathbf{f}, \mathbf{g})$, we introduce the discrete loss function

$$\mathcal{L}^*(v) := \begin{cases} \|f + \Delta v\|^*_{L_\gamma(\Omega)} + \|g - v\|^*_{H^{1/2}(\partial\Omega)}, & d \geq 3, \\ \\ [1 + \log(\widetilde{m})]\|f + \Delta v\|^*_{L_\gamma(\Omega)} + \|g - v\|^*_{H^{1/2}(\partial\Omega)}, & d = 2, \end{cases} \tag{7.1}$$

which is defined whenever v and Δv are continuous on $\overline{\Omega}$. Notice that to compute $\mathcal{L}^*(v)$ we use only the values of v at the data sites.

We want to show that making $\mathcal{L}^*(v)$ small guarantees that v is a good approximation to u in the $H^1(\Omega)$ norm. As in [4], such a result requires model class assumptions on f and g. Keeping in mind the remarks of §3.4, we make the model class assumptions $\mathcal{F} = U(\mathcal{B})$ of (3.24) and $\mathcal{G} = Tr(U(\overline{\mathcal{B}}))$ of (3.28).

Recall that, in the case $d \geq 3$ and all $0 < p \leq \infty$, or $d = 2$ and $1 < p \leq \infty$, for a model class $\mathcal{F}$ of the form (3.24), the uniform optimal recovery rate for $\mathcal{F}$ in $H^{-1}(\Omega)$ is $\asymp \widetilde{m}^{-s/d}$, $\widetilde{m} \geq 1$ where s is the excess regularity of $\mathcal{F}$ in $L_\delta(\Omega)$. For a model class $\mathcal{G}$ of the form (3.28), its uniform optimal recovery rate in $H^{1/2}(\partial\Omega)$ is $\asymp \overline{m}^{-(\bar{s}-1)/(d-1)}$,where $\bar{s} - 1$ is the excess regularity in $H^{1/2}(\partial\Omega)$. These model class assumptions on f, g imply that the solution u to (1.1) is in a model class $\mathcal{U}$ which has a uniform optimal recovery rate $\max\{\widetilde{m}^{-\frac{s}{d}}, \overline{m}^{-(\bar{s}-1)/(d-1)}\}$. We introduce the following notation for a function v

$$\|v\|_{\mathcal{U}} := \max\{\|\Delta v\|_{\mathcal{B}}, \|Tr(v)\|_{Tr(\overline{\mathcal{B}})}\}. \tag{7.2}$$

The following theorem bounds the error $\|u - v\|_{H^1(\Omega)}$ in terms of the discrete loss $\mathcal{L}^*(v)$, provided $\Delta v \in \mathcal{B}$ and $v \in \overline{\mathcal{B}}$.

Theorem 7.2. *Let u be the solution to (1.1) with $f \in \mathcal{F} = U(\mathcal{B})$ of (3.24) and $g \in \mathcal{G} = Tr(U(\overline{\mathcal{B}}))$ of (3.28). Given the data $(\mathbf{f}, \mathbf{g})$ of f and g at grid data sites $(G_{k,r}, \overline{G}_{\bar{k},r})$ with $|G_{k,r}| = \widetilde{m}$ and $|\overline{G}_{\bar{k},r}| = \overline{m}$, consider the functional $\mathcal{L}^*$, defined in (7.1). If v is any continuous function in $H^1(\Omega)$, then*

$$\|u - v\|_{H^1(\Omega)} \lesssim \mathcal{L}^*(v) + [1 + \|v\|_{\mathcal{U}}]\, \mathcal{R}_{\mathcal{U}}(\widetilde{m}, \overline{m}), \tag{7.3}$$

with the constants in $\lesssim$ independent of u, v, $\widetilde{m}$ and $\overline{m}$, where

$$\mathcal{R}_{\mathcal{U}}(\widetilde{m}, \overline{m}) := \begin{cases} \max\{\widetilde{m}^{-\frac{s}{d}}, \overline{m}^{-\frac{\bar{s}-1}{d-1}}\}, & d \geq 3, \\ \max\{\log(\widetilde{m})\widetilde{m}^{-\frac{s}{2}}, \overline{m}^{-(\bar{s}-1)}\}, & d = 2. \end{cases} \tag{7.4}$$

Proof. We only prove this for $d \geq 3$. From (1.3), we have

$$\begin{aligned} \|u - v\|_{H^1(\Omega)} &\lesssim \|f + \Delta v\|_{H^{-1}(\Omega)} + \|g - Tr(v)\|_{H^{1/2}(\partial\Omega)} \lesssim \|f + \Delta v\|_{L_\gamma(\Omega)} + \|g - Tr(v)\|_{H^{1/2}(\partial\Omega)} \\ &\lesssim \left[\|f + \Delta v\|^*_{L_\gamma(\Omega)} + \|g - Tr(v)\|^*_{H^{1/2}(\Omega)}\right] + \left[\|f + \Delta v\|_{\mathcal{B}}\widetilde{m}^{-\frac{s}{d}} + \|g - Tr(v)\|_{Tr(\overline{\mathcal{B}})}\overline{m}^{-\frac{(\bar{s}-1)}{d-1}}\right], \end{aligned}$$

where in the second inequality we used the continuous embedding of $L_\gamma(\Omega)$ into $H^{-1}(\Omega)$ and in the third inequality we used the comparisons between continuous and discrete norms (see Lemma 6.1 and Theorem 6.3). Finally, if we use the facts that $\|f\|_{\mathcal{B}} \leq 1$ and $\|g\|_{Tr(\overline{\mathcal{B}})} \leq 1$, we complete the proof of the theorem. □

Let us make some remarks on this theorem.

Remark 7.3. *If we knew in advance the smoothness class of f, or at least the value of p, and if this value were $1 < p \leq \infty$, then in the case $d = 2$ we could modify appropriately the the loss $\mathcal{L}^*$ so that it is independent of $\widetilde{m}$. This choice would lead to the actual optimal recovery rates $\max\{\widetilde{m}^{-\frac{s}{2}}, \overline{m}^{-(\bar{s}-1)}\}$ for this class (i.e the logarithm will not be present), see the Appendix, §12.7.*

Remark 7.4. *It is natural that $\|v\|_{\mathcal{U}}$ enters into the error bounds since we need some control on v away from the data. However, this means that when we try to obtain a good approximation to u (in $H^1(\Omega)$), through optimization, then the norm $\|\cdot\|_{\mathcal{U}}$ will have to enter the picture [4].*

Remark 7.5. *This error estimator can be used to monitor the error in any optimization scheme that attempts to minimize the loss $\mathcal{L}^*(v)$. For example, suppose that $\Sigma = \Sigma_n$ is a linear space of dimension n or a nonlinear manifold determined by n parameters such that each $S \in \Sigma$ is continuous and has a continuous Laplacian ΔS. If at any stage of the optimization procedure we have an $S \in \Sigma_n$ then we can use (7.3) to bound the error of $\|u - S\|_{H^1(\Omega)}$. Of course, a good estimate requires a bound for $\|S\|_{\mathcal{U}}$.*

Remark 7.6. *Note that for fixed values of $\widetilde{m}$ and $\overline{m}$, the losses $\mathcal{L}$ and $\mathcal{L}^*$ are equivalent up to a constant depending upon the parameters $\widetilde{m}$ and $\overline{m}$. This means that if the original loss $\mathcal{L}$ is driven to 0, i.e. we are perfectly interpolating the data, then both losses give the same error control. However, in the situation where the data is not interpolated, the new loss $\mathcal{L}^*$ gives control on the error according to Theorem 7.2, while the original PINNs loss $\mathcal{L}$ does not.*

In view of the above remarks, we see that the effectiveness of using Σ_n together with the discrete loss $\mathcal{L}^*$ to provide an approximation to u is not simply governed by the error $E(\mathcal{U}, \Sigma_n)_{H^1(\Omega)}$ of approximating the class $\mathcal{U}$ by Σ_n in the $H^1(\Omega)$ but rather by a form of restricted approximation which involves $\|\cdot\|_{\mathcal{U}}$ [4].

8 A numerical optimal recovery algorithm

Let $\mathcal{U}$ be the model class (3.3) determined by the model classes $\mathcal{F}$ and $\mathcal{G}$, where $\mathcal{F}$ is a maximal model class (3.24), (3.26) and $\mathcal{G}$ is a maximal model class (3.28). To build a numerical algorithm for the recovery of $u \in \mathcal{U}_{\text{data}}$, we want to minimize the discrete loss $\mathcal{L}^*$ over a suitable set Σ_n. Let $M \geq 1$ and let $\widetilde{u}$ be such that

$$\widetilde{u} \in \{\operatorname*{argmin}_{\|v\|_{\mathcal{U}} \leq M} \mathcal{L}^*(v)\}. \tag{8.1}$$

Clearly, u is a solution to this problem since $\|u\|_{\mathcal{U}} \leq 1$ and $\mathcal{L}^*(u) = 0$. Thus, $\mathcal{L}^*(\widetilde{u}) = 0$ and it follows from (7.3) that

$$\|u - \widetilde{u}\|_{H^1(\Omega)} \lesssim M\mathcal{R}_{\mathcal{U}}(\tilde{m}, \bar{m}), \quad m \geq 1. \tag{8.2}$$

with the implied constant depending only on r and d. In other words, $\widetilde{u}$ provides a near optimal recovery when $d \geq 3$ and a near optimal recovery up to a $\log m$ factor when $d = 2$. Of course, the minimization problem (8.1) is taken over too large of a set to be numerically viable and so we would like to utilize minimization over a smaller set.

Let $\Sigma_n \subset H^1(\Omega)$ be an approximating set. Here, the primary examples for Σ_n are linear spaces of finite dimension n or a nonlinear manifolds depending on n parameters such as NNs. For $M \geq 1$ (typically taken as some fixed number not dependent on n) we define the set

$$\Sigma_n(M) := \{S \in \Sigma_n : \|S\|_{\mathcal{U}} \leq M\} \subset \Sigma_n, \tag{8.3}$$

and consider the solution $\widehat{u}$ of the minimization problem

$$\widehat{S} \in \{\operatorname*{argmin}_{S \in \Sigma_n(M)} \mathcal{L}^*(S)\}. \tag{8.4}$$

We assume for the moment that $d \geq 3$ and show that $\widehat{S}$ is a near optimal recovery of u from the given data provided the set $\Sigma_n(M)$ is sufficiently good at approximating the elements of our model class $\mathcal{U}$. Later, we explain the changes needed for the case $d = 2$.

In order to present our results, we denote by

$$E(v,\Sigma_n(M)) := \begin{cases} \inf_{S\in\Sigma_n(M)} \left(\|\Delta v - \Delta S\|_{L_\gamma(\Omega)} + \|Tr(v-S)\|_{H^{1/2}(\partial\Omega)}\right) & d\geq 3, \\ \inf_{S\in\Sigma_n(M)} \left((1+\log(\widetilde{m}))\|\Delta v - \Delta S\|_{L_\gamma(\Omega)} + \|Tr(v-S)\|_{H^{1/2}(\partial\Omega)}\right) & d=2. \end{cases} \tag{8.5}$$

the error of simultaneously approximating v and Δv. We remark that in the case $d=2$ and $1<p\leq\infty$, then this analysis can be modified to remove the logarithm. Then, for the model class $\mathcal{U}$ of $H^1(\Omega)$, we denote by

$$E(\mathcal{U},\Sigma_n(M)) := \sup_{v\in\mathcal{U}} E(v,\Sigma_n(M)),$$

the *error of restricted approximation* of the class $\mathcal{U}$ by elements from $\Sigma_n(M)$.

We have the following theorem.

Theorem 8.1. *Let Σ_n be a set in $H^1(\Omega)$ and let $\Sigma_n(M)$ be defined by* (8.3)*for a fixed value $M\geq 1$. If $\widehat{S}$ is any function from* (8.4) *then*

$$\|u-\widehat{S}\|_{H^1(\Omega)} \lesssim M\mathcal{R}_{\mathcal{U}}(\widetilde{m},\overline{m}) + E(u,\Sigma_n(M)), \tag{8.6}$$

with the constant in $\lesssim$ depending only on r,d. Moreover, if the approximation set $\Sigma_n(M)$ is such that $E(\mathcal{U},\Sigma_n(M))\leq CR_{\mathcal{U}}(\widetilde{m},\overline{m})$ for some C depending only on r,d, then $\widehat{S}$ provides a uniform near optimal recovery (up to a logarithmic factor in the case $d=2$) for the model class $\mathcal{U}$.

Proof. We only prove this in the case $d\geq 3$. In the case $d=2$ the same modifications can be made as before. We use (7.3) for $\widehat{S}$ to obtain

$$\|u-\widehat{S}\|_{H^1(\Omega)} \lesssim \mathcal{L}^*(\widehat{S}) + (1+M)\mathcal{R}_{\mathcal{U}}(\widetilde{m},\overline{m}). \tag{8.7}$$

For all $S\in\Sigma_n(M)$, Lemma 6.1 and Theorem 6.4 give the bound

$$\mathcal{L}^*(\widehat{S}) \leq \mathcal{L}^*(S) \lesssim \|\Delta u - \Delta S\|_{L_\gamma(\Omega)} + \|Tr(u-S)\|_{H^{1/2}(\partial\Omega)} + CM\mathcal{R}_{\mathcal{U}}(\widetilde{m},\overline{m}),$$

with the constant in $\lesssim$ depending only on r,d. Taking an infimum over all $S\in\Sigma_M$ gives

$$\mathcal{L}^*(\widehat{S}) \lesssim E(u,\Sigma_n(M)) + M\mathcal{R}_{\mathcal{U}}(\widetilde{m},\overline{m}). \tag{8.8}$$

When this is inserted in (8.7) we obtain the theorem. □

9 CPINNs

The previous section has shown that any minimizer of the loss $\mathcal{L}^*$ over $\Sigma_n(M)$ provides a near optimal recovery of u provided $\Sigma_n(M)$ is sufficiently good at approximating the elements of the model class $\mathcal{U}$. In this section, we specialize these results to the case where $\Sigma := \Sigma_n$ is a space generated by neural networks with n parameters and a prescribed architecture. As a primary example, we can take the activation function σ to be ReLU^k with $k\geq 3$. This guarantees that the elements in Σ_n have continuous Laplacians, i.e., ΔS is continuous whenever $S\in\Sigma_n$. This also guarantees that $\mathcal{L}^*(S)$ is well defined for each $S\in\Sigma_n$. Regarding the architecture of the network, we consider the case of deep networks which is the space usually utilized in PINNs. A similar discussion applies to shallow networks.

Consistent PINNs (CPINNs): *We call any algorithm which minimizes the loss $\mathcal{L}^*$ over a neural network space Σ_n* **consistent PINNs** *and denote it by the acronym* **CPINNs**. *The discussion in Remark 7.6 has given the advantages of using* **CPINNs** *over standard PINNs with the loss $\mathcal{L}$. Here, we do not necessarily require the minimization take place over the constrained space $\Sigma_n(M)$, unless explicitly stated since the latter invokes numerical difficulties in implementation.*

We do not address the question of designing an optimization algorithm that is guaranteed to come close to minimizing the loss $\mathcal{L}^*$ over Σ_n or $\Sigma_n(M)$. We know that the lack of such theoretical guarantees is a bottleneck in rigorously proving convergence results. We analyze the performance of any minimizer or near minimizer of the loss without addressing how such a minimizer is found.

Let $\widehat{u}_n$ denote any minimizer of (8.4). Theorem 8.1 guarantees that $\widehat{u}_n$ is a near optimal recovery of u provided that for the model class $\mathcal{U}$, we can make $E(\mathcal{U}, \Sigma_n(M))$ small by taking n large. Thus, the question is whether Σ_n has favorable restricted approximation properties. There is a large literature of approximation theoretic results concerning how efficiently deep neural networks can approximate various classes of functions (see [19] and the references therein). Let us briefly discuss the specific case of approximating functions from Besov spaces using deep ReLU neural networks. When error is measured in the L_p-norm, optimal rates have been obtained [55, 64, 52]. There has also been recent work on the problem when error is measured in another Besov space, although in this case a smoother activation function must be taken [63, 62].

On the other hand, the question of restricted approximation that is used in the present paper has seemingly not been studied. We conjecture that the techniques developed in [55, 64, 52] can be appropriately modified to give the same rates for restricted approximation with a fixed suitably large value for M (independent of n). Since the proofs are already quite difficult, we leave the proof of this conjecture to future work.

10 Numerical Illustrations

In this section, we present some numerical experiments in support of the theoretical results of the previous sections. The main take away from these theoretical results is that when using optimization of a prescribed loss function over a set Σ_n to numerically compute an approximation $\widehat{u}$ to the solution u of (1.1), then the loss function $\mathcal{L}^*$ given by (1.9) as in **CPINNs** is the proper choice. The purpose of the numerical experiments of this section is to analyze the effect of using $\mathcal{L}^*$ rather than the standard loss $\mathcal{L}$ when taking Σ_n as a neural network space.

As a prelude to presenting our numerical experiments, we make some remarks on minimization of loss functions over Σ_n that are well known to practitioners. Since both losses are non convex when viewed as a function of the parameters of the neural network space Σ_n, there is not a known mathematical procedure which is proven to converge to a minimizer of the loss. Typical approaches used in practice, try various initial guesses of the parameters and use variations of gradient descent with variable step sizes, i.e., learning rates, in an attempt to find a minimizer of the loss. There is no a priori guarantee that any specific recipe will work. In other words, one cannot blindly apply any prescribed procedure with an a priori guarantee of success.

Another drawback of this approach to solving the PDE is that one cannot test its efficacy unless the true solution is in some sense known so that one can measure performance a posteriori. One advantage of using the loss $\mathcal{L}^*$ in place of $\mathcal{L}$ is that our theory provides an a posteriori bound of the error (see (7.3)) that does not require knowledge of the solution u. In other words, when using $\mathcal{L}^*$, one can guarantee an a posterior bound of the error by computing $\mathcal{L}^*(S)$, where S is the output of the numerical optimization procedure. Indeed, one would compute $\|S\|_{\mathcal{U}}$ and from (7.3) we know

$$\|u - S\|_{H^1(\Omega)} \lesssim \mathcal{L}^*(S) + (1 + \|S\|_{\mathcal{U}}) R_{\mathcal{U}}(\widetilde{m}, \overline{m}). \tag{10.1}$$

Here the second term is the recovery error. All quantities on the right side of (10.1) only involve S and so one has a bound for the error without knowing u.

Let us also note that in practice, for the PDE setting and in other learning scenarios, one does not try to impose the model class restriction in the optimization routine since it is computationally too expensive. We follow this paradigm in our numerical procedures.

Finally, we want to point out that the minimization of the new loss $\mathcal{L}^*$ differs from the standard loss $\mathcal{L}$ in advocating the use of:

- a discrete $L_\gamma(\Omega)$ norm (we take $\gamma = 1.1$ when $d = 2$) rather than a discrete $L_2(\Omega)$ for the PDE residual;

- a discrete $H^{1/2}(\partial\Omega)$ norm rather than a discrete $L_2(\partial\Omega)$ norm for the boundary condition residual.

Although it is not possible to completely separate the effect of these two modifications, we consider two different test cases corresponding to a harmonic solution u with non-vanishing boundary conditions and a singular solution with vanishing boundary conditions; see Sections 10.2. In each case, we compare the results of the minimization of the new loss function (referred as CPINNs) and the standard loss (referred as PINNs).

10.1 Practical Algorithm

We now describe the practical algorithm implemented for the minimization of the losses $\mathcal{L}^*$ and $\mathcal{L}$. We remark that PINNs algorithms rely on the inherent assumption that the minimization processes are successful. As noted above, there is no prescribed recipe for the initial choice of parameters, the activation functions, the optimization algorithm and the hyperparameters that will guarantee convergence. Rather, one is left to experimenting with different choices and employing different heuristics until satisfactory convergence of this method is obtained.

There is no way to explore all of the possible minimization algorithms. Moreover, once a particular numerical optimization procedure is chosen to implement, one is always subjected to the possibility that some other choice would have been better and demonstrated different results. So,our numerical results must be viewed from this perspective.

Our approach in this section is to find one procedure that works reasonably well in minimizing the loss (1.9). We remark that in order to conform the common practice of minimizing the squared residual in PINNs, we actually used a loss function obtained by squaring each of the terms in (1.9) and (1.8) in our experiments, i.e. we minimize the losses (sq is shorthand for squared)

$$\mathcal{L}^*_{sq} := \left[\frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} |f(\mathbf{x}_i) + \Delta v(\mathbf{x}_i)|^\gamma\right]^{2/\gamma} + \left[\frac{1}{\overline{m}^2}\sum_{\substack{i,j=1\\ i\neq j}}^{\overline{m}} \frac{|[g-v](\mathbf{z}_i) - [g-v](\mathbf{z}_j)|^2}{|\mathbf{z}_i - \mathbf{z}_j|^2}\right] + \left[\frac{1}{\overline{m}}\sum_{j=1}^{\overline{m}} |g(\mathbf{z}_j) - v(\mathbf{z}_j)|^2\right],$$

and

$$\mathcal{L}_{sq} := \left[\frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} [\Delta v(\mathbf{x}_i) + f(\mathbf{x}_i)]^2\right] + \left[\frac{1}{\overline{m}}\sum_{i=1}^{\overline{m}} [v(\mathbf{z}_i) - g(\mathbf{z}_i)]^2\right].$$

Here we take a uniform grid of collocation points and $\widetilde{m}$ is the total number of collocation points (where we penalize the residual of the equation), while $\overline{m}$ is the number of collocation points which lie on the boundary (where we penalize the residual of the boundary values).

We then apply the exact same procedure to minimize both the original PINNs loss $\mathcal{L}_{sq}$ and the CPINNs loss $\mathcal{L}^*_{sq}$. We observe from the final loss values that our procedure obtains sufficient minimization of both loss functions, so that our results enable us to compare the effect of using the two different loss functions $\mathcal{L}_{sq}$ and $\mathcal{L}^*_{sq}$. These results demonstrate a significant improvement in accuracy when minimizing the loss $\mathcal{L}^*_{sq}$ instead of $\mathcal{L}_{sq}$.

Finally, we remark that the running time of the training the we have implemented is very high and is not remotely competitive with traditional numerical methods for this problem. Improving the efficiency of the training of PINNs on both $\mathcal{L}_{sq}$ and $\mathcal{L}^*_{sq}$ is an important and active area of research (see for instance [44, 14, 37, 48, 66]), and incorporating these recently developed techniques into our algorithm is an area of future work.

In our experiments, we use a deep neural network with the following structure. We use a deep network with $L = 8$ hidden layers of width $W = 100$. The first layer uses a tanh activation function, while the remaining layers are residual layers [33] with the ReLU^3 activation function. The network is initialized as follows:

- The parameters in the first layer are randomly initialized from a normal distribution with variance 1.

- For the residual layers with ReLU^3 activation function, the parameters are randomly initialized from a normal distribution with variance
$$\sqrt{2}\left(\frac{2}{15}\right)^{1/6}\frac{1}{\sqrt{LW}}.$$
In analogy with Xavier and He initialization [26, 32], this ensures that the variance of the output of each layer gets multiplied by a factor of $(1+2/L)$ in each residual layer, which guarantees that the expected magnitude of the neural network function at initialization is bounded.

- In the final layer the weights are initialized randomly from a normal distribution with variance $1/\sqrt{W}$, which ensures that the network output has the same variance as the output of the final hidden layer.

We train the networks with gradient descent with an initial step size of 10^{-3} decreasing by a factor of 2 every 4000 steps and momentum 0.9 for a total of $40,000$ steps. We also rescale the velocities (i.e. weighted averages of the previous gradients according to the momentum) in each layer to have the same maximum as the parameters in that layer, we have found that this makes the training somewhat more stable and improves the final loss values obtained. The full code and all experimental details can be found at https://github.com/jwsiegel2510/consistent-PINNs.

We also note that although the $H^{1/2}(\partial\Omega)$ norm is nonlocal, the implementation of the discrete $H^{1/2}(\partial\Omega)$-norm present in $\mathcal{L}^*_{sq}$ can easily be vectorized and is fairly efficient in our practical implementation. Of course, the computation and optimization of the discrete $H^{1/2}(\partial\Omega)$-norm scales quadratically with the number of collocation points on the boundary, while the norm used in the standard loss $\mathcal{L}_{sq}$ scales only linearly. In our examples, the number of collocation points in the interior scales quadratically with the number of points on the boundary, so that using the discrete $H^{1/2}(\partial\Omega)$-norm on the boundary is comparably expensive to the penalty on the interior. In addition, the evaluation of the neural network at the collocation points typically dominates the calculation, rather than the calculation of the norms or their gradients from these values. However, when using a large number of collocation points on the boundary optimizing the new loss function $\mathcal{L}^*_{sq}$ may be less efficient than optimizing the original $\mathcal{L}_{sq}$ loss. This can potentially be mitigated by using a truncated version of the discrete $H^{1/2}(\partial\Omega)$-norm, but we leave this to future work.

10.2 Results

In our first experiment, we solve the Poisson equation

$$-\Delta u = f \text{ in } \Omega,\ u = g \text{ on } \partial\Omega, \tag{10.2}$$

where f and g are chosen so that the exact solution is given by

$$u(x,y) = e^x\cos(\pi y), \qquad (x,y)\in\Omega = (0,1)^2.$$

For this solution u, we have $f=0$ and the boundary values g are non-zero. The results of this experiment with different numbers of collocation points $\widetilde{m}$ can be found in Table 1.

In our second experiment, we solve the Poisson equation (10.2) where f and g are chosen so that the exact solution is given by

$$u(x,y) = 1000x(1-x)y(1-y)\left((x-0.5)^2+(y-0.5)^2\right)^{9/4}, \qquad (x,y)\in(0,1)^2.$$

This solution is chosen such that it is not smooth (although its second derivatives are still continuous) to test the ability of the algorithm to handle lower regularity. The results of this experiment can be found in Table 2.

We remark that the H^1 relative errors in both experiments are reported with respect to the discrete $H^1(\Omega)$-norm

$$\|v\|^*_{H^1(\Omega)} := \sqrt{\left(\|v\|^*_{L_2(\Omega)}\right)^2 + \left(\||\nabla v|\|^*_{L_2(\Omega)}\right)^2}$$

$\widetilde{m}$	$\overline{m}$	CPINNs H^1 Rel error	CPINNs Loss $\mathcal{L}^*_{sq}$	PINNs H^1 Rel error	PINNs Loss $\mathcal{L}_{sq}$
25	16	2.02%	$2.4 \cdot 10^{-5}$	1.77%	$9 \cdot 10^{-6}$
100	36	0.15%	$3 \cdot 10^{-6}$	0.39%	$5 \cdot 10^{-6}$
225	56	0.06%	$2 \cdot 10^{-6}$	0.55%	$1.9 \cdot 10^{-5}$
400	76	0.07%	$6 \cdot 10^{-6}$	0.58%	$2.5 \cdot 10^{-5}$
625	96	0.11%	$1.1 \cdot 10^{-5}$	0.48%	$2 \cdot 10^{-5}$
900	116	0.07%	$5 \cdot 10^{-6}$	0.34%	$1.3 \cdot 10^{-5}$

Table 1: Results of the numerical experiments for the harmonic solution u. We see that the CPINNs loss function $\mathcal{L}^*_{sq}$ results in about $3-5$ times lower error compared with the original PINNs loss $\mathcal{L}_{sq}$.

using 500 points in each direction.

From these results, we see that using the new loss function $\mathcal{L}^*_{sq}$ consistently gives errors that are about $3-10$ times smaller compared with the original PINNs loss. From the final loss values achieved, we conclude that this is not due to the training algorithm performing better on the new loss, but is rather due to the fact that $\mathcal{L}^*_{sq}$ gives a better control on the solution error than $\mathcal{L}_{sq}$. We remark that by weighting the terms in $\mathcal{L}_{sq}$ differently (i.e. by choosing different values of λ in (1.7)), one may be able to improve the results for the original PINNs loss, as we have not rigorously tested this. However, our method does not require such tuning as all terms in $\mathcal{L}^*_{sq}$ should naturally carry the same weight.

We also remark that in our first experiments, the solution is very smooth. This has the consequence that the optimization error is the dominant source of error once $\widetilde{m} \geq 100$, i.e. in this regime the sampling error is already significantly smaller than the optimization error. This is why the error does not continue to decrease as the number of collocation points is increased beyond this point. By using a more effective optimization method which drives the loss $\mathcal{L}^*_{sq}$ this balance can be shifted and lower error can be obtained. On the other hand, in our second experiment, the solution u is not particularly smooth. In this case, the sampling error is dominant all the way up to $\widetilde{m} = 900$ collocation points and we see the error of our method decrease all the way down to this point. This is due to the fact that the optimal sampling rate decreases much more slowly because of the non-smoothness of u.

$\widetilde{m}$	$\overline{m}$	CPINNs H^1 Rel error	CPINNs Loss $\mathcal{L}^*_{sq}$	CPINNs H^1 Rel error	CPINNs Loss $\mathcal{L}_{sq}$
25	16	52.4%	$1.4 \cdot 10^{-4}$	46.7%	$2.6 \cdot 10^{-3}$
100	36	22.3%	$1.4 \cdot 10^{-3}$	28.0%	$6.6 \cdot 10^{-3}$
225	56	9.6%	$4.2 \cdot 10^{-3}$	26.7%	$1.6 \cdot 10^{-2}$
400	76	7.2%	$5.9 \cdot 10^{-3}$	19.2%	$7.7 \cdot 10^{-3}$
625	96	5.5%	$6.8 \cdot 10^{-3}$	25.4%	$2.3 \cdot 10^{-2}$
900	116	3.9%	$3.5 \cdot 10^{-3}$	23.6%	$2.8 \cdot 10^{-2}$

Table 2: Results of the numerical experiments for the non-smooth solution u. We see that the CPINNs loss function $\mathcal{L}^*_{sq}$ results in about $5-10$ times lower error compared with the original PINNs loss $\mathcal{L}_{sq}$.

11 Concluding Remarks

We have investigated collocation methods for solving elliptic PDEs of the form (1.1). A collocation method seeks to generate an approximation to the true solution u of (1.1) from point value information of the right side f and the boundary value g. Under model class assumptions of the form $f \in \mathcal{F}$ and $g \in \mathcal{G}$ which are both unit balls of Besov spaces, we showed that there is an optimal error (called the error of optimal recovery) in recovering u in the $H^1(\Omega)$ norm. We determined this optimal recovery error for the various model classes.

We then turned to the study of theoretical and numerical algorithms which would yield an optimal recovery. The most prominent of these methods is to minimize a certain loss function $\mathcal{L}$ that measures how well we have fit the data observations. We have shown that the typical loss function used in practice (based on least squares minimization) is not the correct loss and have proposed the correct loss function $\mathcal{L}^*$ defined in (7.1). We then showed that any minimizer of the correct loss $\mathcal{L}^*$ over a suitable set $\Sigma_n(M)$ yields a near optimal recovery of u.

We have then gone on to show that a modfication of PINNs which we call Consistent PINNs (CPINNs) leads to a provable near optimal recovery algorithm. The modifications needed are to use the loss $\mathcal{L}^*$ in place of the loss $\mathcal{L}$ usually used in PINNs. Secondly, the minimization needs to be taken over the constrained set $\Sigma_n(M)$ rather than all of Σ_n. Finally, we provide numerical experiments which demonstrate that the CPINNs loss function significantly outperforms the original PINNs loss function on elliptic PDEs.

We now wish to make some further comments that one can view as serious caveats to the above theory. The first and most serious objection to the above development is that we do not provide a numerical algorithm for finding a minimizer or near minimizer of the loss $\mathcal{L}^*$ over $\Sigma_n(M)$. This is the same objection that can be made for almost all learning problems and certainly for PINNs. The usual numerical method used to try and solve the minimization problem in PINNs, as well as other learning problems, is to use gradient descent or some modification of gradient descent, to solve the minimization of the loss. Unfortunately, there is no proof of convergence or accuracy when applying gradient descent to non-convex losses such as $\mathcal{L}$ or $\mathcal{L}^*$. This has not impeded the use of gradient descent in practice with much empirical success. We have utilized such techniques as well in our numerical examples.

We do not want to enter into a discussion of the art of using gradient descent in optimization. However, we want to address two issues that arise relevant to this paper. The first of these is the fact that we must optimize the loss over $\Sigma_n(M)$ rather than Σ_n itself to gain the proof of optimality. This is a serious numerical burden. We have ignored this issue in our numerical experiments much like one ignores the lack of a provable gradient descent algorithm. One could implement the restricted approximation term by adding a penalty to the loss $\mathcal{L}^*$ to guarantee that at each step of the optimization one remains in the constrained set $\Sigma_n(M)$ (see [4]) but this does not ease the computational burden and one is still faced with a lack of convergence of the numerical optimization.

Finally, let us point out another issue which is understood by practitioners but not usually pointed out in the theory. Neural network spaces are a very unstable manifold. By this we mean that changing parameters very slightly can have a huge effect on the output of the neural network. That is, the mapping from parameters to output in neural networks is very unstable. This instability appears more in deep networks than in shallow networks. Avoiding this lack of stability is an art in practice that manifests itself in judiciously choosing the starting parameters and the learning rates (step size in gradient descent). The theoretical aspects of this lack of stability, and how to avoid the instability theoretically, is addressed in a series of recent papers [8, 29, 28, 30]. There remains a serious gap between theoretical algorithms for learning and their numerical implementation.

12 Appendix

In this appendix, we provide proof of the results stated in sections §2 and §3.

12.1 Local approximation by polynomials

There are many important results on the approximation of functions in Besov classes. We will use approximation by piecewise polynomials. We begin by describing local polynomial approximation.

For any integer $r \geq 1$, we let $\mathcal{P}_r^d := \mathcal{P}_r$ denote the linear space of algebraic polynomials of order r (total degree $r-1$), namely,

$$\mathcal{P}_r := \left\{ \sum_{|\mathbf{k}|_1 < r} a_{\mathbf{k}} \mathbf{x}^{\mathbf{k}},\ a_{\mathbf{k}} \in \mathbb{R} \right\}, \quad \text{where} \quad \mathbf{x}^{\mathbf{k}} := x_1^{k_1} \cdots x_d^{k_d}, \quad \mathbf{k} := (k_1, \ldots, k_d), \quad k_j \geq 0, \quad |\mathbf{k}|_1 := \sum_{j=1}^{d} k_j.$$

Note that $\omega_r(P,t)_p = 0$, $t \geq 0$, for all $P \in \mathcal{P}_r$. If I is any cube in $\mathbb{R}^d$ and $f \in L_p(I)$, $0 < p \leq \infty$, we denote by

$$E_r(f,I)_p := \inf_{P \in \mathcal{P}_r} \|f - P\|_{L_p(I)}, \tag{12.1}$$

the error of approximation of f on I in the $\|\cdot\|_{L_p(I)}$ norm by the elements of $\mathcal{P}_r$. A well known result in approximation theory, commonly referred to as Whitney's theorem, says that for any $f \in L_p(I)$ with I a cube with sidelength ℓ_I, we have

$$cE_r(f,I)_p \leq \omega_r(f,\ell_I)_{L_p(I)} \leq CE_r(f,I)_p, \tag{12.2}$$

with the constants c, C depending only on r, d and also p if p is close to 0. Whitney's theorem usually only refers to the lower inequality in (12.2). However, the upper inequality follows trivially since

$$\omega_r(f,\ell_I)_{L_p(I)} = \omega_r(f-P,\ell_I)_{L_p(I)} \leq C\|f-P\|_{L_p(I)},$$

holds for any polynomial $P \in \mathcal{P}_r$.

It is useful to use the following modified form of Whitney's theorem. For any cube $I \subset \mathbb{R}^d$, any $0 < p < \infty$, and $f \in L_p(I)$, we define

$$\widetilde{w}_r(f,t)^p_{L_p(I)} := t^{-d} \int\limits_{\mathbf{h} \in [-t,t]^d} \int\limits_{I_{r\mathbf{h}}} |\Delta^r_{\mathbf{h}}(f,\mathbf{x})|^p \, d\mathbf{x}\, d\mathbf{h}, \tag{12.3}$$

where $I_{r\mathbf{h}} := \{\mathbf{x} : [\mathbf{x}, \mathbf{x} + r\mathbf{h}] \subset I\}$. This is called *the averaged modulus of smoothness of* f. It is known that w_r is equivalent to $\widetilde{w}_r$ (see §2 of [18]),

$$c\widetilde{w}_r(f,t)_{L_p(I)} \leq \omega_r(f,t)_{L_p(I)} \leq C\widetilde{w}_r(f,t)_{L_p(I)}, \quad 0 < t \leq 1, \tag{12.4}$$

where again the constants c, C depend only on r and p and can be taken absolute when r is fixed and $0 < p_0 \leq p \leq \infty$ with p_0 fixed. Thus, Whitney's theorem holds with ω_r replaced by $\widetilde{w}_r$

$$cE_r(f,I)_p \leq \widetilde{w}_r(f,\ell_I)_{L_p(I)} \leq CE_r(f,I)_p. \tag{12.5}$$

Before proceeding further, let us remark on why we introduce the averaged modulus of smoothness $\widetilde{w}_r$. The advantage of $\widetilde{w}_r$ over ω_r is that $\widetilde{w}^p_r$ it is set subadditive. We shall use this in the following form. Let $\Omega = (0,1)^d$ and let $\mathcal{I}$ be a collection of subcubes of Ω which form a partition of Ω. Then, from the definition of $\widetilde{w}_r$. we have

$$\sum_{I \in \mathcal{I}} \widetilde{w}_r(f,t)^p_{L_p(I)} \leq \widetilde{w}_r(f,t)^p_{L_p(\Omega)}, \quad t > 0. \tag{12.6}$$

This same subadditivity holds when $\mathcal{I}$ is replaced by a set $\mathcal{T}$ of simplices which form a partition of Ω.

If $I \subset \Omega$ is a cube, we say that $P_I \in \mathcal{P}_r$ is a near best $L_p(I)$ approximation to f with constant $\lambda \geq 1$ if

$$\|f - P_I\|_{L_p(I)} \leq \lambda E_r(f,I)_p. \tag{12.7}$$

It is shown in Lemma 3.2 of [16] that if $P_I \in \mathcal{P}_r$ is near best in $L_p(I)$ with constant λ, then it is also near best in $L_{\overline{p}}(I)$ whenever $\overline{p} \geq p$, i.e.,

$$\|f - P_I\|_{L_{\overline{p}}(I)} \leq C\lambda E_r(f,I)_{\overline{p}}, \tag{12.8}$$

with the constant C depending only on r, d and p. This constant does not depend on I or $\bar{p}$.

Another important remark is that any near best approximation P_I with constant λ is near best on any larger cube J which contains I in the sense that

$$\|f - P_I\|_{L_p(J)} \leq C\lambda E_r(f,J)_p, \tag{12.9}$$

where now C depends additionally on $|J|/|I|$, see Lemma 3.3 in [16]. Note that even though Lemma 3.2 and Lemma 3.3 in [16] are stated for polynomials of coordinate degree $< r$, they also hold for polynomials of total degree $< r$.

In summary, a near best $L_p(I)$ approximation is also near best on large cubes J containing I and larger values $\overline{p} \geq p$. We shall use these facts going forward.

12.2 Polynomial norms and inequalities

We need good bounds on the constants that appear when comparing norms. For this, we recall the following equivalences, see (3.2) of [18] or Lemma 3.1 and Lemma 3.2 in [17]. For any cube I in $\mathbb{R}^d$ and any function $g \in L_p(I)$, we introduce the normalized L_p norm

$$\|g\|^{\#}_{L_p(I)} := |I|^{-1/p}\|g\|_{L_p(I)}. \tag{12.10}$$

For any cube $I \subset \mathbb{R}^d$ and any polynomial $P \in \mathcal{P}_r$, and any $0 < p, q \leq \infty$, we have

$$\|P\|^{\#}_{L_q(I)} \asymp \|P\|^{\#}_{L_p(I)}, \tag{12.11}$$

with absolute constants of equivalency provided $q, p \geq p_0$ with $p_0 > 0$ fixed.

One can also compare Besov norms of polynomials. For example, we will use the fact that if $P \in \mathcal{P}_\ell$ and I is a cube in $\mathbb{R}^d$, then for every $s > 0$, we have

$$|P|_{B^s_q(L_p(I))} \leq C\ell_I^{-s}\|P\|_{L_p(I)}, \tag{12.12}$$

with a constant independent of I. This is proved by dilation (see e.g. Corollary 5.2 of [16]).

12.3 Besov spaces and piecewise polynomial approximation

We now recall how membership in $B^s_p(\Omega)$ guarantees a rate of approximation by piecewise polynomials. In actuality the membership of a function in a Besov space can be characterized by its rate of approximation by piecewise polynomials. However, we only need the results that prove that a function in a Besov space can be approximated by piecewise polynomials with a certain accuracy and therefore only concentrate on proving results of this type.

For each $k \geq 0$, we define $\mathcal{D}_k$ to be the partition of Ω into dyadic cubes I of side length 2^{-k}. Here we take a dyadic cube $I \in \mathcal{D}_k$ to be the tensor product of the dyadic intervals $[(j-1)2^{-k}, j2^{-k})$, $1 \leq j \leq 2^k$. We then define

$$\mathcal{S}_k = \mathcal{S}_k(r)$$

to be the space of all piecewise polynomials of order r that are subordinate to the partition $\mathcal{D}_k$. In other words, a function $S \in \mathcal{S}_k$ if an only the restriction of S on each $I \in \mathcal{D}_k$ belongs to $\mathcal{P}_r$.

If $I \in \mathcal{D}_k$, let $P_I \in \mathcal{P}_r$ be the polynomial of best $L_p(I)$ approximation to f and let us define

$$S_k := S_k(f) := \sum_{I \in \mathcal{D}_k} P_I \chi_I \in \mathcal{S}_k, \tag{12.13}$$

where χ_I is the characteristic function of I. Notice that $S_k(x)$ is defined pointwise for each $x \in \Omega$, We extend S_k to all of $\overline{\Omega}$ by continuity.

The following lemma holds.

Lemma 12.1. *Let $0 < p \leq \infty$, $s > 0$, and let $r \geq 2$ be any fixed integer strictly larger than s. If a function $f \in B^s_p = B^s_\infty(L_p(\Omega))$, with $\Omega = (0,1)^d$, then*

$$\operatorname{dist}(f, \mathcal{S}_k(r))_{L_p(\Omega)} \leq C|f|_{B^s_p(\Omega)}2^{-ks}, \quad k \geq 0, \tag{12.14}$$

with C depending only on p and s.

Proof. Consider S_k defined in (12.13). From Whitney's theorem, we have

$$\|f - P_I\|_{L_p(I)} \leq C\omega_r(f, 2^{-k})_{L_p(I)} \leq C\widetilde{w}_r(f, 2^{-k})_{L_p(I)}, \tag{12.15}$$

where $\widetilde{w}_r$ is the averaged modulus of smoothness.

If $p < \infty$, we raise both sides of (12.15) to the power p and then sum over $I \in \mathcal{D}_k$ to obtain

$$\|f - S_k\|_{L_p(\Omega)}^p \le C^p \sum_{I \in \mathcal{D}_k} \widetilde{w}_r(f, 2^{-k})_{L_p(I)}^p \le C^p \widetilde{w}_r(f, 2^{-k})_{L_p(\Omega)}^p \le C^p |f|_{B_p^s(\Omega)}^p 2^{-ksp}, \tag{12.16}$$

where we used the subadditivity (12.6) and the fact that $\mathcal{D}_k$ is a partition of Ω. This proves (12.14) in the case $p < \infty$. When $p = \infty$, these inequalities follow directly from (12.15) and the fact that $\|\cdot\|_{L_\infty(I)} \le \|\cdot\|_{L_\infty(\Omega)}$ for each $I \subset \Omega$. □

12.4 Piecewise polynomial approximation in $L_\tau(\Omega)$

All constants C appearing in this section depend at most on s and p and may change at each occurrence.

Theorem 12.2. *Let S_k be defined as in* (12.13). *If $s > d/p$ and $p \le \tau \le \infty$, then we have*

$$\|f - S_k\|_{L_\tau(\Omega)} \le C|f|_{B_p^s(\Omega)} 2^{-k(s-d/p+d/\tau)}, \quad f \in B_p^s(\Omega). \tag{12.17}$$

Proof. Let us fix any $f \in B_p^s(\Omega)$ and consider the corresponding $S_k = S_k(f)$, see (12.13). It was proven in Lemma 12.1 that

$$\|f - S_k\|_{L_p(\Omega)} \le C|f|_{B_p^s(\Omega)} 2^{-ks}, \quad k \ge 0.$$

Let $R_0 := S_0$ and for each $\mathbf{x} \in \overline{\Omega}$, we define

$$R_k(\mathbf{x}) := S_k(\mathbf{x}) - S_{k-1}(\mathbf{x}), \quad k \ge 1. \tag{12.18}$$

The functions R_k are defined for all $\mathbf{x} \in \overline{\Omega}$ and are in $\mathcal{S}_k$, $k \ge 0$,

$$\|R_k\|_{L_p(\Omega)} \le C[\|f - S_k\|_{L_p(\Omega)} + \|f - S_{k-1}\|_{L_p(\Omega)}] \le C|f|_{B_p^s(\Omega)} 2^{-ks}, \quad k \ge 1, \tag{12.19}$$

and of course $\|R_0\|_{L_p(\Omega)} \le C\|f\|_{L_p(\Omega)}$. It follows that

$$f = \sum_{k=0}^{\infty} R_k \tag{12.20}$$

with the series converging in $L_p(\Omega)$. We now consider the following cases for τ.
Case 1: $\tau = \infty$. On each dyadic cube $I \in \mathcal{D}_k$, we have

$$R_k(\mathbf{x}) = Q_I(\mathbf{x}) := P_I(\mathbf{x}) - P_{\overline{I}}(\mathbf{x}), \quad \mathbf{x} \in I,$$

where $\overline{I} \in \mathcal{D}_{k-1}$ is the parent of I. From Whitney's theorem and (12.11), we have for every $I \in \mathcal{D}_k$, $k \ge 1$,

$$\begin{aligned} \|Q_I\|_{C(I)} &\le C|I|^{-1/p}\|Q_I\|_{L_p(I)} \le C2^{kd/p}[\|f - P_I\|_{L_p(I)} + \|f - P_{\overline{I}}\|_{L_p(I)}] \\ &\le C2^{kd/p}\omega_r(f, 2^{-k+1})_{L_p(\overline{I})}. \end{aligned} \tag{12.21}$$

This gives that for each $\mathbf{x} \in \overline{\Omega}$, we have

$$|R_k(\mathbf{x})| \le C2^{kd/p}\omega_r(f, 2^{-k+1})_{L_p(\Omega)} \le C|f|_{B_p^s(\Omega)} 2^{-k(s-d/p)}, \quad k \ge 1. \tag{12.22}$$

Thus the series (12.20) converges in $L_\infty(\overline{\Omega})$ and also pointwise to a limit function $\widetilde{f}$ and for each $\mathbf{x} \in \overline{\Omega}$, we have

$$|\widetilde{f}(\mathbf{x}) - S_k(\mathbf{x})| \le \sum_{j>k} |R_j(\mathbf{x})| \le C|f|_{B_p^s(\Omega)} 2^{-k(s-d/p)}, \quad k \ge 0. \tag{12.23}$$

Since the same bound holds for $\|f - S_k\|_{L_\infty(\overline{\Omega})}$, we have $f = \widetilde{f}$, a.e. on $\overline{\Omega}$, and we have proven the theorem in the case $\tau = \infty$.

In going further, we refer to $\widetilde{f}$, which is defined pointwise on $\overline{\Omega}$ as the representer of f. We shall see in Theorem 2.3 that $\widetilde{f}$ is continuous and in Lip α, $\alpha = s - d/p$.

Case 2: $\tau < \infty$. Similarly to (12.21) and using the comparison of polynomial norms of (12.11), we find that

$$\begin{aligned}
\int_\Omega |R_k|^\tau &= \sum_{I\in\mathcal{D}_k} \int_I |Q_I|^\tau \le C^\tau 2^{kd(\tau/p-1)} \sum_{I\in\mathcal{D}_k} \|Q_I\|^\tau_{L_p(I)} \le C^\tau 2^{kd(\tau/p-1)} \sum_{I\in\mathcal{D}_k} \widetilde{w}_r(f,2^{-k+1})^\tau_{L_p(\overline{I})} \\
&\le C^\tau 2^{kd(\tau/p-1)} \left[\sum_{I\in\mathcal{D}_k} \widetilde{w}_r(f,2^{-k+1})^p_{L_p(\overline{I})}\right]^{\tau/p} \le 2^{d\tau/p} C^\tau 2^{kd(\tau/p-1)} \left[\sum_{\overline{I}\in\mathcal{D}_{k-1}} \widetilde{w}_r(f,2^{-k+1})^p_{L_p(\overline{I})}\right]^{\tau/p} \\
&\le 2^{d\tau/p} C^\tau 2^{kd(\tau/p-1)} \widetilde{w}_r(f,2^{-k+1})^\tau_{L_p(\Omega)},
\end{aligned}$$

where we used the subadditivity (12.6). In other words, for any $\tau \ge p$, we have

$$\|R_k\|_{L_\tau(\Omega)} \le C 2^{kd(1/p-1/\tau)} |f|_{B^s_p(\Omega)} 2^{-ks}, \quad k \ge 1. \tag{12.24}$$

Since $f - S_k = \sum_{j>k} R_j$, when $\tau \ge 1$, we can add these estimates to arrive at

$$\|f - S_k\|_{L_\tau(\Omega)} \le \sum_{j>k} \|R_j\|_{L_\tau(\Omega)} \le C|f|_{B^s_p(\Omega)} 2^{-k(s-d/p+d/\tau)},$$

which is the desired inequality. When $p < \tau < 1$, we use (12.24) to obtain

$$\|f - S_k\|^\tau_{L_\tau(\Omega)} \le \sum_{j>k} \|R_j\|^\tau_{L_\tau(\Omega)} \le C^\tau |f|^\tau_{B^s_p(\Omega)} \sum_{j>k} 2^{-j\tau(s-d/p+d/\tau)} \le C^\tau |f|^\tau_{B^s_p(\Omega)} 2^{-k\tau(s-d/p+d/\tau)},$$

which completes the proof in this case. □

Remark 12.3. *Theorem* 12.2 *is valid for more general* $S_k \in \mathcal{S}_k$*, for example, for* $S_k = \sum_{I\in\mathcal{D}_k} P_I \chi_I$*, where*

$$\|f - P_I\|_{L_\tau(I)} \le C 2^{kd(1/p-1/\tau)} \widetilde{w}_r(f,2^{-k})_{L_p(I)}, \quad I \in \mathcal{D}_k(\Omega), \tag{12.25}$$

where C *does not depend on* I*, see* (12.21).

12.4.1 Proof of Theorem 2.3

For any $f \in B^s_p(\Omega)$, consider the corresponding $S_k(f)$, defined in (12.13) and the function $\widetilde{f}$, see (12.23). As shown in Theorem 12.2, $f = \widetilde{f}$ a.e.. We want to observe that the function $\widetilde{f}$ is continuous on Ω and moreover, (2.19) holds. Indeed, it is enough to consider $0 < t \le 1$. Let $\mathbf{x} \in \Omega$ and $|\mathbf{h}| \le t$ be such that $[\mathbf{x}, \mathbf{x} + r\mathbf{h}] \subset \Omega$. Let J be the smallest cube that contains $[\mathbf{x}, \mathbf{x} + r\mathbf{h}]$. Then, we have $\ell_J \le rt$. We argue as in the proof of (12.22) and (12.23), replacing Ω by J to show that there is a polynomial $P_J \in \mathcal{P}_r$ such that

$$\sup_{\mathbf{x}\in J} |\widetilde{f}(\mathbf{x}) - P_J(\mathbf{x})| \le C|f|_{B^s_p(J)} \ell_J^{s-d/p} \le C|f|_{B^s_p(\Omega)} t^{s-d/p}.$$

It follows that

$$|\Delta_{r\mathbf{h}}(\widetilde{f},\mathbf{x})| \le C|f|_{B^s_p(\Omega)} t^{s-d/p},$$

uniformly for $\mathbf{x} \in \Omega$. This proves that $\widetilde{f}$ is continuous and (2.19) holds. □

12.5 Polynomial interpolation

In this section, we prove the results for piecewise polynomial approximation stated in §2.

12.5.1 Proof of Theorem 2.1

In view of Theorem 12.2, it is enough to show that for any fixed $k \geq 0$, we have

$$\|S_k(f) - S_k^*(f)\|_{L_\tau(\Omega)} \leq C|f|_{B_p^s(\Omega)} 2^{-k(s-d/p+d/\tau)}, \tag{12.26}$$

with C here and for the remainder of the proof always denoting a constant not depending on f or k, but depending on r and d. If $I \in \mathcal{D}_k$, we let P_I be the best $L_p(I)$ approximation to f by elements of $\mathcal{P}_r$ and similarly let P_I^* be the best $C(I)$ approximation to f from $\mathcal{P}_r$. Recall that by definition $S_k(f) = \sum_{I\in\mathcal{D}_k} P_I\chi_I$. It follows that if $T \in \mathcal{T}_k$ and $T \subset I$, then we have

$$\|S_k^*(f) - S_k(f)\|_{L_\infty(T)} \leq \|L_T(f - P_I)\|_{L_\infty(T)} \leq C\|f - P_I\|_{L_\infty(T)} \leq C\|f - P_I\|_{L_\infty(I)}, \tag{12.27}$$

and therefore

$$\|S_k^*(f) - S_k(f)\|_{L_\infty(I)} \leq C\|f - P_I\|_{L_\infty(I)}. \tag{12.28}$$

We know from Theorem 12.2 that $\|f - P_I\|_{L_\infty(I)} \leq C|f|_{B_p^s(\Omega)} 2^{-k(s-d/p)}$ for each $I \in \mathcal{D}_k$, and thus

$$\|S_k(f) - S_k^*(f)\|_{L_\infty(\Omega)} \leq C|f|_{B_p^s(\Omega)} 2^{-k(s-d/p)}, \tag{12.29}$$

which completes the proof of the theorem when when $\tau = \infty$.

To handle the case $p \leq \tau < \infty$, we first give an improved bound for $\|f - P_I\|_{L_\infty(I)}$ when $I \in \mathcal{D}_k$. Let $\mathbf{x} \in I$ be any fixed point. For each index $j \geq k$, let $J_j \in \mathcal{D}_j$ be the dyadic cube that contains $\mathbf{x}$ and $Q_j := P_{J_{j+1}} - P_{J_j}$, $j \geq k$, $I = J_k$. Then, arguing as in (12.21), we have

$$|f(\mathbf{x}) - P_I(\mathbf{x})| \leq \sum_{j\geq k} \|Q_j\|_{L_\infty(J_{j+1})} \leq C\sum_{j\geq k} |J_{j+1}|^{-1/p}\|Q_j\|_{L_p(J_{j+1})} \leq C\sum_{j\geq k} 2^{jd/p}\|f - P_{J_j}\|_{L_p(J_j)}. \tag{12.30}$$

Since $\|f - P_{J_j}\|_{L_p(J_j)} \leq \widetilde{w}_r(f, 2^{-j})_{L_p(I)}$, this gives

$$\|f - P_I\|_{L_\infty(I)} \leq C\sum_{j\geq k} 2^{jd/p}\widetilde{w}_r(f, 2^{-j})_{L_p(I)}, \tag{12.31}$$

which is the improved bound we want.

We now consider two cases.

Case $p \leq 1$: From (12.28) and (12.31) we derive

$$\begin{aligned} \|S_k(f) - S_k^*(f)\|_{L_\infty(I)} &\leq C\|f - P_I\|_{L_\infty(I)} \leq C\sum_{j\geq k} 2^{jd/p}\widetilde{w}_r(f, 2^{-j})_{L_p(I)} \\ &\leq C\left[\sum_{j\geq k} 2^{jd}\widetilde{w}_r(f, 2^{-j})_{L_p(I)}^p\right]^{1/p}. \end{aligned} \tag{12.32}$$

The set subadditivity (12.6) of $\widetilde{w}_r$ then gives

$$\begin{aligned} \|S_k(f) - S_k^*(f)\|_{L_p(\Omega)}^p &\leq C2^{-kd}\sum_{I\in\mathcal{D}_k} \|S_k(f) - S_k^*(f)\|_{L_\infty(I)}^p \leq C2^{-kd}\sum_{j\geq k} 2^{jd}\sum_{I\in\mathcal{D}_k} \widetilde{w}_r(f, 2^{-j})_{L_p(I)}^p \\ &\leq C2^{-kd}\sum_{j\geq k} 2^{jd}\widetilde{w}_r(f, 2^{-j})_{L_p(\Omega)}^p \leq C|f|_{B_p^s(\Omega)}^p 2^{-kd}\sum_{j\geq k} 2^{jd}2^{-jsp} \leq C|f|_{B_p^s(\Omega)}^p 2^{-ksp}. \end{aligned}$$

This proves (12.26) for the case $\tau = p$. For general $\tau \in (p, \infty)$, we use (12.29) and obtain

$$|S_k(f) - S_k^*(f)|^\tau \leq |S_k(f) - S_k^*(f)|^{\tau-p}|S_k(f) - S_k^*(f)|^p \leq C|f|_{B_p^s(\Omega)}^{\tau-p} 2^{-k(s-d/p)(\tau-p)}|S_k(f) - S_k^*(f)|^p. \tag{12.33}$$

We now integrate both sides of this inequality and find that

$$\|S_k(f) - S_k^*(f)\|_{L_\tau(\Omega)}^\tau \le C|f|_{B_p^s(\Omega)}^\tau 2^{-k(s-d/p)(\tau-p)} 2^{-ksp} \le C|f|_{B_p^s(\Omega)}^\tau 2^{-k(s-d/p+d/\tau)\tau}. \tag{12.34}$$

as desired.

Case $p \ge 1$**:** From (12.31), we find for any $\beta \in (0, s-d/p)$ that

$$\|f - P_I\|_{L_\infty(I)} \le C\{\sum_{j\ge k} 2^{-j\beta p'}\}^{1/p'} \{\sum_{j\ge k} [2^{j(d/p+\beta)} \widetilde{w}_r(f, 2^{-j})_{L_p(I)}^p]^p\}^{1/p}, \tag{12.35}$$

where p' is the conjugate index to p, i.e., $1/p + 1/p' = 1$. Arguing as in (12.33), we arrive at

$$\|S_k(f) - S_k^*(f)\|_{L_p(\Omega)} \le C|f|_{B_p^s(\Omega)} 2^{-ks}. \tag{12.36}$$

This proves (12.26) for $\tau = p$. In other words, we know (12.26) for $\tau = p$ and $\tau = \infty$. We complete the proof for general $\tau \in [p, \infty)$ as in (12.34). This completes the proof of the theorem. □

12.5.2 Proof of Theorem 2.2

Since $B_q^s(L_p(\Omega))$ is embedded in $B_p^s(\Omega) := B_\infty^s(L_p(\Omega))$, it is enough to prove the theorem when $f \in B_p^s(\Omega)$, i.e., when $q = \infty$. We know from Theorem 2.1 that

$$\|f - S_k^*(f)\|_{L_2(\Omega)} \le C|f|_{B_p^s(\Omega)} 2^{-k(s-d/p+d/2)}, \quad k \ge 0. \tag{12.37}$$

We define

$$R_k^*(f) := S_k^*(f) - S_{k-1}^*(f), \quad k \ge 1. \tag{12.38}$$

It follows from (12.37) that

$$\|R_k^*(f)\|_{L_2(\Omega)} \le C|f|_{B_p^s(\Omega)} 2^{-k(s-d/p+d/2)}, \quad k \ge 1. \tag{12.39}$$

Since on each simplex $T \in \mathcal{T}_k$, the function $R_k^*(f)$ is a polynomial Q_T from $\mathcal{P}_r$, we have

$$\|Q_T\|_{H^1(T)} \le C2^k \|Q_T\|_{L_2(T)}, \quad T \in \mathcal{T}_k. \tag{12.40}$$

Since the function $R_k^*(f)$ is continuous on Ω, we have (see Theorem 2.1.2 in [7])

$$\|R_k^*(f)\|_{H^1(\Omega)}^2 \le C \sum_{T\in\mathcal{T}_k} \|R_k^*(f)\|_{H^1(T)}^2 \le C2^{2k} \sum_{T\in\mathcal{T}_k} \|R_k^*(f)\|_{L_2(T)}^2 = C2^{2k} \|R_k^*(f)\|_{L_2(\Omega)}^2.$$

This gives

$$\|R_k^*(f)\|_{H^1(\Omega)} \le C2^k \|R_k^*(f)\|_{L_2(\Omega)} \le C|f|_{B_p^s(\Omega)} 2^{-k(s-1-d/p+d/2)}, \qquad k \ge 1. \tag{12.41}$$

Writing,

$$f - S_k^*(f) = \sum_{j>k} R_j^*(f),$$

and then applying H^1 norms gives the theorem because $s - d/p + d/2 > 1$. □

12.6 The proofs of the Theorems on optimal recovery

This subsection will be devoted to the proof of the theorems on optimal recovery stated in §3.

12.6.1 The proof of Theorem 3.1

It will be convenient to consider only the case $d \geq 2$ to avoid changes in notation. We leave the case $d = 1$ to the reader. The proofs are divided into two parts: a proof of the upper bounds and then a proof of the lower bounds.

Proof of the upper bounds in Theorem 3.1: A proof of (i) and (ii) can already be found in [49], but for completeness, we give the following proof using the results of §2. Recall our notation $B^s_p(\Omega) := B^s_\infty(L_p(\Omega))$ and the fact that all other Besov spaces $B^s_q(L_p(\Omega))$, with the same values for p and s, are contained in $B^s_p(\Omega)$. Therefore, it is enough to prove the upper bounds when $\mathcal{B} = B^s_p(\Omega)$. To prove these upper bounds, we need to exhibit a set of $\widetilde{m}$ data sites at which we achieve the claimed bound. We take these data sites as the tensor product grid points $G_{k,r}$ of $\overline{\Omega}$ where $r > \max(s, 1)$. This will yield the upper bounds when $\widetilde{m} = (2^k r)^d$. The upper bounds for the other values of $\widetilde{m}$ follow from the fact that $R^*_{\widetilde{m}}$ is monotonically nonincreasing as $\widetilde{m} \to \infty$

The cases (i) and (ii): Let $\mathcal{T}_k = \mathcal{T}_k(\Omega)$ be the simplicial partition of Ω into simplices T as in §12.5. We have shown in that section, that from the data $(f_i)_{i=1}^{\widetilde{m}}$ we can create a continuous piecewise polynomial

$$S = S^*_k(f) = \sum_{T \in \mathcal{T}_k} L_T(f)\chi_T, \tag{12.42}$$

where $L_T(f)$ is a polynomial of order r ($L_T(f) \in \mathcal{P}_r$), gotten by interpolating f at the data points $T \cap G_{k,r}$. We have shown (see (2.17)) that if $\tau \geq p$ then

$$\|f - S\|_{L_\tau(\Omega)} \preceq |f|_{B^s_p(\Omega)} 2^{-kd\alpha_\tau} \preceq |f|_{B^s_p(\Omega)} m^{-\alpha_\tau}.$$

This bound also holds for $\tau < p$ since then $B^s_p(\Omega)$ is continuously embedded in $B^s_\tau(\Omega)$. The function S only depends on the data $(f_i)_{i=1}^{\widetilde{m}}$. This shows that every function $\widetilde{f} \in \mathcal{F}_{\rm data}$ is within an $L_\tau(\Omega)$ distance $Cm^{-\alpha_\tau}$ of S and hence the proof of the upper bound in the cases (i) and (ii) follow.

The case (iii): This is similar to case (ii) except now we use Theorem 2.2.

The case (iv): We shall need the following lemma.

Lemma 12.4. *Let $\Omega = (0,1)^d$, $d \geq 3$, and let $\gamma = \gamma(d)$ be defined as in* (3.12)*. Then, the space $L_\gamma(\Omega)$ is embedded into $H^{-1}(\Omega)$ and we have*

$$\|\widetilde{f}\|_{H^{-1}(\Omega)} \leq C\|\widetilde{f}\|_{L_\gamma(\Omega)}, \quad \widetilde{f} \in L_\gamma(\Omega), \tag{12.43}$$

with C depending only on d. In the case $d = 2$, then for any $\gamma(2) = 1 < \tau \leq \infty$, the space $L_\tau(\Omega)$ embeds into $H^{-1}(\Omega)$ and we have the bound

$$\|\widetilde{f}\|_{H^{-1}(\Omega)} \leq C\frac{\tau}{\tau - 1}\|\widetilde{f}\|_{L_\tau(\Omega)}, \quad \widetilde{f} \in L_\tau(\Omega), \tag{12.44}$$

where C is an absolute constant.

Proof. Consider first the case $d \geq 3$, in which case $\gamma(d) > 1$. Let γ' be the dual index of γ, that is, $\frac{1}{\gamma'} = \frac{1}{2} - \frac{1}{d}$, so that $\gamma' < \infty$. If $v \in U(H^1_0(\Omega))$, then by the Sobolev embedding theorem we have

$$\|v\|_{L_{\gamma'}(\Omega)} \lesssim \|v\|_{H^1(\Omega)} \lesssim |v|_{H^1_0(\Omega)} := \||\nabla v|\|_{L_2(\Omega)}, \quad v \in H^1_0(\Omega), \tag{12.45}$$

where the last inequality is Poincare's inequality. From Hölder's inequality, it follows that for any $v \in H^1_0(\Omega)$ and any $\widetilde{f} \in L_\gamma(\Omega)$, we have

$$\int_\Omega \widetilde{f}v \leq \|\widetilde{f}\|_{L_\gamma(\Omega)}\|v\|_{L_{\gamma'}(\Omega)} \lesssim \|\widetilde{f}\|_{L_\gamma(\Omega)}|v|_{H^1_0(\Omega)}. \tag{12.46}$$

This gives

$$\|\widetilde{f}\|_{H^{-1}(\Omega)} := \sup_{v\in U(H^1_0(\Omega))} \int_\Omega \widetilde{f} v \lesssim \|\widetilde{f}\|_{L_\gamma(\Omega)}, \quad \widetilde{f} \in L_\gamma(\Omega), \tag{12.47}$$

which proves the lemma when $d \geq 3$.

When $d = 2$, the Sobolev embedding fails at the endpoint since $\gamma' = \infty$. However, for any $\tau > 1$ we have that $\tau' < \infty$, where τ' is the dual index to τ. In this case we have a compact Sobolev embedding

$$\|v\|_{L_{\tau'}(\Omega)} \leq C_{\tau'} \|v\|_{H^1(\Omega)}, \tag{12.48}$$

where one can show that the constant $C_{\tau'} = C\tau' = C\tau/(\tau-1)$ for an absolute constant C. For completeness, we sketch this argument. Given a function $f \in H^1(\Omega)$, we consider a multiscale decomposition of f

$$f = \sum_{n=0}^{\infty} (f_n - f_{n-1}), \tag{12.49}$$

where f_n denotes a piecewise linear interpolation of f on a triangulation $\mathcal{T}_n$ made of $C2^{2n}$ triangles of diameter $c2^{-n}$ (here we set $f_{-1} = 0$). We use a standard polynomial interpolation bound and scaling argument, combined with the fact that the linear polynomials are finite dimensional (essentially the Bramble-Hilbert Lemma) to see that

$$\begin{aligned} \|f_n - f_{n-1}\|^{\tau'}_{L_{\tau'}(\Omega)} = \sum_{T\in\mathcal{T}_{n-1}} \|f_n - f_{n-1}\|^{\tau'}_{L_{\tau'}(T)} &\leq C2^{2n(\tau'/2-1)} \sum_{T\in\mathcal{T}_{n-1}} \|f_n - f_{n-1}\|^{\tau'}_{L_2(T)} \\ &\leq C2^{2n(\tau'/2-1)} 2^{-\tau' n} \sum_{T\in\mathcal{T}_{n-1}} \|f\|^{\tau'}_{H^1(T)}. \end{aligned} \quad \text{span}$$

Taking τ'-th roots, we then get using that the $\ell_{\tau'}$-norm is bounded by the ℓ_2-norm since $\tau' \geq 2$ that

$$\begin{aligned} \|f_n - f_{n-1}\|_{L_{\tau'}(\Omega)} \leq C2^{-2n/\tau'} \left(\sum_{T\in\mathcal{T}_{n-1}} \|f\|^{\tau'}_{H^1(T)} \right)^{1/\tau'} &\leq C2^{-2n/\tau'} \left(\sum_{T\in\mathcal{T}_{n-1}} \|f\|^2_{H^1(T)} \right)^{1/2} \\ &= C2^{-2n/\tau'} \|f\|_{H^1(\Omega)}, \end{aligned} \quad \text{span}$$

where C is independent of τ'. Using the multiscale decomposition, this now implies that

$$\|f\|_{L_{\tau'}(\Omega)} \leq C\|f\|_{H^1(\Omega)} \sum_{n=0}^{\infty} 2^{-2n/\tau'} = \frac{C}{1-2^{-2/\tau'}} \|f\|_{H^1(\Omega)} \leq C\tau' \|f\|_{H^1(\Omega)}$$

for a constant C independent of τ'. Utilizing (12.48), the same duality argument used for $d \geq 3$ now completes the proof when $d = 2$. □

We can now prove the upper bound in the case (iv) when error is measured in $H^{-1}(\Omega)$. Let us consider first the case where $d \geq 3$. Since the exponent α_{-1} does not change over the $p \geq \gamma$, it is enough to consider the case when $p \leq \gamma$. Indeed, if $p > \gamma$ then $U(B^s_q(L_p(\Omega)))$ is contained in $U(B^s_\infty(L_\gamma(\Omega))) =: U(B^s_\gamma(\Omega))$. Therefore the upper bound follows once we have established the case $\mathcal{F} = U(B^s_q(L_p(\Omega)))$ with $p \leq \gamma$. Similarly, when $p \leq \gamma$, then $U(B^s_q(L_p(\Omega))) \subset U(B^s_\infty(L_p(\Omega))$. Accordingly, in going further we only need to consider the case when $\mathcal{F} = U(B^s_\infty(L_p(\Omega)) =: U(B^s_p(\Omega))$ with $p \leq \gamma$.

Let f be any function in $\mathcal{F} := U(B^s_\infty(L_p(\Omega))$ with $s > d$ and $p \leq \gamma$. We know that $f \in C(\Omega)$. Given the data $\mathbf{f}$, we define S as in (12.42). From (2.17), we have

$$\|f - S\|_{L_\gamma(\Omega)} \preceq |f|_{B^s_p(I)} 2^{-kd\alpha_{-1}}. \tag{12.50}$$

We now use Lemma 12.4 to obtain

$$\|f - S\|_{H^{-1}\Omega} \leq C\|f - S\|_{L_\gamma(\Omega)} \leq Cm^{-\alpha_{-1}}|f|_{B^s_p(\Omega)}. \tag{12.51}$$

The function S only depends on the data. This shows that every element in $\mathcal{F}_{\text{data}}(f)$ is within an $H^{-1}(\Omega)$ distance $Cm^{-\alpha_{-1}}$ of S and hence proves that the Chebyshev radius of this set does not exceed $Cm^{-\alpha_{-1}}$. Since f was arbitrary we obtain the same bound for $R^*(\mathcal{F})$. This concludes the proof of the upper bound in case (iv) when $d \geq 3$.

If $d = 2$, we must slightly modify the above argument. Suppose first that $p > 1$. In this case, we have

$$\|f - S\|_{L_p(\Omega)} \preceq |f|_{B^s_p(I)} 2^{-kd\alpha_{-1}}, \tag{12.52}$$

since the sampling numbers in L_p and L_1 are the same in this case. Thus, Lemma 12.4 implies that (setting $\tau = p$)

$$\|f - S\|_{H^{-1}\Omega} \leq C_p\|f - S\|_{L_p(\Omega)} \leq C_p m^{-\alpha_{-1}}|f|_{B^s_p(\Omega)}. \tag{12.53}$$

On the other hand, if $p \leq 1$, we see that

$$\|f - S\|_{L_\tau(\Omega)} \preceq |f|_{B^s_p(I)} 2^{-kd(\alpha_{-1}-1+1/\tau)}. \tag{12.54}$$

Applying Lemma 12.4 we get that

$$\|f - S\|_{H^{-1}\Omega} \leq C_p\|f - S\|_{L_p(\Omega)} \leq \frac{C}{\tau - 1} m^{-\alpha_{-1}} m^{1-1/\tau}|f|_{B^s_p(\Omega)}. \tag{12.55}$$

Since this is true for all $\tau > 1$ we finally optimize in τ, which gives

$$\min_{\tau>0} \frac{m^{1-1/\tau}}{\tau - 1} \leq C\log(m) \tag{12.56}$$

by setting $\tau = 1 + (\log(m))^{-1}$. □

Proof of the lower bounds in Theorem 3.1. We shall now prove the lower bounds in Theorem 3.1. The proofs of lower bounds all take the following form. Suppose that $\mathbf{x}_i$, $i = 1, \ldots, m$, are any m data sites. In order to prove a lower bound for $R^*(\mathcal{F})_X$, we construct a function η in $\mathcal{F}$ so that

(a) $\eta(\mathbf{x}_i) = 0, \quad i = 1, \ldots, m,$

(b) $\|\eta\|_X \geq cm^{-\alpha}$,

where α is the appropriate index for X. Since both η and the zero function satisfy zero data the bound (b) gives the lower bound we want for $R^*(\mathcal{F})_X$.

We proceed to explain how to construct an appropriate function η for each of the choices of X. Let φ be a smooth non-negative function on $\mathbb{R}^d$ which vanishes outside Ω and additionally satisfies

$$\|\varphi\|_{L_\infty(\mathbb{R}^d)} = 1 \quad \text{and} \quad \varphi(x) \geq 1/2, \quad x \in \Omega_0, \tag{12.57}$$

where $\Omega_0 := [1/4, 3/4]^d$. Of course, there are many such functions φ but to be more specific, and for use further, we assume $\varphi(\mathbf{x}) = \phi(x_1) \cdots \phi(x_d)$ where ϕ is a univariate function with these properties (for $d = 1$).

We choose φ with these properties which has the smallest norm

$$M_{s,p,q} := \|\varphi\|_{B^s_q(L_p(\Omega))}. \tag{12.58}$$

So, φ depends on s, p, q and $M_{s,p,q}$ is a fixed constant since the parameters s, p, q are fixed.

Now consider any cube $I \subset \Omega$ and let ξ_I be the smallest vertex of I and as usual ℓ_I is its side length. We define the function

$$\varphi_I := \varphi_{I,s,p,q}(x) := \ell_I^{s-\frac{d}{p}} \varphi(\ell_I^{-1}(x - \xi_I)), \quad x \in \mathbb{R}^d. \tag{12.59}$$

This function vanishes outside I and on its boundary. Moreover, one easily checks that for all cubes I, we have

$$\|\varphi_I\|_{B^s_q(L_p(\Omega))} = M_{s,p,q}. \tag{12.60}$$

Here the norm is independent of I.

Because of the monotonicity of $R^*_m(\mathcal{F})_X$, it is enough to prove the lower bound when $m = 2^{kd}$ as we now assume for m. Suppose that m sample points $X_m := \{\mathbf{x}_1, ..., \mathbf{x}_m\} \in \Omega$ are given. Consider the regular tensor product grid $G_{k+2,2}$ with spacing $2^{-(k+2)}$. The number of cubes in this grid is equal to $4^d m$ by construction and thus at least $(4^d - 1)m$ cubes do not contain any sample points in its interior. Denote by $\mathcal{I} := \mathcal{I}(X_m)$ the set of cubes which do not contain any sample points, so that $|\mathcal{I}| \geq (4^d - 1)m = C_d m$ and the volume of each cube $I \in \mathcal{I}$ is $|I| \geq c_d m^{-1}$. For any cube $I \in \mathcal{I}$, we define

$$\eta_I := M^{-1}\varphi_I, \quad M = M_{s,p,q}. \tag{12.61}$$

It follows that each η_I is in $\mathcal{F}$. Also, each of the η_I vanishes at each of the data sites. We shall use the η_I to prove the lower bounds of Theorem 3.1.

The lower bound in case (i): Let us fix s, p, q and consider the functions η_I defined by (12.61). We take I as any fixed cube from $\mathcal{I}$ and define $\eta := \eta_I$. Then η vanishes at each of the data sites. Also, it follows from the definition of η that with $M = M_{s,p,q}$, we have

$$\|\eta\|_{L_\infty(\Omega)} = M^{-1}\|\varphi_I\|_{L_\infty(\Omega)} = M^{-1}\ell_I^{s-\frac{d}{p}} \geq cm^{-(\frac{s}{d}-\frac{1}{p})} = cm^{-\alpha_C}.$$

This gives that $R^*(\mathcal{F})_{C(\Omega)} \geq cm^{-\alpha_C}$ and thereby proves the lower bound in case (i).

The lower bound in case (ii): We consider separately the cases $\tau < p$ and $\tau \geq p$. If $\tau \geq p$, then we take $\eta = \eta_I$ where I is any single fixed cube in $\mathcal{I}$. For $M = M_{s,p,q}$, we have from (12.57) that

$$\|\eta\|_{L_\tau(\Omega)} = M^{-1}\|\varphi_I\|_{L_\tau(\Omega)} \geq cM^{-1}\ell_I^{s-\frac{d}{p}+\frac{d}{\tau}} \geq cm^{-(\frac{s}{d}-\frac{1}{p}+\frac{1}{\tau})} = cm^{-\alpha_\tau},$$

since $1/p - 1/\tau \geq 0$. It follows that $R^*(\mathcal{F})_{L_\tau(\Omega)} \geq cm^{-\alpha_\tau}$ in the case $\tau \geq p$.

Next consider the case $\tau < p < \infty$. Since $U(B^s_q(L_\infty(\Omega))) \subset U(B^s_q(L_p(\Omega)))$, it is sufficient to prove the lower bound in the case $p = \infty$, i.e., when $\mathcal{F} = U(B^s_q(L_\infty(\Omega)))$ with $0 < q < \infty$ arbitrary but fixed. We define

$$\eta = \kappa \sum_{I\in\mathcal{I}} \eta_I, \tag{12.62}$$

with κ a fixed constant chosen so that $\eta \in U(B^s_q(L_\infty(\Omega)))$. Here, η_I is defined as in (12.61) with $p = \infty$. We next derive a lower bound for κ. The terms η_I in (12.62) each have L_∞ norm bounded by $M2^{-ks}$ and they have disjoint supports. Therefore, $\|\eta\|_{L_\infty(\Omega)} \leq M2^{-ks}$. For each $r \geq 1$, these functions are in $C^r(\Omega)$ with $C^r(\Omega)$ norms not exceeding $M2^{-ks}2^{kr}$ where M depends only on the choice of r. We take $r := \lceil s \rceil + 1$ and find that the modulus of smoothness of η satisfies

$$\omega_r(\eta, t)_\infty \leq M\kappa 2^{-ks} \min(1, 2^{kr}t^r). \tag{12.63}$$

Therefore, breaking the integral into the integral over $[0, 2^{-k}]$ and $(2^{-k}, 1]$ we find that

$$\int_0^1 [t^{-s}\omega_r(\eta, t)_\infty]^q \frac{dt}{t} \leq \kappa^q M^q 2^{-kq(s-r)} \int_0^{2^{-k}} t^{(r-s)q}\frac{dt}{t} + \kappa^q M^q 2^{-kqs} \int_{2^{-k}}^1 t^{-sq}\frac{dt}{t} \leq \kappa^q C^q, \tag{12.64}$$

with C a fixed constant. This shows that we can take $\kappa \geq c$ where c depends only on s. Since $\eta \geq \kappa 2^{-ks}$ on a set of measure $\geq 1/2$, we have that $\|\eta\|_{L_\tau(\Omega)} \geq c2^{-ks} \geq cm^{-s/d}$ which finishes the proof of the lower bound in case (ii).

The lower bound in case (iii): By assumption we have $p \le 2$. We take $\eta = \eta_I$ for any fixed $I \in \mathcal{I}$. We estimate

$$\|\eta\|_{H^1(\Omega)} = M^{-1}\|\varphi_I\|_{H^1(\Omega)} \ge cM^{-1}\ell_I^{s-1-\frac{d}{p}+\frac{d}{2}} \ge cm^{-(\frac{s}{d}-\frac{1}{d}-\frac{1}{p}+\frac{1}{2})}, \tag{12.65}$$

which proves the lower bound in (iii) for $R^*(\mathcal{F})_{H^1(\Omega)}$.

The lower bound in case (iv): We consider first the case where $p \le \gamma$, where γ is given by (3.12). In this case, we choose $\eta = \eta_I$ where I is any cube in $\mathcal{I}$. The function η vanishes at all of the data sites. By construction, we have that $\eta \in \mathcal{F}$ and the lower bound

$$\eta(x) \ge cm^{-\frac{s}{d}+\frac{1}{p}}, \quad x \in I_0, \tag{12.66}$$

where $I_0 = [\xi_I - \ell_I/4, \xi_I + 1/4]^d$ is the subcube corresponding to Ω_0 in I (see (12.57)).

We want to bound the $H^{-1}(\Omega)$ norm of η from below. For this, we choose a function v from the unit ball of $H_0^1(\Omega)$ which vanishes outside of I and is large on I_0. Namely, we take

$$v(x) := c\ell_I^{1-\frac{d}{2}}\varphi(\ell_I^{-1}(x - \xi_I)). \tag{12.67}$$

where now c is chosen as a constant depending only on d so that $\|v\|_{H_0^1(\Omega)} = 1$. As in the case for η_I, we have

$$v(x) \ge c\ell_I^{1-\frac{d}{2}} \ge cm^{-\frac{1}{d}+\frac{1}{2}}, \quad x \in I_0. \tag{12.68}$$

It follows that

$$\|\eta\|_{H^{-1}(\Omega)} \ge \int_\Omega \eta(x)v(x)\,dx = \int_{I_0} \eta(x)v(x)\,dx \ge cm^{-\frac{s}{d}+\frac{1}{p}}m^{-\frac{1}{d}+\frac{1}{2}}|I_0|. \tag{12.69}$$

Since $|I_0| \ge cm^{-1}$ the right side of (12.69) is

$$cm^{-\frac{s}{d}+\frac{1}{p}-\frac{1}{d}-\frac{1}{2}} = cm^{-\frac{s}{d}+\frac{1}{p}-\frac{1}{\gamma}}.$$

This proves that $R^*(\mathcal{F})_{H^{-1}(\Omega)} \ge cm^{-\alpha_{-1}}$ in the case $p \le \gamma$ since $1/p - 1/\gamma \ge 0$.

We finally consider the case when $p > \gamma$. This is handled in a similar way to case (ii). Namely, we can assume $\mathcal{F} = U(B_q^s(L_\infty(\Omega)))$ where s, q are fixed. We take η as in (12.62). We know this function is in $\mathcal{F}$ and vanishes at all of the data sites. To provide a lower bound the $H^{-1}(\Omega)$ norm of η,

$$v(x) = c\varphi(x), \tag{12.70}$$

where c is a constant chosen so that $\|v\|_{H_0^1(\Omega)} = 1$. By construction, we have $v(x) \ge 0$ for all $x \in \Omega$ and $v(x) \ge c$ for $x \in \Omega_0$. Let $\mathcal{I}_0$ be the set of $I \in \mathcal{I}$ such that $I \subset \Omega_0$. Since η is also non-negative, we estimate

$$\|\eta\|_{H^{-1}(\Omega)} \ge \int_\Omega \eta(x)v(x)dx \ge \int_{\Omega_0} \eta(x)v(x)dx \ge \int_{\Omega_0} \eta(x)dx \ge \kappa \sum_{I\in\mathcal{I}_0}\int_I \eta_I \ge c2^{-ks}2^{-kd}\#(\mathcal{I}_0), \tag{12.71}$$

because $\kappa \ge c$ and $\eta_I \ge 2^{-ks}$ on I_0. We need to estimate the cardinality of the set $\mathcal{I}_0$. Observe that the number of cubes I in the original grid which are contained in Ω_0 is $2^{-d}(4^d m) = 2^d m$. Since $\#(\mathcal{I}) \ge (4^d - 1)m$, we see that

$$\#(\mathcal{I}_0) \ge 2^d m - m = (2^d - 1)m.$$

Since $m = 2^{kd}$, placing this lower bound of $\#(\mathcal{I}_0)$ into (12.71), we find that

$$\|\eta\|_{H^{-1}(\Omega)} \ge cm^{-\frac{s}{d}}. \tag{12.72}$$

This implies that $R^*(\mathcal{F})_{H^{-1}(\Omega)} \ge cm^{-\alpha_{-1}}$ in the case $p > \gamma$ since $1/p - 1/\gamma < 0$ and therefore completes the proof of the lower bound in (iv). This completes the proof of Theorem 3.1. □

12.6.2 The proof of Theorem 3.2

We first prove (ii) which also gives the upper bound in (3.19). From the definition of the trace norm, we have

$$\|g - \overline{S}_k(g)\|_{H^{1/2}(\partial\Omega)} \leq \|v - S_k\|_{H^1(\Omega)} \leq C[r2^k]^{-t+1} \leq Cm^{-\beta}, \tag{12.73}$$

where the next to last inequality uses the estimate (2.18) and also the fact that $m \asymp 2^{k(d-1)}$. This proves (ii).

It follows that

$$R_m(g)_{H^{1/2}(\partial\Omega)} \leq Cm^{-\beta}, \tag{12.74}$$

for the values of m that equal $2d[(r-1)2^k]^{d-1}$. Since $g \in \mathcal{G}$ was arbitrary, we obtain

$$R_m^*(\mathcal{G})_{H^{1/2}(\partial\Omega)} \leq Cm^{-\beta}, \tag{12.75}$$

for the above values of m. From the monotonicity of R_m^* we obtain (12.75) for all m. Thus, we have proven the upper bound in (3.19).

We next prove the lower inequalities in (3.19). From the monotonicity of $R_m^*(\mathcal{G})_{H^{1/2}(\partial\Omega)}$, it is enough to prove this lower bound for $m = 2^{k(d-1)} - 1$ whenever k is any non-negative integer. The following reasoning is the same as in the proof of the lower inequalities in Theorem 3.1.

Let $\mathcal{Z} := \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$, $m = 2^{k(d-1)} - 1$, be any proposed set of data sites on $\partial\Omega$. Let $F := \{\mathbf{x} \in \Omega : \mathbf{x} \cdot e_1 = 0, e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$, be the face of $\partial\Omega$ corresponding to the points $\mathbf{x} \in \overline{\Omega}$ whose first coordinate is equal to zero. Consider the set $\mathcal{D}_k(F)$ of ($d-1$ dimensional) dyadic cubes of F. Since there are $2^{k(d-1)}$ cubes in $\mathcal{D}_k(F)$, it follows that there is a $\overline{J} \in \mathcal{D}_k(F)$ such that $\overline{J}$ contains none of the data sites from $\mathcal{Z}$ in its interior. We will now construct an appropriate function $\eta \in \mathcal{G}$ which vanishes at each of the data sites.

Let J be the d dimensional cube in $\mathcal{D}_k(\overline{\Omega})$ which has $\overline{J}$ as a face and let φ_J be the function defined in (12.59) with the parameters $\bar{s}, d, \bar{p}, \bar{q}$. We know that $M^{-1}\varphi_J \in U(B_{\bar{q}}^{\bar{s}}(L_{\bar{p}}(\Omega))$ when $M := M_{\bar{s},\bar{p},\bar{q}}$. If ξ_J' is the center of J, then the point $\xi_J' - (2^{-k-1}, 0, \ldots, 0)$ is the center of $\overline{J}$.

We now define

$$v(\mathbf{x}) := M^{-1}\varphi_J(\mathbf{x} - (2^{-k-1}, 0, \ldots, 0)), \quad x \in \Omega, \tag{12.76}$$

which is also a function in $U(B_{\bar{q}}^{\bar{s}}(L_{\bar{p}}(\Omega))$. Hence, the function

$$\eta := T_{\partial\Omega} v \tag{12.77}$$

is in our model class $\mathcal{G}$ and η vanishes at all of the data sites $\mathcal{Z}$.

We now show that for the intrinsic $H^{1/2}(\partial\Omega)$ norm we have

$$\|\eta\|_{H^{1/2}(\partial\Omega)} \geq cm^{-\beta}, \tag{12.78}$$

with c not depending on m. This will prove the lower bound we seek. First consider the function φ defined in (12.57). Let $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$. Because of the tensor product structure of φ described after (12.57), the trace η_0 of φ onto the hyperplane $\mathbf{x} \cdot e_1 = 1/2$ is the $d-1$ version of φ. We define

$$\overline{M} := |\eta_0|_{H^{1/2}(\partial\Omega)} > 0. \tag{12.79}$$

By a change of variables and the fact that $m = 2^{k(d-1)} - 1$, it follows that

$$|\eta|_{H^{1/2}(\partial\Omega)} = \overline{M}M^{-1}2^{-k(\bar{s}-d/\bar{p})}2^{k/2}2^{-k(d-1)/2} \geq cm^{-\beta}, \tag{12.80}$$

where $c > 0$ does not depend on m. This proves the lower bound and completes the proof of the theorem. □

12.6.3 Proof of Theorem 3.3

We first prove the right inequality of (3.21). Given $m \geq 1$, let $\mathcal{X}$ be a collection of $\widetilde{m} := \lfloor \frac{m}{2} \rfloor$ data sites in $\overline{\Omega}$ for optimal sampling of the functions in $\mathcal{F}$. Similarly let $\mathcal{Z}$ be a collection of $\overline{m} := \lfloor \frac{m}{2} \rfloor$ data sites in $\partial\Omega$ for optimal sampling of the functions in $\mathcal{G}$. Given $u \in \mathcal{U}$ let $f = -\Delta u$ and g be the trace of u on $\partial\Omega$. We know that there is an $\widehat{f} \in H^{-1}(\Omega)$ which is an optimal $H^{-1}(\Omega)$ recovery of f using the point samples at $\mathcal{X}$. That is, we have $\|\widetilde{f} - \widehat{f}\|_{H^{-1}(\Omega)} \leq Cm^{-\alpha}$ whenever $\widetilde{f} \in \mathcal{F}$ shares the same data as f. Similarly, let $\widehat{g}$ be an optimal $H^{1/2}(\partial\Omega)$ recovery of g using the data at $\mathcal{Z}$. Then, we have $\|\widetilde{g} - \widehat{g}\|_{H^{1/2}(\partial\Omega)} \leq Cm^{-\beta}$ whenever $\widetilde{g}$ is a function in $\mathcal{G}$ which shares the same data as g. Now, define $\widehat{u}$ as the solution to (1.1) with right side $\widehat{f}$ and boundary value $\widehat{g}$. Let $\widetilde{u}$ be any function from $\mathcal{U}$ which shares the same data as u. From (1.3), we have

$$\|\widetilde{u} - \widehat{u}\|_{H^1(\Omega)} \leq Cm^{-\min(\alpha,\beta)},$$

where C is independent of m and the data. Since this holds for all $u \in \mathcal{U}$, this proves the upper inequality in (3.21)

To prove the lower inequality we consider two cases. First suppose that $\alpha \leq \beta$. Given any set $\mathcal{X}$ of m data sites in $\overline{\Omega}$ at which we sample the $f \in \mathcal{F}$, we can find two functions $f_1, f_2 \in \mathcal{F}$ that have the same values on $\mathcal{X}$ and yet $\|f_1 - f_2\|_{H^{-1}(\Omega)} \geq cm^{-\alpha}$ where c does not depend on m. For $i = 1, 2$, we let u_i satisfy (1.1) with right side f_i and with $g_i \equiv 0$. Then, (1.3) says that

$$\|u_1 - u_2\|_{H^1(\Omega)} \geq cm^{-\alpha} = cm^{-\min(\alpha,\beta)}.$$

This proves the lower inequality in (3.21) in this case. A similar argument applies when $\beta < \alpha$. □

12.7 The proof of Theorem 7.2 in the case $d = 2$

In this section we provide the proof of Theorem 7.2 in when $d = 2$. We discuss two cases.

Case 1: $p = 1$

Let $\varepsilon > 0$. According to Lemma 12.4, Lemma 6.1 and Theorem 6.3, we have

$$\begin{aligned}
\|u - v\|_{H^1(\Omega)} &\lesssim \|f + \Delta v\|_{H^{-1}(\Omega)} + \|g - Tr(v)\|_{H^{1/2}(\partial\Omega)} \\
&\lesssim \varepsilon^{-1}\|f + \Delta v\|_{L_{1+\varepsilon}(\Omega)} + \|g - Tr(v)\|_{H^{1/2}(\partial\Omega)} \\
&\lesssim \varepsilon^{-1}\|f + \Delta v\|^*_{L_{1+\varepsilon}(\Omega)} + \|g - Tr(v)\|^*_{H^{1/2}(\Omega)} \\
&+ \left[\|f + \Delta v\|_{\mathcal{B}}\varepsilon^{-1}\widetilde{m}^{1-1/(1+\varepsilon)}\widetilde{m}^{-\frac{s}{2}} + \|g - Tr(v)\|_{Tr(\overline{\mathcal{B}})}\overline{m}^{-(\bar{s}-1)}\right].
\end{aligned}$$

Optimizing ε as in Lemma 12.4 gives a choice of $\varepsilon = [\log(\widetilde{m})]^{-1}$, for which

$$\varepsilon^{-1}\widetilde{m}^{1-1/(1+\varepsilon)} \lesssim \log(\widetilde{m})$$

and thus we have

$$\begin{aligned}
\|u - v\|_{H^1(\Omega)} &\lesssim \log(\widetilde{m})\|f + \Delta v\|^*_{L_{1+\varepsilon}(\Omega)} + \|g - Tr(v)\|^*_{H^{1/2}(\Omega)} \\
&+ \left[\|f + \Delta v\|_{\mathcal{B}}\log(\widetilde{m})\widetilde{m}^{-\frac{s}{2}} + \|g - Tr(v)\|_{Tr(\overline{\mathcal{B}})}\overline{m}^{-(\bar{s}-1)}\right] \\
&\lesssim \mathcal{L}^*(v) + [1 + \|v\|_{\mathcal{U}}]\,\mathcal{R}_{\mathcal{U}}(\widetilde{m}, \overline{m}).
\end{aligned}$$

Case 2: $1 < p \leq \infty$

$$
\begin{aligned}
\|u-v\|_{H^1(\Omega)} &\lesssim \|f+\Delta v\|_{H^{-1}(\Omega)} + \|g - Tr(v)\|_{H^{1/2}(\partial\Omega)} \\
&\lesssim C(p)\|f+\Delta v\|_{L_p(\Omega)} + \|g - Tr(v)\|_{H^{1/2}(\partial\Omega)} \\
&\lesssim \left[C(p)\|f+\Delta v\|^*_{L_p(\Omega)} + \|g - Tr(v)\|^*_{H^{1/2}(\partial\Omega)}\right]. \\
&+ \left[C(p)\|f+\Delta v\|_{\mathcal{B}}\widetilde{m}^{-\frac{s}{2}} + \|g - Tr(v)\|_{Tr(\overline{\mathcal{B}})}\overline{m}^{-(\bar{s}-1)}\right] \\
&\lesssim \left[C(p)\|f+\Delta v\|^*_{L_p(\Omega)} + \|g - Tr(v)\|^*_{H^{1/2}(\Omega)}\right] + [1+\|v\|_{\mathcal{U}}]\,\mathcal{R}_{\mathcal{U}}(\widetilde{m},\overline{m}).
\end{aligned}
$$

Notice that in this case we could have chosen $\mathcal{L}^*$ as in the case of $d=3$ with $\gamma = p$ and obtained the optimal recovery rate for this class. However, this choice would result in a loss functional depending on $\mathcal{F}$.

References

[1] D. Arnold and J. Saraven, *On the asymplotic convergence of spline-collocation methods for partial differential equations*, SIAM J. Num. Anal. **21** (1984), 459–472.

[2] Genming Bai, Ujjwal Koley, Siddhartha Mishra, and Roberto Molinaro, *Physics informed neural networks (pinns) for approximating nonlinear dispersive pdes*, arXiv preprint arXiv:2104.05584 (2021).

[3] Jens Berg and Kaj Nyström, *A unified deep artificial neural network approach to partial differential equations in complex geometries*, Neurocomputing **317** (2018), 28–41.

[4] Peter Binev, Andrea Bonito, Ronald DeVore, and Guergana Petrova, *Optimal learning*, Calcolo **61** (2024), no. 1, 15.

[5] Andrea Bonito, Claudio Canuto, Ricardo H Nochetto, and Andreas Veeser, *Adaptive finite element methods*, to appear in Acta Numerica (2024).

[6] Shengze Cai, Zhiping Mao, Zhicheng Wang, Minglang Yin, and George Em Karniadakis, *Physics-informed neural networks (PINNs) for fluid mechanics: A review*, Acta Mechanica Sinica **37** (2021), no. 12, 1727–1738.

[7] Philippe G Ciarlet, *The finite element method for elliptic problems*, SIAM, 2002.

[8] Albert Cohen, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk, *Optimal stable nonlinear ap- proximation*, Foundations of Computational Mathematics **22** (2022), 607–648.

[9] ———, *Optimal stable nonlinear approximation*, Foundations of Computational Mathematics **22** (2022), no. 3, 607–648.

[10] Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli, *Scientific machine learning through physics–informed neural networks: Where we are and what's next*, Journal of Scientific Computing **92** (2022), no. 3, 88.

[11] Stephan Dahlke and Ronald A DeVore, *Besov regularity for elliptic boundary value problems*, Communications in Partial Differential Equations **22** (1997), no. 1-2, 1–16.

[12] Stephan Dahlke, Erich Novak, and Winfried Sickel, *Optimal approximation of elliptic problems by linear and nonlinear mappings i*, Journal of Complexity **22** (2006), no. 1, 29–49.

[13] ———, *Optimal approximation of elliptic problems by linear and nonlinear mappings ii*, Journal of Complexity **22** (2006), no. 4, 549–603.

[14] Caio Davi and Ulisses Braga-Neto, *Pso-pinn: Physics-informed neural networks trained with particle swarm optimization*, arXiv preprint arXiv:2202.01943 (2022).

[15] R. DeVore and G. Lorentz, *Constructive approximation*, Springer-Verlag, 1993.

[16] R. DeVore and V. Popov, *Interpolation of Besov spaces*, Transactions AMS **305** (1988), no. 1, 397–414.

[17] R. DeVore and R. Sharpley, *Maximal functions measuring smoothness*, Memoirs of AMS **293** (1984).

[18] ———, *Besov spaces on domains in R^d*, TAMS **335** (1993), no. 2, 843–364.

[19] Ronald DeVore, Boris Hanin, and Guergana Petrova, *Neural network approximation*, Acta Numerica **30** (2021), 327–444.

[20] MWMG Dissanayake and Nhan Phan-Thien, *Neural-network-based approximations for solving partial differential equations*, communications in Numerical Methods in Engineering **10** (1994), no. 3, 195–201.

[21] Patrick Dondl, Johannes Müller, and Marius Zeinhofer, *Uniform convergence guarantees for the deep ritz method for nonlinear problems*, Advances in Continuous and Discrete Models **2022** (2022), no. 1, 49.

[22] Nathan Doumèche, Gérard Biau, and Claire Boyer, *Convergence and error analysis of pinns*, arXiv preprint arXiv:2305.01240 (2023).

[23] Chenguang Duan, Yuling Jiao, Yanming Lai, Dingwei Li, Jerry Zhijian Yang, et al., *Convergence rate analysis for deep ritz method*, Communications in Computational Physics **31** (2022), no. 4, 1020–1048.

[24] Dimitrios Gazoulis, Ioannis Gkanis, and Charalambos G Makridakis, *On the stability and convergence of physics informed neural networks*, arXiv preprint arXiv:2308.05423 (2023).

[25] David Gilbarg and Neil S Trudinger, *Elliptic partial differential equations of second order*, vol. 224, Springer, 1977.

[26] Xavier Glorot and Yoshua Bengio, *Understanding the difficulty of training deep feedforward neural networks*, Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[27] Pierre Grisvard, *Elliptic problems in nonsmooth domains*, SIAM, 2011.

[28] Guergana and Przemyslaw Wojtaszczyk, *Limitations on approximation by deep and shallow neural networks*, Journal of Machine Learning Research **24** (2023), no. 353, 1–38.

[29] ———, *Lipschitz widths*, Constructive Approximation **57** (2023), 759–805.

[30] ———, *Neural networks: deep, shallow, or in between?*, arXiv:2310.07190 (2023).

[31] Houstis Hadjidimos, A., J. E., Rice, and E. Vavalis, *Analysis of iterative line spline collocation methods for elliptic partial differential equations*, SIAM J. MAtrix Analysis and Applications **21** (2000), no. 2.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[33] ———, *Deep residual learning for image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[34] E. Houslis, J. Rice, and E. Vavalis, *Spline-collocation methods/or elliptic partial differential equations*, Advances in Compo Methodsfor PDE's. Vol. V, R. Vichnevetsky, R. S. Stepleman (Editors), 1984.

[35] David Jerison and Carlos E Kenig, *The inhomogeneous dirichlet problem in lipschitz domains*, Journal of functional analysis **130** (1995), no. 1, 161–219.

[36] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang, *Physics-informed machine learning*, Nature Reviews Physics **3** (2021), no. 6, 422–440.

[37] Alena Kopaničáková, Hardik Kothari, George E Karniadakis, and Rolf Krause, *Enhancing training of physics-informed neural networks using domain decomposition–based preconditioning strategies*, SIAM Journal on Scientific Computing (2024), S46–S67.

[38] David Krieg, Erich Novak, and Mathias Sonnleitner, *Recovery of Sobolev functions restricted to iid sampling*, Mathematics of Computation **91** (2022), no. 338, 2715–2738.

[39] Isaac E Lagaris, Aristidis Likas, and Dimitrios I Fotiadis, *Artificial neural networks for solving ordinary and partial differential equations*, IEEE transactions on neural networks **9** (1998), no. 5, 987–1000.

[40] Isaac E Lagaris, Aristidis C Likas, and Dimitris G Papageorgiou, *Neural-network methods for boundary value problems with irregular boundaries*, IEEE Transactions on Neural Networks **11** (2000), no. 5, 1041–1049.

[41] Jianfeng Lu and Yulong Lu, *A priori generalization error analysis of two-layer neural networks for solving high dimensional schrödinger eigenvalue problems*, Communications of the American Mathematical Society **2** (2022), no. 1, 1–21.

[42] Yulong Lu, Jianfeng Lu, and Min Wang, *A priori generalization analysis of the deep ritz method for solving high dimensional elliptic partial differential equations*, Conference on learning theory, PMLR, 2021, pp. 3196–3241.

[43] Zhiping Mao, Ameya D Jagtap, and George Em Karniadakis, *Physics-informed neural networks for high-speed flows*, Computer Methods in Applied Mechanics and Engineering **360** (2020), 112789.

[44] Levi D McClenny and Ulisses M Braga-Neto, *Self-adaptive physics-informed neural networks*, Journal of Computational Physics **474** (2023), 111722.

[45] Siddhartha Mishra and Roberto Molinaro, *Estimates on the generalization error of physics-informed neural networks for approximating PDEs*, IMA Journal of Numerical Analysis **43** (2023), no. 1, 1–43.

[46] Dorina Mitrea, Marius Mitrea, and Lixin Yan, *Boundary value problems for the laplacian in convex and semiconvex domains*, Journal of Functional Analysis **258** (2010), no. 8, 2507–2585.

[47] Johannes Müller and Marius Zeinhofer, *Error estimates for the deep ritz method with boundary penalty*, Mathematical and Scientific Machine Learning, PMLR, 2022, pp. 215–230.

[48] ———, *Achieving high accuracy with PINNs via energy natural gradient descent*, International Conference on Machine Learning, PMLR, 2023, pp. 25471–25485.

[49] Erich Novak and Hans Triebel, *Function spaces in lipschitz domains and optimal rates of convergence for sampling*, Constructive approximation **23** (2006), 325–350.

[50] Maziar Raissi, Paris Perdikaris, and George E Karniadakis, *Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations*, Journal of Computational physics **378** (2019), 686–707.

[51] Thomas Sauer and Yuan Xu, *A case study in multivariate Lagrange interpolation*, Approximation Theory, Wavelets and Applications **454** (1995), 443–452.

[52] Zuowei Shen, Haizhao Yang, and Shijun Zhang, *Optimal approximation rate of ReLU networks in terms of width and depth*, Journal de Mathématiques Pures et Appliquées **157** (2022), 101–135.

[53] Y. Shin, D. Jerome, and G. Karniadakis, *On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type PDEs*, Commun. Comput. Phys. **28** (2020), no. 5, 2042–2074.

[54] Yeonjong Shin, Zhongqiang Zhang, and George Em Karniadakis, *Error estimates of residual minimization using neural networks for linear PDEs*, Journal of Machine Learning for Modeling and Computing **4** (2023), no. 4.

[55] Jonathan W Siegel, *Optimal approximation rates for deep ReLU neural networks on Sobolev and Besov spaces*, Journal of Machine Learning Research **24** (2023), no. 357, 1–52.

[56] Jonathan W Siegel, Qingguo Hong, Xianlin Jin, Wenrui Hao, and Jinchao Xu, *Greedy training algorithms for neural networks and applications to PDEs*, Journal of Computational Physics **484** (2023), 112084.

[57] E. Stein, *Singular integrals and differentiability properties of functions*, Princeton University Press, 1970.

[58] Kejun Tang, Xiaoliang Wan, and Chao Yang, *Das-pinns: A deep adaptive sampling method for solving high-dimensional partial differential equations*, Journal of Computational Physics **476** (2023), 111868.

[59] Jan Vybíral, *Sampling numbers and function spaces*, Journal of Complexity **23** (2007), no. 4-6, 773–792.

[60] E Weinan and Bing Yu, *The deep Ritz method: A deep learning-based numerical algorithm for solving variational problems*, Communications in Mathematics and Statistics **6** (2018), no. 1.

[61] Jinchao Xu, *Finite neuron method and convergence analysis*, Communications in Computational Physics **28** (2020), no. 5, 1707–1745.

[62] Yahong Yang, Yue Wu, Haizhao Yang, and Yang Xiang, *Nearly optimal approximation rates for deep super RELU networks on Sobolev spaces*, arXiv preprint arXiv:2310.10766 (2023).

[63] Yahong Yang, Haizhao Yang, and Yang Xiang, *Nearly optimal VC-dimension and pseudo-dimension bounds for deep neural network derivatives*, Advances in Neural Information Processing Systems **36** (2024).

[64] Dmitry Yarotsky, *Optimal approximation of continuous functions by very deep ReLU networks*, Conference on learning theory, PMLR, 2018, pp. 639–649.

[65] Marius Zeinhofer, Rami Masri, and Kent-André Mardal, *A unified framework for the error analysis of physics-informed neural networks*, arXiv preprint arXiv:2311.00529 (2023).

[66] Shiheng Zhang, Jiahao Zhang, Jie Shen, and Guang Lin, *An element-wise RSAV algorithm for unconstrained optimization problems*, arXiv preprint arXiv:2309.04013 (2023).

Andrea Bonito, Department of Mathematics, Texas A&M University, College Station, TX 77843, email: bonito@tamu.edu.

Ronald DeVore, Department of Mathematics, Texas A&M University, College Station, TX 77843, email: rdevore@tamu.edu.

Guergana Petrova, Department of Mathematics, Texas A&M University, College Station, TX 77843, email: gpetrova@tamu.edu.

Jonathan W. Siegel, Department of Mathematics, Texas A&M University, College Station, TX 77843, email: jwsiegel@tamu.edu.